# Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis

**Suzanne L. Slocum-Gori · Bruno D. Zumbo**

**Abstract** Whenever one uses a composite scale score from item responses, one is tacitly assuming that the scale is dominantly unidimensional. Investigating the unidimensionality of item response data is an essential component of construct validity. Yet, there is no universally accepted technique or set of rules to determine the number of factors to retain when assessing the dimensionality of item response data. Typically factor analysis is used with the eigenvalues-greater-than-one rule, the ratio of first-to-second eigenvalues, parallel analysis, root-mean-square-error-of-approximation, or hypothesis testing approaches involving chi-square tests from Maximum Likelihood or Generalized Least Squares estimation. The purpose of this study was to investigate how these various procedures perform individually, and in combination, when assessing the unidimensionality of item response data via a computer simulated design. Conditions such as sample size, magnitude of communality, distribution of item responses, proportion of communality on second factor, and the number of items with non-zero loadings on the second factor were varied. Results indicate that there was no one individual decision-making method that identified unidimensionality under all conditions manipulated. Given the low communalities, all individual decision-making methods failed to detect unidimensionality for the combination where sample size was small, magnitude of communality was low, and item distributions were skewed. A set of guidelines and a new statistical methodology are provided for researchers.

S. L. Slocum-Gori (✉)
School of Population and Public Health, Faculty of Medicine, University of British Columbia, 2206 East Mall, Vancouver, BC V6T 1Z3, Canada
e-mail: sslocum@bccancer.bc.ca

B. D. Zumbo
Measurement, Evaluation, & Research Methodology Program, Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

## 1 Introduction

Millions of people under various circumstances are tested every year. Inferences made from the results of these tests are used to make high-stake decisions in education and social-welfare systems, as well as for diagnosis and treatment in health care systems. The test scores, in particular, are used for research and policy purposes. The inferences made from these test scores are highly dependent upon the accuracy and validity of the interpretation of the test results. According to the Standards for Educational and Psychological Testing (American Psychological Association et al. 1999), validity refers to "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9). As Hubley and Zumbo (1996) state, a test score is not necessarily directly linked to the construct that is being measured, but should actually be considered one (of various) indicators of the construct. Furthermore, Messick (1975) and Guion (1977) claim that in order to adequately assess the meaningfulness of the inference(s) one makes from test scores, one should have some kind of confirmation of what the test score itself actually exhibits. Therefore, in order to properly assess the appropriateness, meaningfulness, and usefulness of the inferences made from test scores, one needs to take construct validity into consideration (see recent reviews by Kane 2006; Lissitz 2009; Sireci 2009; Zumbo 2007, 2009).

Construct validity seeks agreement between a theoretical concept and a specific measurement (i.e., test). Messick (1988) has claimed that one major concern of the validity of inferences is the unanticipated or negative consequences of test score interpretation, which can often be traced to construct under-representation or construct-irrelevant variance. If a measurement instrument excludes essential facets of a construct, or if the instrument includes facets that may be irrelevant, the inferences made from the test scores could involve a variety of inappropriate consequences. For example, if an instrument intended to only measure depression, and actually included secondary facets of anxiety, the inferences made from the test scores could lead to inappropriate diagnosis and treatment. One way, among many, to prevent inappropriate consequences from test score interpretation is to focus on the theoretical dimensions of the construct a test is intending to measure.

Building on an earlier psychometric tradition (e.g. Humphreys 1952, 1962; Stout 1987), Slocum-Gori et al. (2009) recently made the case that that measures used in quality of life and happiness research are essentially unidimensional: inherently tapping minor dimensions. A case was made that psycho-social variables need not meet the standard of strict unidimensionality, but that the interpretation of the total scale score is not compromised because the additional dimensions are relatively minor. The notion of essential unidimensionality, with one dominant dimension, substantially complicates the dimensionality issue for psycho-social measures because one must be able to discern essential unidimensionality from genuine multi-dimensionality.

Investigating the dimensionality (i.e., the structure of a specific phenomenon) of tests is an essential component of construct validity. When the items of a test are summed to one total score, there is a tacit assumption that the test is measuring one dimension and that this dimension is measured on a continuum. If, for example, a measure is actually multi-dimensional yet is being treated as strictly unidimensional, interpretation of the test scores can be invalid and entail harmful consequences, as shown above with the depression example.

### 1.1 Problem Under Investigation

Investigating the dimensionality of test data is one of the most commonly encountered activities in day-to-day research. The structure (i.e., dimensionality) of item response data

is composed of a certain number of latent factors, sometimes also referred to as latent variables in the methodological research literature. The decision of how many factors to retain in a factor analysis not only explains the structure of the underlying phenomena, but also accounts for the relationships between the test items. This verdict is critical in that it tends to be one of the most common problems facing researchers who assess the dimensionality of measures (Crawford 1975; Fabrigar et al. 1999). When assessing dimensionality, the goal is to obtain factor solutions that are reliable and mirror the population factor structure. Evidently, it has been noted by several researchers that this verdict can lead to researchers deciding on an erroneous number of factors (e.g. Crawford 1975; Fabrigar et al. 1999; Russell 2002). Moreover, the consequences of over- and under-extracting the number of factors can lead to inappropriate inferences and decisions (e.g., high-stakes measurement decisions or inappropriate interventions). Yet, there is no universally accepted technique or set of rules to determine the number of factors to retain when assessing the dimensionality of item response data.

Previous research has shown that many of the decision-making rules and indices that are commonly used to identify dimensionality do not work efficiently under certain conditions (e.g., small sample sizes), over-and under-extract the number of factors and have a limited range of accuracy (Gorsuch 1983; Fabrigar et al. 1999; Russell 2002; Slocum-Gori et al. 2009). Most importantly, however, observations made by Hattie (1985) and Lord (1980) still hold today. Specifically, Hattie noted that there has been no empirical study that has examined the efficiency of combinations of these rules, indices and methods. Lord declared that there is a need for an index, rule or algorithm to ascertain unidimensionality, in particular.

Furthermore, it has been recommended that researchers apply multiple criteria when deciding on the appropriate number of factors to retain (Gessaroli and De Champlain 1996; Fabrigar et al. 1999; Davison and Sireci 2000). In fact, researchers are utilizing multiple criteria in practice, but the rationale behind the combinations is unclear and the selection of combinations may not necessarily prove to be effective for that particular combination of sample size and item characteristics. Although there are numerous rules and indices that researchers can utilize in order to determine the dimensionality of item response data, there are a select few that have been investigated in the literature. These rules have been noted as being commonly used and, under certain conditions, successful in determining dimensionality (Hattie 1984; Fabrigar et al. 1999; Russell 2002). Such rules and indices include (1) the chi-square statistic from Maximum Likelihood (ML) factor analysis, (2) the Chi-square statistic from Generalized Least Squares (GLS) factor analysis, (3) the eigenvalues-greater-than-one rule, (4) the ratio-of- first- to-second-eigenvalues- greater-than-three rule, (5) the ratio-of- first- to-second-eigenvalues-greater-than-four rule, (6) parallel analysis (PA) using continuous data, (7) parallel analysis (PA) using ordinal or rating scale (Likert) data, (8) the Root Mean Square of Approximation (RMSEA) index from ML estimation, and (9) the Root Mean Square of Approximation (RMSEA) index from GLS estimation.

The purpose of this study is to extend previous research by investigating how the nine decision-making rules and indices perform individually and in combination under varying conditions (e.g., different sample sizes or magnitudes of communality) when assessing the underlying unidimensionality of item response data that is often found in psychological, educational and health measurements (e.g., subjective well-being, depression, or motivation). These rules and indices are widely used in practice and exhibit strong empirical support, although under certain circumstances (e.g., small sample size) have performed unsuccessfully. The overall objective of the present study is to provide guidelines to assist the social and behavioral science researcher in the decision-making process of retaining

factors in an assessment of unidimensionality, which will focus on informing day-to-day research practice.

The scoring of item response data rests on an implicit principal: when test items are summed to one total scale score, there is a tacit assumption that a test is unidimensional (i.e., one factor) leading to the common psychometric research question: Do the items measure just one latent variable? (Zumbo et al. 2002). The inferences from a total scale score are made in regards to the *one* latent variable (i.e., factor). Lord's (1980) claim, that there is a need for an index or rule to define unidimensionality, is one that is critical. Because there is not a universal and sound index or methodology to assess unidimensional measures, one may question the validity of the inferences made from total scale scores. This study focused on developing a (universal) methodology in order to assess unidimensionality.

### 1.2 Dimensionality

Underlying phenomenon is considered to be the reason why observed variables are correlated in the first place. The underlying phenomena can reflect one or more dimensions. Dimensionality refers to the structure of a specific phenomenon (Pett et al. 2003). Unidimensionality, in particular, refers to one dominant latent variable or phenomena. There are several statistical procedures that provide a structural analysis of a selected set of observed variables (e.g., factor analysis (FA) or multidimensional scaling). Ultimately, these procedures ideally obtain an appropriate number of dimensions to justify the use of composite scores and to explain the pattern of correlations among observed variables. Dimensions (i.e., latent variables) are known to be constructed variables that come prior to the observed variables (i.e., test items). That is, it is assumed that if two test items are correlated, they have something unobserved in common (i.e., the latent dimension). Despite the significance of the theoretical and procedural aspects of assessing the dimensionality of measures via FA, there is not an acceptable index to represent the unidimensionality of a set of test items.

### 1.3 Strict and Essential Unidimensionality

Psycho-educational and health measures often utilize composite scale scores in order to make inferences. Although some of the psycho-educational and health measures meet strict unidimensionality, most of these measures are actually essential unidimensional (Slocum-Gori et al. 2009). According to Humphreys (1952, 1962), in order to measure any psychological latent variable of interest, the inclusion of numerous minor latent variables is not merely desirable but unavoidable. The inclusion of secondary minor latent variables is often referred to as 'essential unidimensionality'. Strict unidimensionality, on the other hand, is defined as one dominant latent variable with no secondary minor dimensions. Both essential and strict unidimensionality were investigated in this study. Strict unidimensionality was investigated via manipulation of the magnitude of the communalities ($h^2$). Essential unidimensionality was also examined by varying the magnitude of the communalities, but the proportion of communality on the second factor and the number of test items with non-zero loadings on the second factor was also manipulated for essential unidimensionality. The methodology of this study provides an operational definition for *essential* unidimensionality: manipulation of the magnitude of communality, proportion of communality on the second factor, and the number of test items with non-zero loadings on the second factor.

## 1.4 Research Objectives

The key aim of this research was to fill the gaps and attempt to resolve some of the discrepancies in the research literature as to what rules or combination of rules and indices are superior when retaining factors. The overall objective was to provide guidelines to assist researchers in the decision-making process of retaining factors. The primary contributions of this research were to provide (1) a comprehensive direct comparison of a variety of decision-making rules and indices. This allowed us to determine if there was, in fact, one superior method; and (2) an investigation of whether multiple rules and indices (e.g., eigenvalues-greater-than-one rule and ML RMSEA) aid in the decision-making process of retaining factors. This strategy was suggested in the literature as early as 25 years ago (Hattie 1984), but had never been investigated to date. Ultimately, the goal was to develop a new statistical methodology that involved multiple criteria when determining the number of factors to retain.

## 1.5 Scope of the Study

The scope of this study limited the selection of rules and indices. Because the focus of this study was based on unidimensional measures, the decision-making rules and indices that are used in day-to-day research, Likert-type item formats, and the availability of statistical-software, many other possible methods were not discussed or investigated in this study. Furthermore, assessment methods specifically appropriate for tests containing dichotomous items were not considered (e.g., tetrachoric correlation matrices). Finally, the conditions of interest were limited to those found in psychological assessment, ignoring issues and conditions that go along with large scale assessment (e.g., IRT-item parameters).

# 2 Methodology

## 2.1 Strict Unidimensionality

When the observed variables (i.e., test items) generate factor loadings that are high for only one factor, a strict unidimensional model is reflected. In order to reflect this notion of strict unidimensionality, a statistical model that was based on the magnitude of communality was applied. The communality is the sum of squared factor loadings for a variable across factors (Tabachnick and Fidell 1996) Therefore, a factor model cannot have high loadings and low communalities simultaneously. The factor loadings for the strict unidimensional models were generated via magnitude of communalities. Based on the calculation of a communality estimate, the factor loadings were generated by the following equation:

$$\text{Factor loading} = (\%h^2) \times \sqrt{h^2}.$$

Because there is only one factor for strict unidimensional models, 100% of the communality is allocated to factor one. The factor loadings were then imputed into the simulation design for factor analytic procedures.

In order to investigate strict unidimensional measures, three different conditions were varied: (1) magnitude of communality, (2) distribution of test items, and (3) sample size. There were two levels of communalities, two skewness indices for the distribution of test items, and three different sample sizes.

## 2.2 Conditions for the Magnitude of Communality

It has been found in several studies that factor analytic solutions are influenced by the magnitude of communality (Preacher and MacCallum 2002; MacCallum et al. 1999, 2001; Fabrigar et al. 1999). Preacher and MacCallum (2002) selected high values for the communality estimates, whereas MacCallum et al. (1999, 2001) selected a wide range for the magnitude of communalities. In order to reflect the methodology of both these simulations, the magnitudes of the communality selected for this simulation were of a wide variability and included a high degree ($h^2 = 0.20$ and 0.90).

## 2.3 Conditions for the Sample Size

The data consisted of three different sample sizes: 100, 450, 800. In order to reflect current practice, the findings from literature reviews indicate a minimum sample size of 95, which was rounded to 100 for this study. The median value of 433 from the literature was rounded to 450. The difference was then taken between 100 and 450, which was 350. This difference was added to 450 to get 800. Although Preacher and MacCallum (2002) and Schonemann (1981) utilized sample sizes as small as $n = 10$ and $n = 30$, a sample size of 100 is considered small in the psycho-educational context.

## 2.4 Conditions for Distributions

The distribution of the test items was simulated so that both symmetric and skewed conditions were included. Many psychological variables (e.g., intelligence) have bell-shaped (unimodal and symmetric) distributions (Glenberg 1996). However, distributions studied in social and behavioral sciences are often strongly skewed (Aron and Aron 2002). Therefore, both skewed and symmetric distributions were included. The skewness indicators and threshold values for a five-point Likert scale were based off of Di Stefano's (2002) methodology. Continuous data were initially generated. The first skewness indicator of zero was chosen to represent a normal distribution of item responses. The area under the curve was reported for approximately 5, 21, 48, 21, and 5% of the responses for ordered categories one through five. A second skewness indicator of 2.50 was produced to represent extreme responses to items. For the nonnormal item distributions, the percentage of responses (i.e., thresholds) in each of the five categories was approximately 75, 15, 5, 3, and 2 for categories one through five.

## 2.5 Essential Unidimensionality

In order to investigate essential unidimensional measures, five different conditions were varied: (1) magnitude of communality, (2) distribution of test items, sample size, (3) proportion of communality on the second factor, and (4) the number of test items with nonzero loadings on the second factor. The first three conditions were introduced above within the strict unidimensional methodology. In review, there were two magnitudes of communality ($h^2 = 0.20$ and 0.90), two distributions of test item (skewness = 0.00 and 2.50), and three different sample sizes ($n = 100, 450, 800$). Likewise, there were also three proportions of communality on the second factor and two different numbers of test items with nonzero loadings on the second factor manipulated for essential unidimensional models.

## 2.6 Conditions for the Proportion of Communality on Secondary Factor

In order to generate secondary dimensions of varying strengths, the proportion of communalities on the secondary minor factor varied as follows: 0.05, 0.30, and 0.50. Any value greater than 50% would indicate another dominant factor, and the current study investigated unidimensional measures only. The first factor received the higher proportions of the communality, except for the cases when factor one and the secondary minor factor received equal proportions of the communalities (i.e., 0.50).

## 2.7 Conditions for the Number of Test Items with Non-Zero Loadings on the Second Factor

The question of how many items need to load onto a factor in order for that factor to be considered a meaningful latent construct has been a question that remains unanswered. Fabrigar et al. (1999) claimed that mirroring the population factor structure is not necessarily a function of the sample size or the number of observed variables (i.e., test length), but rather is influenced by the magnitude of communality as well as overdetermination. Overdetermination is referred to as the degree to which each factor is clearly represented by a sufficient number of variables and is often assessed by the number of factors to the number of variables ratio ($p{:}r$). Highly overdetermined factors are those that exhibit high factor loadings on at least three or four variables (MacCallum et al. 1999). The more items that load on the second factor, the higher the chances for a second factor to become overdetermined, depending on whether the factor loadings are high (which is the result of the communality and proportion of communality on the second factor). The number of items with non-zero loadings on the second factor will vary as follows: three and six. These items (which will have factor loadings) will represent the varying magnitudes of secondary minor dimensions. Any number of items greater than six with non-zero loadings would represent another dominant factor, and the current study examined unidimensional measures only.

## 2.8 Factors Held Constant

There were several experimental factors that were not investigated in this study: the correlation between the first and second factor, the number of test items per scale, and the number of scale points.

## 2.9 Dependent Variables: Decision-making Rules and Indices

There are numerous decision-making rules and indices that are used to determine the number of factors to retain when assessing dimensionality. The simultaneous use of multiple decision-making rules has been recommended and practiced (Boyd and Gorsuch 2003; Fabrigar et al. 1999; Thompson and Daniel 1996). Therefore, the dependent variables for this simulation study included optimal combinations (i.e., simultaneous use) as well as the individual application of various decision-making rules and indices. As specified previously, the rules and indices included the following: (1) the Chi-square statistic from Maximum Likelihood (ML) factor analysis, (2) the Chi-square statistic from Generalized Least Squares (GLS) factor analysis, (3) the eigenvalues-greater-than-one rule, (4) the ratio-of- first- to second eigenvalues-greater-than three rule, (5) the ratio-of- first-to-second eigenvalues-greater-than four rule, (6) PA using continuous data, (7) PA using

Likert (i.e., ordinal or rating scale) data, (8) the Root Mean Square of Approximation (RMSEA) index from ML, and (9) the Root Mean Square of Approximation (RMSEA) index from GLS. The selection of these rules was based on widely available statistical packages, computational restrictions, as well as previous simulation and theoretical studies.

## 3 Procedures

### 3.1 Generation of Item Response Data

For the present simulation study, population covariance matrices were created under a variety of conditions for both strict and essential unidimensional measures, as illustrated above. The population data were produced based on factor loadings, which varied according to the strict and essential unidimensional cell structure. For each condition investigated, a population covariance matrix was computed via factor analysis. For each population matrix, 100 samples were generated with a specific sample size.

Continuous item responses were then transformed into Likert responses, in the population, for both symmetric and skewed distributions using the approach by Di Stefano (2002), as described above. This process was replicated to create the comparative data for PA Likert and PA continuous data, with all the factor loadings being 0.00, resulting in a correlation matrix with all the off-diagonal elements equal to zero.

### 3.2 FA and PCA Analyses

FA and PCA analyses were conducted for each replication, which included saving various output data using SPSS 13.0 Output Management System (OMS). The following procedures were performed on each replication: (1) FA using ML estimation, (2) FA using GLS estimation, (3) PCA, (4) PCA on the random (PA) Likert data, and (5) PCA on the random (PA) continuous data.

In summary, the simulation study was comprised of a 2 (magnitude of communality) × 2 (skewness of items) × 3 (sample size) completely crossed factor design resulting in 12 different cells or sets of conditions for *strict unidimensionality*. The simulation was also comprised of a 2 (magnitude of communality) × 2 (skewness of items) × 3 (sample size) × 3 (proportion of communality on second factor) × 2 (number of test items with nonzero loadings on the second factor) completely crossed factor design resulting in 72 different cells or sets of conditions for *essential unidimensionality*. The factor model was determined by the conditions of the cell. There were 100 replications per cell. The syntax for the FA and PCA was generated using the Output Management System in SPSS 13.0. Factor analysis of matrices was performed by ML and GLS estimation methods. Exploratory factor analysis (EFA) was conducted on all Pearson product-moment (PPM) correlation matrices via SPSS 13.0. PPM correlation matrices were selected in order to reflect current research practices.

SPSS was used, along with the Pearson correlation matrix (which is the only option available in SPSS and many other widely used statistical packages) because we aimed to reflect day-to-day research practice. The recommendation in the statistical literature to use a polychoric correlation matrix with 5-point rating scale data has not been widely taken up in day-to-day research practice. This lack of uptake may simply be a reflection of the limited availability of appropriate correlation matrices in commonly used software like SPSS or SAS (i.e., the tetrachoric or polychoric correlation matrices). Typically, we would

recommend the use of a polychoric correlation matrix in the factor analysis, however this requires specialized software such as LISREL, MPlus, or EQS.

## 4 Data Analyses of the Simulation Results

### 4.1 The Individual Decision-Making Rules and Indices

In order to analyze the performance of the individual nine decision-making rules and indices for both strict and essential unidimensional measures, an accuracy index was computed for each of the nine decision-making rules and indices, for each case and each replication. This index coded each of the nine decision-making rules and indices as either being accurate (i.e., correctly identified a unidimensional model) or inaccurate (i.e., did not select a unidimensional model) for each case. An overall accuracy rate or proportion (i.e., how often a decision-making rule or index was correct in identifying unidimensionality) ranging from 0.00 to 1.00 was generated for each of the nine decision-making rules and indices. Descriptive statistics of the accuracy index in terms of cell means were generated in order to describe performance according to the various conditions. For example, an accuracy rate for each of the decision-making rules and indices was computed for a strict unidimensional measure with a skewness of 2.50, sample size of 100, and a magnitude of communality as 0.20.

### 4.2 How the Simulation Design Factors Affect the Individual Decision-Making Rules and Indices

We were interested in investigating the main effects of the independent variables on the performance of the individual rules and indices. Such variables included sample size, levels of skewness, and the magnitude of communality for both strict and essential unidimensional measures. The proportions of communality on the second factor and the number of items with non-zero loadings on the second factor were also examined for essential unidimensional measures. The main effects were explored via binary logistic regression for both strict and essential unidimensionality. The Wald chi-square test from the binary logistic regression was analyzed. A binary logistic regression was run separately for each of the dependent variables (i.e., nine individual decision-making rules and indices). The independent variables consisted of the corresponding conditions for strict and essential unidimensionality.

### 4.3 Did Any of the Nine Decision-making Rules and Indices Perform Best?

We were interested in whether any one of the nine decision-making rules or indices performed best, in terms of detecting unidimensionality, in *all* combinations of conditions explored. We examined the optimal performances of the decision-making rule(s) under specific conditions. This entailed, for example, determining which decision-making rule(s) performed best when the distribution of test items was skewed, communalities were low, and sample size was large. A list of possible combinations was constructed. In other words, if the decision-making rules and indices were considered optimal for certain conditions, they were added to the list of possible combinations.

Several criteria needed to be determined in order to approach an understanding of which rule or index performed best. For example, the definition of "best" and "optimal" needed

to be identified. An appropriate definition for *best* and *optimal* in this context was to use a *criterion-referenced* interpretation. This type of interpretation makes no direct reference to other decision-making rules or indices. Instead, rules and indices are evaluated for their proportion of correct responses. In other words, how often a decision-making rule or index was correct in identifying unidimensionality was used to define *best* or *optimal*. An accuracy rate or *proportion of correct responses* was calculated for each of the decision-making rules and indices. A minimum criterion or cut-off value for this accuracy rate was selected in order to define *best* and *optimal*.

## 4.4 Criteria for Best and Optimal Performance

Currently, there are no objectively defined or accepted rules that state how well a decision-making index needs to perform in order to be considered *best* or *optimal* (i.e., successful). A decision-making rule or index was considered *best* if it exceeded a 50% criterion in *all* cell conditions of the design. A decision-making rule or index was considered *optimal* if it exceeded the 50% criterion for specific conditions of the cell design. However, in order to raise the standards for the application of combinations, a second criterion was developed.

Taking from Cohen's (1988) statistical power criterion, the individual decision-making rules and indices were required to meet an 80% criterion in order to be considered for a combination. As mentioned previously, a list of possible combinations was constructed from the results. In other words, once individual rules and indices met the first criterion (i.e., exceeded 50%), they were assessed in order to determine if the 80% criterion was met. If the 80% criterion was met, the rule or index was included in a combination.

Cohen (1988) developed a statistical rule of thumb for the behavioral sciences, which states that a minimum power of 80% is recommended for a research study. The statistical power of a test is the long-term probability of rejecting the null hypothesis ($H_o$) given a specified alpha criterion ($\alpha$), sample size ($N$), and measure of effect size ($f^2$). Although this study is not utilizing statistical power per se, the same 80% criterion was used to determine whether decision-making rules and indices were optimal for combinations. The 80% criterion raised the standards for what is considered a successful decision-making rule or index. In order for a rule to be considered successful, the rule should perform significantly better than 50%.

## 4.5 Did A Combination of the Nine Decision-making Rules and Indices Perform Best?

Given recommendations in the methodological literature to use combinations of rules and indices, we were interested in investigating whether a combination of the individual decision-making rules or indices provided a methodology that performed better than any one individual rule or index. In addition, the question of whether any one combination performed *best* in all conditions was also explored. That is, we focused on examining combinations for optimal performance in specific sets of conditions. This entailed, for example, determining which combination(s) performed *optimally* when the distribution of test items was skewed, communalities were low, and sample size was large.

Combinations were formed based on whether the individual decision-making rules or indices met the 80% criterion. In order to test these combinations, overall accuracy rates for the combinations were generated and then assessed. The accuracy rates for the combinations were the proportion correctly identified for unidimensionality for various sets of conditions. In order for a combination to detect unidimensionality correctly, at least one of the decision-making rules and indices in the combination had to identify unidimensionality

accurately. The accuracy rates for the combinations ranged from 0.00 to 1.00. Cell means of the accuracy rates for the combinations were generated. These cell means were used to determine if *best* and *optimal* performance was met.

Taking from the methodology described above, *best* or *optimal* performance of the combinations was determined by utilizing the 80% criterion. Keeping the standards high (i.e., significantly greater than the 50% criterion), a combination was considered *best* if it met the 80% criterion for *all* conditions. Likewise, a combination was considered *optimal* if it met the 80% criterion for specific conditions. For example, a combination for a strict unidimensional measure with a skewness of 2.50, sample size of 100, and a magnitude of communality as 0.20 was considered *optimal* if it met the 80% criterion for these specific conditions. In addition, all the combinations were assessed to determine if any combination performed better than any one individual rule or index. This was done by comparing the accuracy rates of the combinations to the accuracy rates of the individual rules and indices.

## 5 Results and Conclusions

### 5.1 Summary for Strict Unidimensionality

Overall, both PA methods provided the strongest (i.e., highest accuracy rates) and most consistent (across various conditions) performance for strict unidimensionality. There were several conditions in which the chi-square GLS outperformed chi-square ML, and both performed relatively poor for skewed distributions. All three eigenvalue rules performed poorly when magnitude of communality was 0.20 and favorably when magnitude of communality was 0.90. The RMSEA indices were inconsistent. On the whole, when sample size was 100, distributions were skewed, and magnitude of communality was 0.20, all nine decision-making rules generated considerably low accuracy rates, except for Chi-square GLS (0.69).

There were no main effects of the independent variables found on the PA methods, whereas the other seven decision-making rules and indices had several main effects. The eigenvalues-greater-than-one rule and the ratio-of-first-to-second-eigenvalues-greater-than three rule had a main effect of sample size only. The ratio-of-first-to-second-eigenvalues-greater-than four rule had a main effect of magnitude of communality only.

There was no superior or *best* decision-making rule or index for *all* conditions of strict unidimensionality. There were several optimal decision-making rules and indices for various sets of conditions, and these were formulated into combinations. The list of new combination rules can be found in Table 1 for Strict Unidimensionality.

There were four different sets of conditions in which the combinations performed better than the nine individual rules and indices. The one set of conditions that posed a problem for individual decision-making rules was when sample size was 100, distributions were skewed, and magnitude of communality was 0.20, as mentioned above. However, three combinations met the 80% criterion in *all* conditions explored, including this one problem-set of conditions.

### 5.2 Summary for Essential Unidimensionality

As seen in strict unidimensional measures, both PA methods provided the strongest and most consistent performance. The Chi-square GLS generated the highest accuracy rates when sample size was 100, distributions were skewed, and magnitude of communality was

**Table 1**  List of new combination rules for strict unidimensionality

| Rule number | Simulation-design conditions | Combined rules and indices |
|---|---|---|
| Rule 1 | $n = 100$, $h^2 = 0.90$, skew = 0.00 | Chi-square GLS, PA continuous, PA Likert, eigenvalues > 1, eigenvalues ratio > 4[a] |
| Rule 2 | $n = 100$, $h^2 = 0.20$, skew = 0.00<br>$n = 450$, $h^2 = 0.20$, skew = 0.00 | Chi-square ML, Chi-square GLS, PA continuous, PA Likert |
| Rule3 | $n = 100$, $h^2 = 0.90$, skew = 2.50<br>$n = 450$, $h^2 = 0.90$, skew = 0.00<br>$n = 450$, $h^2 = 0.90$, skew = 2.50<br>$n = 800$, $h^2 = 0.90$, skew = 2.50 | PA continuous, PA Likert, eigenvalues > 1, eigenvalues ratio > 4[a] |
| Rule 4 | $n = 450$, $h^2 = 0.20$, skew = 2.50 | PA continuous, PA Likert |
| Rule 5 | $n = 800$, $h^2 = 0.90$, skew = 0.00 | PA continuous, PA Likert, eigenvalues > 1, eigenvalues ratio > 4, RMSEA ML, RMSEA GLS[a] |
| Rule 6 | $n = 800$, $h^2 = 0.20$, skew = 0.00 | Chi-square ML, Chi-square GLS, PA continuous, PA Likert, eigenvalues ratio > 3, RMSEA ML, RMSEA GLS |
| Rule 7 | $n = 800$, $h^2 = 0.20$, skew = 2.50 | PA continuous, PA Likert, RMSEA ML, RMSEA GLS |

[a] The ratio of first-to-second eigenvalues-greater-than three rule also met the 80% criterion, but it was not included in the combination because if the ratio of first-to-second eigenvalues-greater-than four was successful, then so would the eigenvalues-greater-than three rule. This rule was excluded because of redundancy

0.20. All three eigenvalue rules performed generally poorly when magnitude of communality was 0.20 and favorably when magnitude of communality was 0.90, and the RMSEA indices were inconsistent, which was the same results found with strict unidimensionality. Likewise, on the whole, when sample size was 100, distributions were skewed, and magnitude of communality was 0.20, all nine decision-making rules generated considerably low accuracy rates, except for Chi-square GLS.

All nine decision-making rules and indices had main effects on all conditions, except for the ratio-of-first-to-second-eigenvalues-greater-than three rule and the ratio-of-first-to-second-eigenvalues-greater-than four rule. The ratio-of-first-to-second-eigenvalues-greater-than three rule had main effects of proportion of communality on second factor and number of items loading on second factor, and the ratio-of-first-to-second-eigenvalues-greater-than four rule had main effects of sample size, skewness and proportion of communality on second factor. Because the number of cells was larger for the essential unidimensional investigation, more replications were conducted, and therefore the overall sample size was much larger for essential unidimensional models ($n = 7200$). Consequently, even the smallest of differences may be generating statistically significant main effects of the conditions.

Both rules for the ratio of the first-to-second-eigenvalues (i.e., greater than three and four) had the highest accuracy rates among the nine decision-making rules for the proportion of communality on the second factor as 0.50 when distributions were skewed, magnitude of communality was 0.90 (regardless of sample size and the number of items loading on the second factor). Similarly, the RMSEA indices generated the highest accuracy rates among the nine decision-making rules for the proportion of communality on the second factor as 0.50 when distributions were skewed, magnitude of communality was

**Table 2** List of new combination rules for essential unidimensionality

| Rule number | Simulation-design conditions | Combined rules and indices |
|---|---|---|
| Rule 1 | $n = 100$, $h^2 = 0.90$, skew = 2.50<br>$n = 450$, $h^2 = 0.90$, skew = 0.00<br>$n = 450$, $h^2 = 0.90$, skew = 2.50<br>$n = 800$, $h^2 = 0.90$, skew = 2.50 | PA continuous, PA Likert, eigenvalues > 1, eigenvalues ratio > 4[a] |
| Rule 2 | $n = 800$, $h^2 = 0.90$, skew = 0.00 | PA continuous, PA Likert, eigenvalues > 1, eigenvalues ratio > 4, RMSEA ML, RMSEA GLS[a] |
| Rule 3 | $n = 800$, $h^2 = 0.20$, skew = 2.50 | PA continuous, PA Likert, RMSEA ML, RMSEA GLS |
| Rule 4 | $n = 100$, $h^2 = 0.20$, skew = 0.00<br>$n = 450$, $h^2 = 0.90$, skew = 2.00 | Chi-square ML, Chi-square GLS, PA continuous, PA Likert, |
| Rule 5 | $n = 450$, $h^2 = 0.20$, skew = 0.00<br>$n = 800$, $h^2 = 0.90$, skew = 0.00 | Chi-square ML, Chi-square GLS, PA continuous, PA Likert, RMSEA ML, RMSEA GLS |
| Rule 6 | $n = 100$, $h^2 = 0.90$, skew = 0.00 | Chi-square GLS, eigenvalues > 1, eigenvalues ratio > 4, PA continuous, PA Likert[a] |

[a] The ratio of first-to-second eigenvalues-greater-than three rule also met the 80% criterion, but it was not included in the combination because if the ratio of first-to-second eigenvalues-greater-than four was successful, then so would the greater-than three rule. This rule was excluded because of redundancy

0.20, and sample size was 800. There was no superior or *best* decision-making rule or index for *all* conditions of essential unidimensionality.

## 5.3 Combination Rules

There were several optimal decision-making rules and indices for various sets of conditions, and these were formulated into combinations. The list of new combination rules can be found in Table 2 for Essential Unidimensionality. Rule 4, Rule 5 and Rule 6 (i.e., new rules) met the 80% criterion in *all* sets of conditions explored. Therefore, there were three *superior* or *best* combinations for essential unidimensionality. There were four different sets of conditions in which the combinations performed better than the nine individual rules and indices. When sample size was 100, distributions were skewed, and magnitude of communality was 0.20, all individual decision-making methods failed to meet unidimensionality. However, there were three *superior* or *optimal* combinations (i.e., new rules), which meant that the 80% criterion was met in all sets of conditions investigated, including this problem-set of conditions.

We will now discuss the Chi-square, eigenvalue, parallel analysis (PA), and RMSEA methods in turn.

## 5.4 Chi-square Statistic

Fabrigar et al. (1999) and others claimed that the chi-square statistic was extremely sensitive to sample size. This was partially true in the current simulation. There was a main

effect of sample size on Chi-square GLS, but not chi-square ML, for strict unidimensionality. Both ML and GLS Chi-square tests had main effects of sample size for the essential unidimensional investigation.

## 5.5 Eigenvalue Rule

It is stated in the literature that the eigenvalues-greater-than-one rule continuously performs poorly, especially for small sample sizes. This assumption proved to be partially true in this study. In the current simulation study, for strict unidimensional measures, this rule seemed to perform inadequately when sample sizes were small *and* when communalities were low. When communalities were high, however, even for small sample sizes, the eigenvalues-greater-than-one rule had a 100% accuracy rate. Likewise, when assessing essential unidimensionality, for small sample size and low communalities, the eigenvalues-greater-than-one rule performed extremely poorly (accuracy rates of 0.00), but when communalities increased to 0.90, this rule met the 80% criterion in all cases, except when the proportion of communality on the second factor was 0.50.

This rule was previously found to be most effective when sample sizes were large (Gorsuch 1983), as was the case with the current findings. Hattie (1984) found the eigenvalues-greater-than-one rule to overestimate the number of factors for unidimensional cases. Fabrigar et al. (1999) reported that there had been no study that found the eigenvalues-greater-than-one rule to work. However, the results of this simulation show that this rule actually performs quite well under certain sets of conditions. Researchers continue to widely apply this rule. It seems appropriate at this point to assess whether researchers are actually using this rule under appropriate sets of conditions (see guidelines below).

## 5.6 PA Methods

For both strict and essential unidimensionality, both PA methods provided the strongest accuracy rates overall. The PA methods did not perform well with sample sizes of $n = 100$. Likewise, Crawford and Koopman (1973) found that sample size and different factoring methods influenced the effectiveness of PA methods. Overall, however, this method was highly recommended in the literature for making decisions about the number of factors to retain.

## 5.7 RMSEA Indices

For both strict and essential unidimensional measures, RMSEA indices provided erratic results. However, for essential unidimensional measures, the RMSEA values generated the highest accuracy rates among the nine decision-making rules for the proportion of communality on the second factor as 0.50 when distributions were skewed, magnitude of communality was 0.20, and sample size was 800. In fact, overall, for non-skewed distributions with large sample sizes and communalities of 0.20, these indices performed quite well. Bryne (1998) found the cut points for the RMSEA index to be unreliable, but according to Browne and Cudeck (1992), RMSEA was a promising approach. Fabrigar et al. (1999) recommended using RMSEA, but also pointed-out that the performance of this index lacked empirical evidence.

## 6 Contributions

The development of a procedural definition of essential unidimensionality was a unique contribution to the research literature. Essential unidimensionality is defined conceptually as one dominant factor with the inclusion of an underlying secondary minor factor(s). As introduced in the methodology, essential unidimensionality was technically defined as the simultaneous manipulation of (1) the magnitude of communality, (2) the proportion of communality on the second factor, and (3) the number of items with non-zero loadings on the second (minor) factor.

As mentioned above, the overall objective of the present study was to provide guidelines to assist researchers. The findings from this study were used to develop a preliminary decision-making methodology for determining unidimensionality (i.e., a set of guidelines). It is vital for researchers in today's technologically-based society, where a variety of data analyses can be conducted quickly and efficiently, to be able to make decisions with confidence in regards to retaining factors and potentially using multiple criteria. By using the outcomes (i.e., guidelines) from this study, researchers can assist with making appropriate inferences and high-stake decisions in policy, education and health care when interpreting the results of measures that consist of item response data.

## 7 Guidelines

These guidelines will serve as advice for the social and behavioral science researchers in the decision-making process of retaining factors in an assessment of unidimensionality. These guidelines are preliminary in that future research is required to further investigate how these guidelines perform using multiple data-sets. This advice is guided by what researchers actually have knowledge of before making decisions. For example, sample size, magnitude of communality, and skewness of the distribution of items is information that a researcher can actually attain, whereas information on population data is not necessarily information acquired by day-to-day researchers. Guidelines are provided for both strict and essential unidimensional measures. The guidelines provide recommendations based on certain values of conditions (e.g., sample sizes of 100, 450, and 800), and researchers will need to determine how closely their sample data approximate the conditions represented in these guidelines.

In summary, the new combination rules provided extremely high accuracy rates, and in most cases performed better or just as well as the individual rules and indices. For that reason, it is recommended to use one or more of the new combination rules that are outlined in Tables 1 and 2 for determining both strict and essential unidimensionality. A set of guidelines is provided below for researchers who prefer using individual rules and indices (see Tables 4, 6). In addition, as illustrated in Tables 3 and 5, even though a researcher may prefer the use of individual rules, there are several sets of conditions in which a new combination rule should be used instead of an individual rule or index. This is due to the combinations performing significantly better than the individual rules for those particular sets of conditions.

### 7.1 Strict Unidimensionality

There was no superior or *best* individual decision-making rule or index for *all* sets of conditions of strict unidimensionality. There were several optimal decision-making rules

**Table 3** Recommended combination rules for strict unidimensionality

| Recommendation | Simulation-design conditions | New combined rules |
|---|---|---|
| 1 | $n = 100$, $h^2 = 0.20$, skew $= 0.00$ | Rule 1, Rule 2, and Rule 6 |
| 2 | $n = 100$, $h^2 = 0.20$, skew $= 2.50$ | Rule 1 and Rule 2 |
| 3 | $n = 450$, $h^2 = 0.20$, skew $= 2.50$ | Any of the new seven rules |
| 4 | $n = 800$, $h^2 = 0.20$, skew $= 2.50$ | Rule 4, Rule 5, Rule 6, and Rule 7 |

Refer to Table 1 in order to view the rules and indices of the specific combined rules outlined in this table

**Table 4** Recommended individual rules for strict unidimensionality

| Recommendation | Simulation-design conditions | Individual rules |
|---|---|---|
| 5 | $n = 100$, $h^2 = 0.90$, skew $= 0.00$<br>$n = 100$, $h^2 = 0.90$, skew $= 2.50$<br>$n = 450$, $h^2 = 0.90$, skew $= 0.00$<br>$n = 450$, $h^2 = 0.90$, skew $= 2.50$<br>$n = 800$, $h^2 = 0.90$, skew $= 2.50$ | Eigenvalues $> 1$, eigenvalues ratio $> 3$, eigenvalues ratio $> 4$, PA continuous, OR PA Likert |
| 6 | $n = 450$, $h^2 = 0.20$, skew $= 0.00$ | PA continuous OR PA Likert |
| 7 | $n = 800$, $h^2 = 0.20$, skew $= 0.00$ | PA continuous, PA Likert, ML RMSEA, OR GLS RMSEA |
| 8 | $n = 800$, $h^2 = 0.90$, skew $= 0.00$ | Eigenvalues $> 1$, eigenvalues ratio $> 3$, eigenvalues ratio $> 4$, PA continuous, PA Likert, ML RMSEA, OR GLS RMSEA |

and indices for various sets of conditions, and these were formulated into combinations. There were four different sets of conditions in which the combinations performed better than the nine individual rules and indices. Therefore, it is recommended that the following combination rule(s) be applied for the following sets of conditions, as shown in Table 3.

In the remaining sets of conditions, the individual rules and indices performed just as well as the combinations. Therefore, if a researcher prefers to use individual methods, it is recommended that the following individual rules and indices be used for the specified set of conditions.

Again, there were three combination rules (Rule 1, Rule 2, Rule 6) that were considered superior, and hence could be applied to all sets of conditions. However, if a researcher prefers to utilize one method, recommendations five through eight in Table 4 are deemed appropriate.

### 7.2 Essential Unidimensionality

There was no superior or *best* individual decision-making rule or index for *all* sets of conditions of essential unidimensionality. There were numerous optimal decision-making rules and indices for various sets of conditions, and these were formulated into combinations. There were four different sets of conditions in which the combinations performed better than the nine individual rules and indices. Therefore, it is recommended that the

**Table 5** Recommended combination rules for essential unidimensionality

| Recommendation | Simulation-design conditions | New combined rules |
|---|---|---|
| 1 | $n = 100$, $h^2 = 0.20$, skew $= 0.00$<br>$n = 100$, $h^2 = 0.20$, skew $= 2.50$ | Rule 4, Rule 5, and Rule 6 |
| 2 | $n = 450$, $h^2 = 0.20$, skew $= 0.00$ | Rule 4, Rule 5, and Rule 6[a] |
| 3 | $n = 450$, $h^2 = 0.20$, skew $= 2.50$ | Any of the new six rules |

Refer to Table 2 in order to view the rules and indices of the specific combined rules outlined in this table

[a] In addition, when the proportion of communality on the second factor was 0.05, both PA methods individually performed just as well

**Table 6** Recommended individual rules for essential unidimensionality

| Recommendation | Simulation-design conditions | Individual rules |
|---|---|---|
| 4 | $n = 100$, $h^2 = 0.90$, skew $= 0.00$<br>$n = 100$, $h^2 = 0.90$, skew $= 2.50$<br>$n = 450$, $h^2 = 0.90$, skew $= 0.00$<br>$n = 450$, $h^2 = 0.90$, skew $= 2.50$<br>$n = 800$, $h^2 = 0.90$, skew $= 0.00$<br>$n = 800$, $h^2 = 0.90$, skew $= 2.50$ | Eigenvalues $> 1$, eigenvalues ratio $> 3$, eigenvalues ratio $> 4$, PA continuous, OR PA Likert[a] |
| 5 | $n = 800$, $h^2 = 0.20$, skew $= 0.00$ | PA continuous, PA Likert, ML RMSEA, OR GLS RMSEA[b] |
| 6 | $n = 800$, $h^2 = 0.20$, skew $= 2.50$ | RMSEA GLS[c] |

[a] When the proportion of communality on the second factor is 0.50, just the rules for the ratio of first-to-second-eigenvalues-greater than thee and four are recommended

[b] When the proportion of communality on the second factor is 0.50, just the RMSEA indices are recommended

[c] When the proportion of communality on the second factor is 0.50, both RMSEA indices are recommended

following combination rule(s) be applied for the specified sets of conditions, as shown in Table 5.

In the remaining sets of conditions, the individual rules and indices performed just as well as the combinations. Therefore, if a researcher prefers to use individual methods, it is recommended that the following individual rules and indices, as shown in Table 6, be used for the specified set of conditions for essential unidimensional measures.

Again, there were three combination rules (Rule 4, Rule 5, Rule 6) that were considered superior, and hence could be applied to all sets of conditions. However, if a researcher would prefer to utilize one method, recommendations four through six in Table 6 are deemed appropriate.

# References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.

Aron, A., & Aron, E. (2002). *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Boyd, K. C., & Gorsuch, R. L. (2003). *Factor replication, factor invariance, and salient loadings: Three objective criteria for number of factors*. Manuscript submitted for publication.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*, 230–258.

Bryne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crawford, C. B. (1975). Determining the number of interpretable factors. *Psychological Bulletin, 82*, 226–237.

Crawford, C. B., & Koopman, P. (1973). A note on Horn's test for the number of factors in factor in analysis. *Multivariate Behavioral Research, 8*, 117–125.

Davison, M. L., & Sireci, S. G. (2000). Multidimensional scaling. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematics modeling* (pp. 323–352). New York: Academic Press.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327–346.

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*, 272–299.

Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying responses to a set of items. *Journal of Educational Measurement, 33*, 157–179.

Glenberg, A. (1996). *Learning from data: An introduction to statistical reasoning*. New York, NY: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1–10.

Hattie, J. (1984). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 20*, 1–14.

Hattie, J. (1985). An empirical study of the various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123*(3), 207–215.

Humphreys, L. G. (1952). Individual differences. *Annual Review of Psychology, 3*, 131–150.

Humphreys, L. G. (1962). The organization of human abilities. *American Psychologist, 17*, 475–483.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education and National Council on Measurement in Education.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Charlotte: Information Age Publishing.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*(4), 611–637.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, CA: Sage.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics, 32*(2), 13–43.

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin, 28*(2), 1629–1646.

Schonemann, P. H. (1981). Power as a function of communality in factor analysis. *Bulletin of the Psychonomic Society, 17*, 57–60.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte: Information Age Publishing.

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction With Life Scale (SWLS). *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement, 92*, 489–496.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins College Publishers.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 213–224.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). The Netherlands: Elsevier Science B.V.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65–82). Charlotte: Information Age Publishing.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The construction and use of psychological tests and measures. In the Psychology theme of the *Encyclopedia of Life Support Systems (EOLSS)*. Oxford, UK: Eolss Publishers. http://www.educ.ubc.ca/faculty/zumbo/talks/Zumbo_Gelin_Hubley_EOLSS.pdf.