The Number of Factors Problem

Timmerman, Marieke E.; Lorenzo-Seva, Urbano; Ceulemans, Eva

Published in:
The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development

DOI:
10.1002/9781118489772.ch11

11

The Number of Factors Problem

Marieke E. Timmerman, Urbano Lorenzo-Seva, and Eva Ceulemans

Introduction

Psychological assessment relies heavily on the use of measurement instruments. Those instruments commonly take the form of questionnaires or scales. Typically, such a scale is designed to measure one or more psychological constructs. In the process of scale evaluation, a common step is to collect item responses in a sample of respondents, and model them using factor analysis or item response analysis. The key idea in both modelling approaches is that the model captures the dimensional structure underlying the observed item scores, and that the dimensional structure can be meaningfully interpreted in terms of the psychological constructs to be measured. The dimensions are generally denoted as factors in factor analysis, and as latent traits in item response analysis. Both approaches assume that individuals' scores on the factors (or the latent traits in item response analysis) are due to individuals' scores on the corresponding latent variables.

The dimensional structure involves two aspects, namely the number of dimensions needed to adequately model the item responses, and the mathematical relationship between the individual position on the dimension(s) and the item scores. The form of the mathematical relationship depends on the model considered. For example, the factor model specifies a linear relationship. Furthermore, it is important to note that the number of dimensions needed for a given set of item responses may depend on the specific model considered, as will become clear next.

To assess the dimensional structure underlying a set of item responses, one could proceed in an exploratory or a confirmatory way. In both ways, the target model to consider, like a common factor model, is typically selected beforehand. The exploratory approach is taken when clear hypotheses about the structure are lacking. Also, in the presence of a hypothesized structure, the exploratory approach is sometimes used as a strong test of the hypothesized structure. However, this interpretation warrants some caution, because the results of an exploratory analysis fully depend on the criterion underlying the analysis. If this criterion does not match the hypothesized structure, the latter will not be retrieved, regardless whether it is present or not in the data. Taking

The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, First Edition. Edited by Paul Irwing, Tom Booth, and David J. Hughes.

the confirmatory approach boils down to testing a specific model, in which each item is only associated with a particular (and often a single) dimension.

The exploratory approach typically involves a multiple step procedure, in which both formal criteria and substantive considerations play a role. First, using some formal criterion, one obtains an indication of the dimensionality; that is, the number of dimensions needed. Some criteria require a series of target models with increasing dimensionality to be fitted, whereas others are based on measures derived from the observed data. Second, the target model with the indicated dimensionality is fitted to the data. Obviously, when the selected model is among the fitted target models in the first step, one can select it from the fitted models. Third, the interpretability of the fitted model is judged. In case of difficulties in interpretation, it is common to consider alternative models with a different number of dimensions as well.

In this chapter, we focus on formal criteria to assess the dimensionality for exploratory factor modelling with the aim to facilitate the selection of a proper criterion in empirical practice. We first introduce the different foundations that underlie the various criteria. Then, we elaborate on the meaning of an indicated dimensionality. Subsequently, we provide an overview of currently available formal criteria, which we selected on the basis of their popularity in empirical practice and/or proven effectiveness. Further, we provide guidance in selecting a dimensionality assessment method in empirical practice. We illustrate the use of the methods with an empirical example.

Categorizing Criteria to Indicate the Number of Factors

To understand the properties of the criteria to indicate the number of factors, and hence to make a well-founded choice in empirical practice, it is useful to categorize the criteria. A first categorization is that the criteria to indicate the number of factors are associated with either principal component analysis (PCA), common factor analysis (CFA), or CFA for ordinal variables (CFA-ord) (see e.g., Mislevy, 1986). In case of PCA, it would be more appropriate to use the term **component** rather than **factor**, but for ease of presentation, we use the term "factor" in the context of PCA as well. PCA and CFA are suitable for modelling continuous item responses, and CFA-ord for polytomous, ordered item responses.

From a theoretical point of view, it is best to have a match between the type of criterion to assess the number of factors and the specific model subsequently used to fit the data. However, in empirical practice, one often finds a discrepancy; most often when the number of common factors (for a CFA or CFA-ord) is being indicated using a PCA-based criterion. This practice is not necessarily problematic, because the number of factors indicated by a nonmatching method often equals that of a matching method. The reason for the equality in dimensionality suggested is that performing PCA, CFA, and CFA-ord on the same empirical data set can yield similar results, especially with respect to the statistic (e.g., fit value) used in the number of factors criterion. However, differences in the indicated number of factors may occur. They are to be expected particularly when the PCA, CFA, and CFA-ord would reveal substantially different results. This may occur when a PCA- or CFA-based method rather than a CFA-ord based method is applied to ordinal items of which (1) the number of categories is five or smaller

(e.g., Dolan, 1994), (2) the univariate distributions of items are asymmetric or with excess of kurtosis (Muthén & Kaplan, 1985, 1992), or (3) when a PCA-based method is used to indicate the number of common factors when the loadings are low (and hence the unique variances high) (e.g., Widaman, 1993) and there are few variables per factor (e.g., Velicer & Jackson, 1990).

A second categorization is according to the definition underlying the criterion. We distinguish statistical test and major factors criteria. Statistical test criteria are based on a strict definition of dimensionality, as the minimum number of latent variables needed to completely describe the scores in a population (Zhang & Stout, 1999). Though this definition is well-founded from a mathematical point of view, a strict adherence to this definition appears of limited use in empirical practice.

Indeed, when applied to empirical data, a statistical test criterion may yield as many latent variables as there are items in the scale. The reason is that such criteria presume the model considered to hold exactly at the population level, implying that no model error is involved (see MacCallum & Tucker, 1991). This contrasts to the view that a model is an approximation to reality (Browne & Cudeck, 1992), a view that appears to be generally accepted nowadays.

In the current context, the absence of model error would imply that the relationship between the latent variables and the items (e.g., linear in the case of a CFA) holds exactly. Even if this would be the case, one would expect the need for many latent variables. The reason is that items within a scale typically show diversity in item content. This diversity can only be captured fully by a model with a large dimensionality. To solve for this, it might seem attractive to limit the diversity in item content. However, the price to be paid is that it results in a measure of conceptually narrow constructs (Cattell & Tsujioka, 1964). This has led to the view that any scale with a reasonable degree of substantive breadth has item diversity (Reise, Waller, & Comrey, 2000). Typically, all items in such a scale share a single (or a few) major dimension(s), whereas small subsets of items share minor parts that are independent of the remaining observed items. Within a factor analysis, the associated major factors are thought to account for a large part, and the minor factors for a small part of the observed variance. This property enables major factors to be identified, and it forms the basis of major factor based criteria.

In empirical practice, statistical test criteria are useful to identify the maximal number of factor to use. Depending on the research aims and the data characteristics, it can be appropriate to consider the major factors only, or the major and a few minor factors. In the latter case, one would end up with a number of factors that are in between the numbers of major and total factors.

The Meaning of Dimensionality

Being the number of dimensions needed to adequately model the item responses, the dimensionality defines the number of axes that span the space in which the subjects are projected. In this low-dimensional space, each subject and each item in the data set can be depicted. Their specific positions are estimated in the analysis. In an exploratory analysis, the low-dimensional space is typically estimated such that the variability in the subject positions in the low-dimensional space is as large as possible. This implies

that the dimensionality for a given data set crucially depends on both the subjects and the items involved (see Reckase, 2009). That is, it depends on the number of dimensions on which the subjects differ, and on the number of dimensions to which the items are sensitive. Typically, the subjects are considered to be a sample from a population. This leads to the important notion that the dimensionality may be population dependent. Furthermore, to detect the dimensions on which the subjects differ, the items included should be sensitive to those dimensions, whereas the sensitivity to the various dimensions should vary across items (Reckase, 2009). Ideally, the dimensions have a substantive meaning, and thus can be interpreted from an explanatory-theory perspective.

For example, suppose that we are interested in levels of anxiety and depression, and that we use a set of items to measure the degrees of anxiety and depression in a sample. We can sensibly distinguish between the two if and only if (1) the subjects in our sample show variability in the degrees of both anxiety and depression, (2) the degrees of anxiety and depression are at most moderately related, and (3) the items are sensitive to anxiety and depression to varying degrees. In empirical practice, the latter requirement is usually satisfied when the items are designed with the aim to tap only a single substantive dimension of interest (i.e., anxiety or depression in this example). It is important to recognize that the other two requirements may not be satisfied; for example, because subjects under study differ considerably in their degree of anxiety, but are all similar in their degree of depression, or because in the sample under study the degrees of anxiety and depression are highly correlated. This would typically yield a single indicated major dimension, which does not match the substantive dimensions of depression and anxiety.

To interpret the dimensional structure, one considers the position of the items in the low-dimensional space. Items in a similar direction (i.e., when depicted as vectors from the origin) are interpreted as being sensitive to the same dimension. The position is expressed in a factor model via the loadings. To ease the interpretation, the axes are typically rotated to so-called simple structure, implying that the axes are positioned such that the direction of each item is as close as possible to only a single axis. Kaiser (1974) denoted this as factor simplicity. The meaning of each rotated axis is derived from the common content of the items associated with that axis. Ideally, each rotated axis matches a substantive dimension.

For example, reconsidering the set of items indicative of the degrees of anxiety and depression, the items may be designed so that two subsets of items are relatively pure indicators of anxiety and depression, respectively, and the remaining items indicate a combination of both. Then, it is useful to position the two axes such that their directions coincide as much as possible with the two subsets of "pure" anxiety and depression items. In this approach, the construct of interest is defined beforehand, and the causal relationship between the levels of depression and anxiety of a given person and his/her responses to the items is reasonable to assume.

In some applications, the dimensions are identified on the basis of items solely, rather than that constructs of interest are defined beforehand. Then, one strictly needs an exploratory analysis approach. A typical example is found in personality research, since the theories on distinguishing personality traits have not been settled yet. For example, the Big Five personality traits have found large adherence, but claims for traits beyond the Big Five have been made (De Raad & Barelds, 2008). In such cases, the dimensional structure of the items under study is the key issue to settle.

Dimensionality Assessment Methods

We now successively review PCA-based methods and CFA-based methods to assess the number of common factors. For each method, we outline the rationale behind the method. We summarize the findings on its quality to indicate the number of (total and/or major) common factors. An overview of the different methods, categorized according to the associated model and the type of criterion, is provided in Table 11.1.

Though from a theoretical point of view PCA-based methods are inappropriate in a CFA context, their fundamental differences are often hidden in empirical practice, because they may yield very similar results. Furthermore, PCA-based methods appear to be even more popular than their CFA-based counterparts to indicate the number of factors in empirical practice (ten Holt, van Duijn, & Boomsma, 2010).

Kaiser Criterion

The Kaiser criterion (Kaiser, 1960) aims at indicating the number of principal components (PCs) of a correlation matrix, and can be viewed as a statistical criterion. The basic idea originates from the observation that if all items involved would be independent, the number of PCs should equal the number of items, and each PC has a variance of one at the population level. Because PCs should reflect dependencies among variables, the Kaiser criterion is to retain PCs with a variance larger than one only. Because the variance of a PC equals the associated eigenvalue of a correlation matrix, the Kaiser criterion is also known as the eigenvalue-over-one criterion.

A fundamental problem with the Kaiser criterion is that it is based on properties at the population level only. It is known that the eigenvalues of the correlation matrix resulting from samples of limited size, even if the variables are uncorrelated in the population, are larger than one (Buja & Eyuboglu, 1992; Horn, 1965). This observation explains the finding in many studies that the Kaiser criterion clearly yields inaccurate indications of the number of PCs and common factors, mostly indicating too many factors (e.g., Revelle & Rocklin, 1979; Zwick & Velicer, 1982, 1986). In spite of its poor performance, the Kaiser criterion appears very popular in empirical practice (Costello &

Table 11.1 Overview of the dimensionality assessment methods discussed, categorized according to the associated model and the type of criterion; q is the true number of common factors.

	Model associated with criterion	
	PCA	*CFA/CFA-ord*
Major factor criterion	Scree test	PA-MRFA and variants (for $q > 1$)
	Horn's PA (for $q > 1$)	Goodness-of-fit measures (as RMSEA, CFI)
		Hull
Statistical criterion	Kaiser criterion	PA-MRFA and variants (for $q = 1$)
	MAP	LRT: Chi-square test
	Horn's PA (for $q = 1$)	LRT: Chi-square difference test

Osborne, 2005), probably simply because it is the default criterion to select the number of factors in popular statistical packages.

Scree test

The scree test (Cattell, 1966) involves the visual inspection of a so-called scree-plot. The scree-plot is a plot of the successive eigenvalues ($m = 1,\ldots,n$, with n the number of items) of the observed correlation or covariance matrix versus the number of the associated eigenvalue. Note that the successive eigenvalues are linearly related to the percentage of explained variance of the successive PCs. Cattell proposed to locate an "elbow," which is the point beyond which the scree-plot follows an approximately straight line, and take the first point on that line as the indicated number of factors. In practice, researchers often take the number of factors before the elbow. This makes sense because the PC associated with the elbow itself already explains relatively little variance. Because the scree test is based on variances explained of PCs, it is PCA-based method, though Cattell explicitly aimed at finding the number of common factors. Furthermore, the scree test optimally balances the fit and number of parameters involved, and can therefore viewed as a major factor criterion.

The scree test has been criticized for its subjectivity. One may expect this to be problematic when a clear elbow is missing in the scree-plot. This may particularly occur when the scree-plot shows a gradual slope, or when multiple elbows can be indicated (Jolliffe, 2002). Based on simulated data, the scree test has been found to be reasonably accurate to indicate the number of major factors. Its performance deteriorated with increasing unique variances, making the method less suitable for items with large unique variances (Zwick & Velicer, 1986).

Minimum Average Partial (MAP)

Based on partial correlations, the MAP (Velicer, 1976) is a statistical criterion for indicating the number of PCs. For a sequence of numbers of PCs, $m = 1,\ldots,n$ (with n the number of items), the partial correlations first decrease, and then increase. The partial correlation for a solution with m PCs is computed as the average squared partial correlations between all pairs of items involved, with each of m components partialled out. The MAP indicates the number of components associated with the minimum partial correlation.

Though MAP was proposed in a PC context, it is noteworthy that it employs the concept of common factors. This is so because the solution with the minimum partial correlations is associated with a residual matrix that closely resembles an identity matrix, and for which at least two variables load high on each component involved.

MAP has been examined in comparative simulation studies based on a common factor model (CFM) including both major and minor factors (Tucker, Koopman, & Linn, 1969). MAP performances in indicating the number of major factors deteriorated when the unique variances increased, with no clear tendency to over- or underindicate the number of factors (Lorenzo-Seva, Timmerman, & Kiers, 2011; Zwick & Velicer, 1986). Because MAP performed poorly in empirically relevant conditions (i.e., certain levels of unique variances), Lorenzo-Seva et al. (2011) discourage the use of MAP to indicate the number of common factors.

Horn's Parallel Analysis (Horn's PA)

In its original form, PA (Horn, 1965) is a method to indicate the number of PCs. It is based on the same principle as the Kaiser criterion, with PA also taking sampling fluctuations into account. The central idea that components to retain must be associated with eigenvalues larger than the eigenvalues of components derived from random data, where the random data involve the same sample size as the sample data.

The procedure followed in Horn's PA can be summarized as follows. First, one obtains the sampling distributions of the random eigenvalues; these are the sampling distributions of the successive eigenvalues ($m = 1, \ldots, n$, with n the number of items) of sample correlation matrices for independent items. Second, one obtains the empirical eigenvalues, which are the eigenvalues of the correlation matrix of the observed data. Third, one successively compares each empirical eigenvalue to the sampling distribution of the random eigenvalues at the same position (i.e., the first empirical eigenvalue is compared to the distribution of the first random eigenvalue, the second empirical eigenvalue to the distribution of the second random eigenvalue, etc.). PA indicates to retain all components of which the empirical eigenvalues are larger than a given threshold in the distribution of random eigenvalues. Horn (1965) proposed using the average as the threshold. Because this threshold appeared to yield a tendency to overestimate the dimensionality, the use of a more stringent threshold was advocated, like the 95% quantile (Buja & Eyuboglu, 1992; Glorfeld, 1995). Nowadays, the 95% threshold appears to be most popular.

Though Horn's PA has been interpreted as an inferential method (i.e., a method to assess the significance of each dimension), this interpretation is appropriate only for the first eigenvalue (Buja & Eyuboglu, 1992). The reason is that one faces inherent dependencies between successive eigenvalues, which results in a loss of statistical power for increasing number of components. The size of the power loss is data dependent and cannot be generally predicted. Because successive eigenvalues of a correlation matrix are directly related to the explained variance of successive PCs, the preceding implies that Horn's PA is a statistical method when it comes to indicating the first principal component. For the remaining successive PCs, Horn's PA suffers from loss of statistical power, and can be expected to function as a major factor criterion.

Horn's PA has been extensively studied, on the basis of empirical data (Cota, Longman, Holden, & Fekken, 1993; Hubbard & Allen, 1987) and simulated data (Glorfeld, 1995; Peres-Neto, Jackson, & Somers, 2005; Turner, 1998; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). The performance of Horn's PA appeared to be generally good, resulting in strong recommendations for PA in empirical practice (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Hayton, Allen, & Scarpello, 2004; Thompson, 2004).

PA-MRFA and other variants, for common factors, and for ordinal items

Noting that Horn's PA is not well-founded to assess the number of common factors (CFs), Humphreys and Ilgen (1969) proposed a variant of Horn's PA to assess the number of CFs. In this procedure, the eigenvalues are obtained from the reduced correlation matrix, with estimates of the communalities on its diagonal. This approach suffers from fundamental problems (Buja & Eyuboglu, 1992; Steger, 2006), among which the fact that the empirical eigenvalue and the random eigenvalues lack a common interpretation

(e.g., in terms of explained variances, as in Horn's PA), and therefore cannot be sensibly compared (Timmerman & Lorenzo-Seva, 2011). A better alternative to indicate the number of CFs is found in PA-MRFA (Timmerman & Lorenzo-Seva, 2011), in which the proportions of explained common variance of empirical and random data are compared, rather than the eigenvalues.

Typically, PA is performed on Pearson correlation matrices, making the approach suitable for modelling continuous item responses. In case of polytomous, ordered item responses, it has been proposed to use polychoric correlations (e.g., Olsson, 1979) as a basis for PA (Cho, Li, & Bandalos, 2009; Weng & Cheng, 2005).

The different PA variants have been compared in simulation studies. On the basis of those results, for polytomous, ordered item responses the use of polychoric correlations rather than Pearson correlations seems indicated (Timmerman & Lorenzo-Seva, 2011). Furthermore, PA-MRFA with 95% threshold is the preferred variant for indicating the number of major CFs (Timmerman & Lorenzo-Seva, 2011), with Horn's PA with 95% threshold as a second best, with still acceptable performance. In empirical practice, one may encounter convergence problems for the polychoric correlations. Then, one could resort to Pearson correlations, where PA-MRFA performed best with mean threshold (Timmerman & Lorenzo-Seva, 2011).

Model Selection Methods in SEM: CFM as a Special Case of SEM

The issue of the number of CFs to retain can be viewed as a model selection problem. That is, one aims at selecting the most appropriate model from a series of CFMs, with different numbers of factors. By noting that an exploratory CFM is a special case of a structural equation model, the methods used for model selection in structural equation modelling (SEM) can be readily applied in selecting the number of CFs (Fabrigar et al., 1999). In the course of years, statistical tests and a vast number of goodness-fit-measures have been proposed. In selecting the number of CFs, the general procedure is to fit a sequence of CFMs, with zero up to some maximal number. For each model, one performs the statistical test, or computes the selected goodness-of-fit measure(s), and applies, implicitly or explicitly, a certain criterion involving those measures to indicate the number of factors. To understand why the tests and goodness-of-fit criteria may indicate different number of factors, it is essential to understand the theoretical underpinnings of the statistical tests and the goodness-of-fit measures.

The statistical tests and most goodness-of-fit measures considered in SEM are based on the chi-square measure. The chi-square measure can be obtained using various estimation methods, of which their suitability depends on the nature of the data at hand. For multivariate normally distributed items, maximum likelihood (ML) estimation is appropriate. For continuous items that are not multivariate normally distributed, robust ML (Yuan & Bentler, 2000), which adjusts standard errors and fit indices, is indicated. For ordered polytomous items, different estimation methods, like unweighted least squares (ULS), weighted least squares (WLS) (Browne, 1984) and robust WLS (Muthén, du Toit, & Spisic, 1997), are available (see Wirth & Edwards, 2007, for a nice review). Though from a theoretical point of view the use of those variants is justified only for the type of items indicated, the story is more shaded in empirical practice. First, the consequences of taking an "inappropriate" estimation method depend on the

severity and type of violation and on the parameter of interest. For example, ML estimates, which are suitable for multivariate normally distributed items, appeared to yield reasonable estimates for ordered polytomous items with five or more categories (Dolan, 1994). Second, the applicability of estimation methods depends on the sample size available. WLS requires extremely large sample sizes, whereas ULS and robust WLS (Muthén et al., 1997) work better in a smaller sample (say, minimally 250 (e.g., Beauducel & Herzberg, 2006)). Comparative simulation studies show ULS and robust WLS performing similarly (Yang-Wallentin, Joreskog, & Luo, 2010).

The statistical tests and the many available goodness-of-fit measures share the purpose of expressing the degree to which the model at hand approximates the population correlation or covariance matrix of the items concerned. They differ in their foundations and the specific aspects of lack of model fit stressed. We limit our discussion on goodness-of-fit measures to the currently popular variants. For a review of the goodness-of-fit measures in SEM, we refer to Hu and Bentler (1999) and Hooper, Coughlan, and Mullen (2008). Now, we will specifically discuss various approaches used in SEM for model selection, and examine their usefulness in indicating the number of CFs. We start with statistical tests, followed by fit measures.

Likelihood ratio tests: the chi-square test and the chi-square difference test

Likelihood ratio tests (LRTs) are statistical tests. Those tests involve the population covariance matrix $\mathbf{\Sigma}$ (i.e., the covariance matrix in the population from which the sample at hand has been randomly drawn) and the model implied covariance matrix $\mathbf{\Sigma}_m$ (i.e., here, the covariance matrix associated with the CFM with at most m CFs). To assess the number of factors, two LRT tests (see e.g., Hayashi, Bentler, & Yuan, 2007) have been put forward. Both tests have the null-hypothesis that the population covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}_m$. The chi-square test tests against the saturated model, that is, has the alternative hypothesis that $\mathbf{\Sigma}$ is any positive definite matrix. The chi-square difference test tests against the model with at least $m + 1$ factors, that is, has the alternative hypothesis that $\mathbf{\Sigma} = \mathbf{\Sigma}_{m+1}$. The chi-square test statistic involved is based on the discrepancy between $\mathbf{\Sigma}_m$ and the observed sample covariance matrix. For both LRTs, rejecting the null-hypothesis for a model with m factors indicates that more than m factors are needed.

A fundamental problem with the LRTs is that they are only appropriate for applying to CFMs with m factors if the number of factors in the population does not exceed m. The reason is that the chi-square statistic of a model with more than the number of population factors no longer follows a chi-square distribution, with as a consequence a tendency to indicate too many factors (Hayashi et al., 2007). Therefore, the chi-square difference test to test for the number of factors in the population appears inappropriate, because it involves computing the chi-square statistics for models with both m and $m + 1$ factors. This problem does not hold for the chi-square test itself, because it only uses the chi-square statistic of the model with m factors.

To assess the number of CFs, one may apply a stepwise forward procedure. That is, one starts with a one-factor model and applies the chi-square test. If this test is rejected, one fits a two-factor model, and uses the chi-square test. If this test is rejected, one fits a three-factor model. One proceeds until the m-factor model is not rejected, and selects the model with m factors.

This stepwise forward procedure implies that one uses a significance test to confirm a null-hypothesis. Therefore, any nonreject should be considered with caution, since a nonreject can be due to either a true null-hypothesis or a lack of power to detect differences. Obviously, the latter is more likely to occur in small sample sizes. Thus, in empirical practice, when considering samples from the same population, one would expect the number of indicated factors to decrease with decreasing sample sizes. This dependency on the sample size is indeed a major problem of the use of LRT in empirical practice.

A further objection that has been raised with respect to LRTs (e.g., Bentler & Bonett, 1980) is that in large sample sizes even tiny deviations from the null-hypothesis are likely to result in a reject. "Tiny deviations" can occur for many different reasons, including distributional misspecification (e.g., assuming a normal distribution of the items scores in case of ordered polytomous items) and the presence of minor factors (e.g., factors on which the individuals show little variance). From a theoretical point of view, the sensitivity of the LRT to misspecifications is appropriate. In fact, it is exactly what would be expected from a method founded on the strict definition of dimensionality. From an empirical point of view, some authors considered the behavior of the LRT to be problematic. They argued that the LRT is also sensitive to model misspecifications that pertain to characteristics that are trivial from an explanatory-theory perspective (Barrett, 2007), that is, characteristics that do not match any substantive dimension. To alleviate the problems with the LRT in evaluating models, various goodness-of-fit measures were developed. The use of those measures to indicate the dimensionality can be viewed as a major factor criterion.

Goodness-of-fit measures

The Root Mean Square Error of Approximation (RMSEA) (Steiger, 1990) appears among the most popular goodness-of-fit measures used in SEM. The RMSEA expresses the discrepancy between the population covariance matrix $\mathbf{\Sigma}$ and the model implied covariance matrix $\mathbf{\Sigma}_m$ per degree of freedom of the model. Thus, the RMSEA expresses the fit of the model relative to its complexity, implying that simpler models are favored over more complex models. Furthermore, the RMSEA is a property defined at the population level.

An RMSEA value of zero indicates a perfect model fit at the population level. Recognizing that any model is an approximation to reality (Browne & Cudeck, 1992), models with an RMSEA larger than zero can be acceptable. RMSEA values smaller than .05 (Browne & Cudeck, 1992) or .06 (Hu & Bentler, 1999) are considered to indicate close fit, and values in the range of .05 to .08 fair fit (e.g., MacCallum, Browne, & Sugawara, 1996). By noting that in empirical practice the RMSEA is estimated on the basis of sample data, and hence is influenced by sampling fluctuations, it is useful to consider a (say, 95%-) confidence interval for the RMSEA (Browne & Cudeck, 1992), rather than its point estimate.

Another popular fit index is the Standardized Root Mean square Residual (SRMR), which summarizes the differences between the sample covariances and the model implied covariances. Unlike the RMSEA, model complexity does not play a role in the SRMR. Furthermore, the SRMR only expresses the fit to the current sample data (rather than to the population, as the RMSEA). An SRMR value of zero indicates

perfect fit. SRMR values of .05 (Sivo, Fan, Witta, & Willse, 2006) or .08 (Hu & Bentler, 1999) are recommended as cut-off values for model selection.

The comparative fit index (CFI) is a measure of improvement of fit, where the model at hand is compared to a nested baseline model. Typically, one uses as a baseline model a model specifying completely uncorrelated items. CFI values are in the range from 0 to 1, with 1 indicating a perfect fit. A value > .95 is considered to indicate good fit (Hu & Bentler, 1999).

Model selection on the basis of goodness-of-fit criteria typically takes place by selecting those models with associated goodness-of-fit criteria below the recommended cut-off values, which were determined on the basis of extensive simulation studies in a SEM context. However, a strict adherence to cut-off values appears inappropriate. On the one hand, strictly applying the cut-off values may yield the selection of unacceptably misspecified models, depending on the specific characteristics of the data (e.g., Marsh, Hau, & Wen, 2004). This is so because also small degrees of misspecification may pertain to aspects that are important from an explanatory-theory perspective, that is, match a substantive dimension. On the other hand, strictly applying the cut-off values may also result in including minor factors, which are unimportant from an explanatory-theory perspective.

As an alternative to strict cut-off values, one may apply the goodness-of-fit values as relative measures (Yuan, 2005). In the exploratory common factor context, this would imply considering goodness-of-fit measures for a series of CFMs with sequentially increasing numbers of factors. Fabrigar et al. (1999, p. 279) suggested to select the model "…which constitutes a substantial improvement in fit over a model with one fewer factor but for which a model with one more factor provides little if any improvement in fit." This suggestion, which is based on the same principle as the scree test, requires a subjective decision on what constitutes a substantial improvement in fit. A method that formalizes this step is the Hull method.

Hull method

The Hull method (Lorenzo-Seva et al., 2011) aims at indicating the number of major, CFs. The Hull method is based on the same principle as the scree test, namely to optimally balance the fit and number of parameters involved. The optimal balance is found using a numerical convex hull-based heuristic (Ceulemans & Kiers, 2006). This heuristic makes a visual inspection, as in the scree test, superfluous. Further, the Hull method limits the possible maximum number of factors to the number of factors indicated by Horn's PA, to guard against overfactoring. The fit of a CFA model can, depending on the estimation method, be expressed in different ways. Based on the results of a comparative simulation study (Lorenzo-Seva et al., 2011), including the CFI, RMSEA, and SRMR as fit measures, it is advised to use the CFI (Bentler, 1990) as a fit measure when chi-square statistics are available, yielding Hull-CFI. When chi-square statistics are unavailable, such as when Principal Axis Factoring is used as an estimation method, it is advised to use as a fit measure the Common part Accounted For (CAF), which expresses the amount of variance of the common part explained by the factor model at hand.

In a comparative simulation study, Hull-CFI and Hull-CAF performed generally very well in indicating the number of major factors, and on average outperformed

alternatives as PA and MAP. The performance of the Hull variants seems to improve as the sample size increases, and the number of observed variables per factor increases. As a consequence, Hull is a suitable approach when large datasets are analyzed: with this kind of dataset, it outperforms such classical methods as PA. However, PA outperformed both Hull variants in case of small sample sizes ($N = 100$) and small numbers of items (i.e., 5) per factor.

Assessment of the Number of Factors in Empirical Practice

To assess the number of factors underlying an empirical data set, we suggest the following strategy. First, one should decide about the most appropriate model for the data at hand. In this chapter, we considered the models associated with PCA, CFA, and CFA-ord. In the evaluation of measurement instruments, typically CFA and CFA-ord are more appropriate than PCA. CFA-ord is generally preferred for ordinal polytomous items, and CFA for continuous items.

Second, one should select the estimation procedure, taking into account the distribution of item scores and the sample size available. Note that the minimally required sample sizes to achieve a reasonable estimate of an exploratory CFM crucially depend on the size of the communalities and the level of overdetermination (i.e., number of items per factor) (e.g., MacCallum, Widaman, Preacher, & Hong, 2001). In general, identifying a large number of factors (relative to the overall number of items) and/or small factors (i.e., factors related to a low number of items per factor) require large sample sizes. For "easy" cases, a sample size of 100 may suffice, whereas for "complicated" cases one may need up to 20 cases per item observed.

Third, one should decide whether the aim of the research requires assessing the number of major factors, the number of major factors and (some) minor factors. To assess the maximum number of factors to possibly consider, thus including minor factors, one can use the chi-square test in a stepwise manner, as explained in the section Likelihood ratio test. To assess the number of major factors, we suggest considering various criteria. In particular, we recommend HULL and PA, both in a variant that is in line with the model and estimation procedure selected. Further, we recommend considering various goodness-of-fit measures, as RMSEA, CFI, and SRMR. This implies examining the minimal number of factors required (associated with the model selected) to meet the recommend cut-off values for acceptable model fit. In our view, those cut-off values should be applied with some lenience. Then, the model(s) with the indicated number(s) of factor should be examined carefully, to judge the interpretability and theoretical plausibility.

Empirical example

To illustrate our strategy, we re-analyze data collected to evaluate the Dutch translation (van Dijk, Timmerman, Martel, & Ramsay, 2011) of the **Montreal Children's Hospital Feeding Scale** (MCH-FS) (Ramsay, Martel, Porporino, & Zygmuntowicz, 2011). This scale is a screening instrument for the detection of feeding problems in young children. Such problems are rather common, with 10–33% of the caregivers reporting feeding problems. The negative consequences range from increased parental

difficulties to hampered physical growth and mental development. Food problems are described by various types of symptoms, as food refusal, irregular eating, noncompliance during mealtime, and "mealtime negativity".

The MCH-FS was developed to provide clinicians with a valid and reliable screening instrument. The primary aim is to quickly identify the severity of feeding problems as reported by primary caregivers during a short consultation session, so that caregivers having severe problems can be referred to specialists. This implies that it is unnecessary to obtain a detailed picture of the nature of the feeding problems, as long as those parents having severe problems will be identified correctly. Because the scale is used in clinical practice, the scale should be as brief and easy to administer as possible. This implies that in this case, both the numbers of items and subscales should be as small as possible, as long as the instrument is sufficiently reliable in identifying severity of feeding problems. Note that for other purposes, like scientific research or a more detailed diagnosis, an instrument that disentangles several symptoms of feeding problems, and hence consists of various subscales, may be more appropriate.

The MCH-FS consists of 14 items on symptoms of feeding problems in the following domains: oral sensory/motor symptoms, appetite, parental concerns, mealtime behaviors, compensatory strategies, and family reactions. The primary feeder is asked to rate each of the 14 items on a seven-point Likert scale.

The Dutch translation of the MCH-FS is the SEP (*Screeningslijst Eetgedrag Peuters*). A normative sample of caretakers filled in the SEP when visiting their local Child Health Center. For details about the sampling procedure, we refer to van Dijk et al. (2011). We consider the scores of a sample of 1,386 caretakers, who completed all items of the SEP for their child, who was in the range from 6 to 208 months of age.

To illustrate the various methods to identify the number of factors underlying a data set, we estimated a series of CFMs, with one up to eight factors. Given that the number of domains covered equals six, the maximum of eight factors appeared to be large enough. All item distributions were right skewed (skewness range .32–2.30). Given the ordinal polytomous nature of the items and their skewness, we estimated CFA-ord models with robust weighted least squares (WLSMV) of polychoric correlations. We used the ESEM procedure of the computer program MPlus (Muthén & Muthén, 1998–2011), which provides both estimates of the CFA-ord models with the requested number of factors, and their associated measures of the chi-square, RMSEA, CFI, and SRMR.

In Table 11.2, the various fit measures associated with the estimated CFA-ord models are presented. The chi-square test result is significant (at $\alpha = .05$) for the CFA-ord models with seven and eight factors, suggesting that the total number of factors underlying the data set equals seven.

When considering the fit measures RMSEA, CFI, and SRMR and applying the recommended cut-off values for good fit, it is salient that mixed results appear. On the basis of the 95% CIs of the RMSEA (with recommended cut-off values of .05–.06 for close fit and .05–.08 for fair fit), two to three factors appear to be indicated. On the basis of the CFI value, with a commonly used cut-off value of .95, two factors are indicated. The SRMR value suggests either two (applying the stringent cut-off value of .05), or one (applying the more lenient cut-off value .08).

On the CFI values, we applied the Hull procedure (using own code). In Figure 11.1 (a), the CFI values versus degrees of freedom taken by the CFM concerned, are

Table 11.2 Fit measures for the CFA-ord model with m = 1 up to eight common factors. VAF is percentage explained variance.

m	χ^2	*Df*	*p-value* χ^2*-test*	*RMSEA [95% CI]*	*CFI*	*SRMR*
1	1,255.64	77	.00	.11 [.10; .11]	.92	.07
2	498.55	64	.00	.07 [.06; .08]	.97	.05
3	241.76	52	.00	.05 [.05; .06]	.99	.03
4	125.79	41	.00	.04 [.03; .05]	.99	.02
5	72.36	31	.00	.03 [.02; .04]	1.00	.01
6	34.52	22	.04	.02 [.00; .03]	1.00	.01
7	16.25	14	.30	.01 [.00; .03	1.00	.01
8	4.44	7	.73	.00 [.00; .02]	1.00	.00

presented. The thus obtained Hull-CFI indicated one factor. Hull-CFI based on a ULS estimation procedure on polychoric correlation matrices with the program Factor (Lorenzo-Seva & Ferrando, 2006) revealed the same result.

Finally, we applied PA-MRFA with polychoric correlation matrices, again using the program Factor. In Figure 11.1(b), the proportion of variance explained by the CFs and the 95th percentile of the variances explained of random data versus the number of factors estimated with MRFA are presented. As can be seen, the percentage of explained variance of the observed data exceeds that of random data only at a number of factors equal to one. Thus, PA-MRFA with a 95% threshold indicates a single factor to retain.

Given the aim of the MCH-FS as a screening instrument, the number of major factors is of primary interest, whereby possibly a few minor factors could be interesting to reveal the structure of the items. The major factor criteria considered suggest one or two factors, with only the RMSEA possibly indicating three. The solutions with one and two factors are well-interpretable. The solution with three factors is more difficult to interpret, because of substantial cross-loadings.

In Table 11.3, the loadings of the one-factor and two-factor models, after Geomin rotation, are presented. In the two-factor solution, Factor 1 clusters behavior that can be labeled as Negative mealtime behavior (e.g., poor appetite, start refusing food). The second factor pertains to a broader range of symptoms that can be labeled as Negative causes and consequences (e.g., holding food in mouth, poor growth). Item 5, involving the duration of mealtimes, has low loadings on both factors. The ordered categories of item 5 pertain to increasing duration of meals (in minutes), and those ordered categories do not necessarily indicate increasing feeding problems, because both very long and very brief mealtimes can be indicative of problematic behavior. This implies that in its current form item 5 is a weak item in the scale. The correlation between the two factors is .64, indicating that there is a rather strong positive relation between the two types of Negative symptoms in the normative population.

In the one-factor solution, the factor can be interpreted as Feeding Problems. Apart from item 5, which was already identified as a weak item, item 11 occurs as an item with a low loading. Given the interpretable relationships shown in the two-factor solution, we conclude that to represent in detail the various types of feeding problems, one needs a two-factor solution. To represent Feeding Problems in general, the one-factor solution suffices.

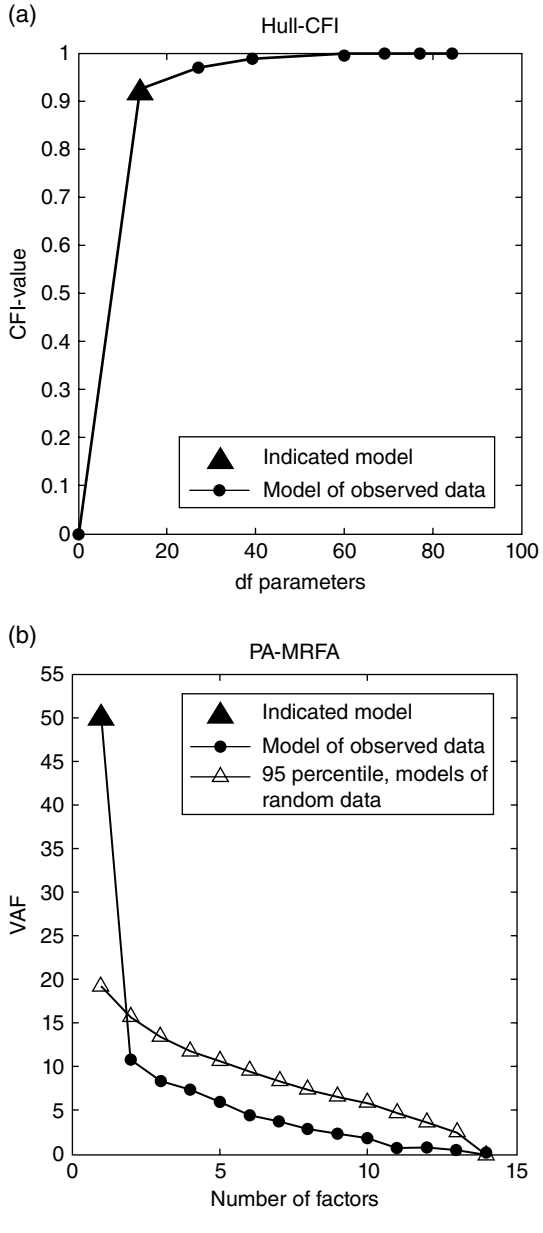

Figure 11.1 (a) Hull-CFI: CFI values versus number of parameters in CFM; (b) PA-MRFA: VAF for models of observed data and 95th percentile of VAF for models of random data versus the number of CFs in the model. VAF is percentage explained variance.

Those results suggest that to meet the goal of a screening instrument, it is sufficient to compute the score on a single total scale, for example by summing the item scores. The reliability of the total scale including all items (estimated on the basis of the one-factor solution (omega) (McDonald, 1999, p. 89)) appears to be rather high, namely 0.88. One may consider removing items 11 and 5, because they add relatively little to the

Table 11.3 Estimated loadings of the CFA-ord, two factors, after Geomin rotation. Items indicated with* are mirrored before analysis. Loadings >0.3 in absolute value are printed in bold type.

	One-factor solution		*Two-factor solution*		
Item number (brief description)	*Loading*	*Unique variance*	*Loading Factor 1*	*Loading Factor 2*	*Unique variance*
7 (gags/spits/vomits)	**.30**	.91	–.15	**.51**	.81
8 (holding food in mouth)*	**.44**	.80	.06	**.44**	.77
11 (poor chewing abilities)	.20	.96	–.28	**.54**	.83
3 (poor appetite)*	**.52**	.73	**.55**	–.01	.70
4 (start refusing food)*	**.69**	.52	**.79**	–.07	.44
1 (difficult mealtimes)*	**.86**	.26	**.90**	.01	.17
2 (worries about feeding)	**.64**	.59	.28	**.45**	.56
12 (poor growth)*	**.52**	.73	.15	**.43**	.71
6 (bad behavior at table)	**.78**	.40	**.71**	.13	.37
5 (long mealtimes)	.27	.93	.30	–.01	.92
9 (follow around/distract)	**.60**	.64	**.37**	.30	.63
10 (force to eat)*	**.74**	.46	**.50**	**.31**	.45
13 (influence relation)*	**.78**	.40	–.02	**.87**	.24
14 (influence family relations)	**.73**	.46	.10	**.73**	.36

reliability of the total scale. In future developments of the scale, one may consider to change item 5.

If the goal would have been to offer a more detailed diagnosis, the use of two subscales would be indicated, on the basis of the factor analysis results. Note that those results could be population dependent. In particular, it might be that even more than two subscales could be sensibly distinguished when a clinical sample, suffering from a varied range of feeding problems, would be considered.

Concluding Remark

In this chapter, we considered various approaches to arrive at the number of factors to use in an exploratory CFA. We included the standard linear CFA, suitable for continuous items, and CFA-ord for ordered polytomous items. As an alternative for polytomous items, exploratory item response theory (IRT) based models could be used. For two-parameter (normal ogive and logistic) IRT models, or constrained versions thereof, the approaches discussed here can be readily applied. This is so because CFA-ord is intimately related to the two-parameter IRT models (Takane & de Leeuw, 1987). For less constrained versions, the models differ, and hence the usefulness of the current approaches cannot be guaranteed.

Acknowledgments

We would like to thank Marijn van Dijk for sharing her data.

References

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*, 186–203.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, *27*(4), 509–540.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.

Cattell, R. B., & Tsujioka, B. (1964). The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales 1. *Educational and Psychological Measurement*, *24*(1), 3–30.

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 133–150.

Cho, S., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*(5), 748–759.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), Retrieved from: http://pareonline.net/pdf/v10n7.pdf.

Cota, A. A., Longman, R. S., Holden, R. R., & Fekken, G. C. (1993). Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educational and Psychological Measurement*, *53*(4), 865–876.

De Raad, B., & Barelds, D. P. H. (2008). A new taxonomy of Dutch personality traits based on a comprehensive and unrestricted list of descriptors. *Journal of Personality and Social Psychology*, *94*(2), 347–364.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326.

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.

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, *55*(3), 377–393.

Hayashi, K., Bentler, P. M., & Yuan, K. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 505–526.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*(2), 191–205.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

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*, 1–55.

Hubbard, R., & Allen, S. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research*, *15*(2), 173–190.

Humphreys, L. G., & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement*, *29*(3), 571–578.

Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer-Verlag.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: a computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments & Computers*, *38*, 88–91.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*(2), 340–364.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, *109*(3), 502–511.

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.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics*, *11*(1), 3–31.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished technical report.*, Retrieved from: http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf (accessed October 2017).

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38* (2), 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 19–30.

Muthén, L. K., & Muthén, B. O. (1998–2011). *MPlus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, *49*(4), 974–997.

Ramsay, M., Martel, C., Porporino, M., & Zygmuntowicz, C. (2011). The Montreal Children's Hospital Feeding Scale: A brief bilingual screening tool for identifying feeding problems. *Paediatrics & Child Health*, *16*(3), 147.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer-Verlag.
Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, *12*(3), 287–297.
Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403–414.
Sivo, S. A., Fan, X., Witta, E. L., & Willse, J. T. (2006). The search for "optimal" cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, *74*(3), 267–288.
Steger, M. F. (2006). An illustration of issues in factor extraction and identification of dimensionality in psychological assessment data. *Journal of Personality Assessment*, *86*(3), 263–272.
Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180.
Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
ten Holt, J. C., van Duijn, M. A. J., & Boomsma, A. (2010). Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling*, *52*(3), 272–297.
Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, D.C.: American Psychological Association.
Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220.
Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(4), 421–459.
Turner, N. E. (1998). The effect of common variance and structure pattern on random data eigenvalues: Implications for the accuracy of parallel analysis. *Educational and Psychological Measurement*, *58*(4), 541–568.
van Dijk, M., Timmerman, M. E., Martel, C., & Ramsay, M. (2011). Towards the development of a Dutch screening instrument for the detection of feeding problems in young children. *Netherlands Journal of Psychology*, *66*(4), 112–119.
Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.
Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In E. Helmes (Ed.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). New York, NY: Kluwer Academic/Plenum Publishers.
Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*(1), 1–28.
Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*(5), 697–716.
Widaman, K. F. (1993). Common factor analysis versus principal component analysis: differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*(3), 263–311.
Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79.
Yang-Wallentin, F., Joreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 392–423.
Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115–148.

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200.
Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213–249.
Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*(2), 253–269.
Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442.