APPROACHES TO IDENTIFYING THE OPTIMAL NUMBER OF FACTORS IN MULTILEVEL EXPLORATORY FACTOR ANALYSIS

A Dissertation

by

YUHONG JI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Oi-man Kwok
Co-Chair of Committee,	Wen Luo
Committee Members,	Lei-Shih Chen
	Hok Chio (Mark) Lai
	Myeongsun Yoon
Head of Department,	Shanna Hagan-Burke

August 2020

Major Subject: Educational Psychology

ABSTRACT

In exploratory factor analysis (EFA), there are numerous methods to extract the optimal number of factors, and these methods can be generally divided into two categories, namely, the model-selection-based approaches and the eigenvalue-based approaches. The model-selection-based approaches exploit the commonly used model-fit indexes and selection criteria, including the Root Mean Square Error of Approximation (RMSEA), the Comparative Fit Index (CFI), the Tucker-Lewis index (TLI), the Standardized Root Mean square Residual (SRMR), the overall model-fit chi-squared test, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the sample-size adjusted BIC (SBIC/SABIC). On the contrary, the eigenvalues-based approaches rely on eigenvalues, including the "eigenvalues > 1" criterion, the scree plot test and parallel analysis (PA).

The purpose of this dissertation is to examine the accuracy of these two approaches in identifying the number of factors in multilevel exploratory factor analysis (MEFA) through two studies, separately. In both studies, multilevel data were designed and simulated.

In the first study, we used both the model-based and the design-based approaches for the analysis. The results of Study 1 disclosed: (a) as the model-based approach showed most of the commonly used fit indexes and selection criteria were effective at identifying the correct number of factors at the within level, except for level-specific SRMR and AIC, while most of them tended to extract fewer factors at the between level,

(b) as the design-based approach showed, most of them were able to identify the true model as the best solution, except for SRMR, AIC, and ΔAIC.

In the second study, we used different eigenvalues extraction techniques for calculating eigenvalues, such as principal axis factoring (PAF), iterated principal axis factor (Iterated PAF), and maximum likelihood (ML). The results of Study 2 revealed: (a) the "eigenvalues > 1" criterion was not effective at searching for the optimal number of factors, (b) PA approach performed well in recovering the correct number of factors at the within level, while the performance of PA was related to sample size and ICC at the between level. In addition, PAF performed the best followed by Iterated PAF and then ML at both levels.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my committee chair and co-chair, Dr. Kwok and Dr. Luo for their continuous guidance and support for my Ph.D. study and research, for their patience, motivation, enthusiasm, and immense knowledge. I would like to thank the rest of my committee members, Dr. Lai, Dr. Yoon, and Dr. Chen, for their insight, comments, and support throughout the course of this research.

Thanks also go to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience.

Finally, my eternal appreciation also towards my loving family, my grandmother, parents, aunt, and uncle for all their love, encouragement, and blessings.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Professors Oi-man Kwok [advisor], Associate Professor Wen Luo [co-advisor], and Associate Professor Myeongsun Yoon of the Department of Educational Psychology, and Associate Professor Lei-Shih Chen of the Department of Health and Kinesiology from Texas A&M University, and Assistant Professor Hok Chio (Mark) Lai of the Department of Psychology form University of Southern California. All work for the dissertation was completed independently by the student.

Graduate study was supported by a fellowship from Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Exploratory Factor analysis (EFA) is a commonly utilized statistical method for a substantial variety of research goals. In the enormous majority of multivariate studies, the potential underlying unobserved (or latent) structures of a set of measured variables are the primary interest of research and thus EFA is used for exploring these possible underlying factors without having an a priori hypothesis of a preconceived structure (Child, 1990; Conway & Huffcutt, 2003; D'Haenens, Van Damme, & Onghena, 2010; Thompson, 2004; Yang & Bozdogan, 2011; Yang & Xia, 2015). EFA is sometimes taken as a first step for developing a hypothesized model (Child, 1990; Burton, & Mazerolle, 2011; Hyman & Sierra, 2012). Researchers might not at first have a clear idea of what the underlying structure of a set of variables or a newly developed scale looks like (Conway & Huffcutt, 2003). For example, Hyman and Sierra (2012) firstly employed EFA as a tool to develop a potential hypothesized model and then applied the confirmatory factor analysis (CFA) to evaluate their hypothesized model with the use of a different and independent random sample. Chan and colleagues (1999) also applied EFA to check the unidimensionality of each subscale of an existing multiscale instrument before analyzing the same scale with the item response theory (IRT). Moreover, unlike a typical CFA, EFA allows retaining nonzero cross-loadings in the measurement model (Fabrigar, Wegener, MacCallum, & Strahan, 1999), which may be more realistic for an observed variable that can have nonzero loading(s) on more than one non-targeted latent factor (Yang & Xia, 2015). Consequently, many researchers recommended conducting an EFA as a proper alternative when a CFA model shows a substantially inadequate fit to the data (Browne, 2001; Gorsuch, 1997; Muthén & Muthén, 2010; Schmitt, Sass, Chappelle, & Thompson, 2018). In only a very few studies, EFA has been used to reduce the complexity of a set of measured variables into a smaller number of dimensions that carried the same number of observed variables without any particular study's interest of the underlying latent structure (Conway & Huffcutt, 2003; Yang & Bozdogan, 2011).

One of the major challenges in the applications of EFA is to identify the optimal number of (latent or unobserved) factors. Since retaining incorrect number (i.e., too many or too few) of latent factors can lead to unstable factor pattern estimates and inaccurate factor solutions, selecting the optimal number of factors is the most important step in an EFA analysis (Preacher, Zhang, Kim, & Mels, 2013). Many studies (e.g., MacCallum, Widaman, Preacher, & Hong, 2001; Schmitt, 2011) have shown that under-factoring (i.e., extracting fewer than the true number of factors) is more problematic than over-factoring (i.e., extracting more than the true number of factors) because the information from a missing factor is redistributed or confounded with other factors (Zwick & Velicer, 1986). Thus, the missing factor that due to under-factoring leads to some of the observed variables falsely loaded on the other factors, results in distorting the factor loadings (Hayton, Allen, & Scarpello, 2004).

Currently, there are numerous methods for deciding how many necessary factors to retain in EFA analysis, and these methods can be generally divided into two categories, namely, the model-selection-based approaches and the eigenvalue-based approaches.

The model-selection-based approaches use the overall model fit statistics that originally developed under the CFA/structural equation modeling (SEM) framework. Preacher and colleagues (2013) proposed a model selection perspective that uses the commonly used fit statistics in SEM to choose the optimal number of factors in EFA. These commonly used fit statistics include the Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), the Comparative Fit Index (CFI; Jöreskog, 1993), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), the Standardized Root Mean square Residual (SRMR; Hu & Bentler, 1999), the overall model-fit chi-squared test/Likelihood Ratio test (Browne & Arminger, 1995), the Akaike information criterion (AIC; Akaike, 1987), the Bayesian information criterion (BIC; Schwarz, 1978), and the sample-size adjusted BIC (SBIC/SABIC; Sclove, 1987) and can be used to determine the number of factors in EFA.

On the contrary, the eigenvalues-based approaches rely on eigenvalues and include some of the traditional approaches such as the "eigenvalues > 1" criterion (a.k.a. the Kaiser Criterion) that determines the number of factors by the number of eigenvalues that larger than 1.0 (Kaiser, 1960). Another traditional approach is the scree plot test that identifies the number of factors based on the location of a severest drop off amongst the eigenvalues from the scree plot (Cattell, 1966). A third approach, parallel analysis (PA, Horn, 1965), compares two different sources of eigenvalues, one from the measured correlation matrix based on the measured variables and the other from another matrix based on random normal-distributed variables.

In recent decades, the traditional EFA has been extended to multilevel EFA (MEFA, Muthén, 1991, 1994) to account for the clustering effects in multilevel/hierarchical data which are very common in social and behavior sciences (Mayberry, Espelage, & Koenig, 2009; Moore et al., 2016; Reise, Ventura, Nuechterlein, & Kim, 2005) and educational studies (D'Haenens, Van Damme, & Onghena, 2010). Data are obtained through cluster sampling or collected on multiple levels, and the study interests are not limited to a particular data level but all the available levels and the corresponding variables (Muthén, 1991, 1994; Hsu, Lin, Kwok, Acosta, & Willson, 2017; Zyphur, Kaplan, & Christian, 2008). For example, in the field of education, pupils are nested within classes, classes are nested within schools, and schools are further nested within school districts. Another example is that in the field of psychiatry, clients are nested within doctors, doctors are nested within clinical departments, and clinical departments are further nested within hospitals. Pupils/clients from the same classes/doctors are likely to be more homogenous than those pupils/clients from different classes/doctors, because of the influences of the same classes/doctors. Therefore, the effect of dependency/clustering needs to be considered for these examples even these studies were only focus on the individual level (Cheung, Leung, & Au, 2006). Ignoring the existence of multilevel data structures may produce improper parameter estimates and statistical inferences, such as biased parameter estimation, deflated standard errors, inflated Type I error and inaccurate overall model goodness-of-fit (Muthén & Satorra, 1995; Raudenbush & Bryk, 2002). Thus, it is important to adopt the multilevel methods that can properly analyze multilevel data and take the issue of non-independence or clustering effect into account during the analysis (Goldstein & McDonald, 1988).

Currently MEFA has already had a range of practical applications. One example is that a study of psychiatric assessment with repeated measurement used MEFA to investigate the instrument structure through separating the within-individual and between-individual variances (Reise, Ventura, Nuechterlein, & Kim, 2005). Another example is a cross-cultural research with individual- and culture-level analyses, and the authors tested three different more-than-one-level models through MEFA in order to find out the number of factors in different levels (Cheung, Leung, & Au, 2006). Yet another example is an educational effectiveness study which has class-level and school-level process items, but it is impracticable to define a priori latent structure for the school-level, so MEFA is applied to evaluate the school process variables which depended on teachers' perceptions (D'Haenens, Van Damme, & Onghena, 2010). Finally, from the field of psychometric properties, researchers used MEFA to discover the factor structure of the Positive Values Scale as part of a school climate survey, and the results showed a model with 2 factor model in the within-level and 1 factor model in the between-level (Huang & Cornell, 2016).

The reasons behind such extensive applications of MEFA are two leading advantages. First, the total variances/covariance(s) are decomposed into the pooled within-cluster and the between-cluster variances/covariance(s) and these two variances/covariance(s) are independent from each other (Zyphur, Kaplan, & Christian, 2008). Second, it allows distinctive factor structures across different levels that can truly show the natural features of data (Reise, Ventura, Nuechterlein, & Kim, 2005; Zyphur, Kaplan, & Christian, 2008), thus researchers can investigate the number of factors for the within-level and between-level separately (D'Haenens, Van Damme, & Onghena, 2010; Huang & Cornell, 2016). Kim, Dedrick, Cao, & Ferron (2016) believed that there were three different or competing models at two-level MEFA: (a) the same number of factors at each cluster and equal loading across clusters, (b) the same number of factors at each cluster but different loadings across clusters, (c) different numbers of factors at two levels.

MEFA is a valuable tool, but it has one big disadvantage that is the uncertainty about how many individuals per cluster and how many higher levels or clusters are needed to handle analysis for both the within and the between levels (Reise, Ventura, Nuechterlein, & Kim, 2005). So far, there are no clear or uniform requirements of how many cluster numbers and cluster sizes are needed in MEFA to investigate the between-level variances. However, if there are many clusters, we can study the between-cluster differences in variable means and covariance structure separately (Ceulemans, Wilderjans, Kiers, & Timmerman, 2016).

Similar to EFA, a crucial concern in the applications of MEFA is to finalize the ideal number of latent factors for both within-level and between-level underlying the observed variables. To the best of our knowledge, no research has been conducted to confirm whether the model-selection-based approaches or the eigenvalues-based approaches applied in single-level EFA also perform well in MFEA. Hence, the purpose of this study was to examine the accuracy of these commonly used approaches in identifying the correct numbers of factors in MEFA within- and between-levels.

CHAPTER II

MODEL-SELECTION-BASED APPROACHES TO IDENTIFYING THE OPTIMAL NUMBER OF FACTORS IN MULTILEVEL EXPLORATORY FACTOR ANALYSIS

Introduction

The key assumption behind exploratory factor analysis (EFA) is the existence of a true common factor model with an accurate number of latent factors that can explain the population factor structure (Preacher, Zhang, Kim, & Mels, 2013). The common factor model is often compared with the component model. While the former attempts to account for measurement error, the latter assumes no measurement error (Schmitt, 2011). Nonetheless, many academics have argued for the true common factor model based on the following two points. First, researchers are not sure there is a true population model, and even if such a model exists, it may only capture limited information from the sample data (MacCallum, 2003; Meehl, 1990). Second, the relationships between observed items and latent factors are seldom truly linear, and the factor loadings are seldom consistently invariant across observations (MacCallum, Browne, & Cai, 2007). Consequently, searching for a true model is pointless, and using the terms *correct* or *true* with regard to the number of factors may also be meaningless. Nevertheless, Preacher et al. (2013) emphasized that detecting the optimal numbers of factors through EFA is still very valuable and functional, because this quasi-true model can approximate the objective truth and can be used in practice, serving some explicitly indicated scientific aims and assisting in realizing model descriptions and predictions.

According to these authors, therefore, the true motivation for using EFA is to obtain the ideal number of latent factors from observed sample data.

Eigenvalues-based approaches, including the “eigenvalues > 1” criterion (a.k.a. the Kaiser Criterion; Kaiser, 1960), the scree plot test (Cattell, 1966), and the parallel analysis (PA, Horn, 1965) are common ways to extract the optimal number of factors. Besides, other approaches are mainly based on model-selection. As Preacher and colleagues (2013) pointed out, identifying the optimal number of factors in EFA is fundamentally a model-selection issue, and one can typically evaluate each model through several statistical tests and report some common fit statistics for specific interpretations. This model-selection-based procedure is typically carried out within a CFA or SEM framework (especially when testing factorial invariance) with the use of model chi-squared difference and delta fit indexes (the difference-of-fit indexes between nested models for comparison, ΔGFI). These indexes include RMSEA, CFI, TLI, and SRMR, and the overall model-fit chi-squared test. Meanwhile, models with different numbers of factors can also be compared with each other, and the best-fitting model can be selected based on other selection criteria; the commonly used fit criteria for model comparisons include AIC, BIC, and SBIC.

RMSEA measures the discrepancy between the hypothesized model and the perfect model by taking model complexity into account, whereas both CFI and TLI assess incremental goodness-of-fit allowing for model complexity. Thus, both CFI and TLI measure the degree to which the hypothesized model is better than the “baseline” model in replicating the observed variance/covariance matrix. Finally, SRMR tests total goodness-of-fit by calculating the standardized discrepancy between the variance/covariance matrix based on the hypothesized model and the observed variance/covariance matrix. The formulas for these indexes are presented below.

$$RMSEA = \sqrt{Max\left(\frac{\chi^2 - df}{df(N-1)}, 0\right)}$$

$$CFI = 1 - \frac{Max[(\chi_H^2 - df_H),0]}{Max[(\chi_I^2 - df_I),0]}$$

$$TLI = \frac{(\chi_I^2/df_I) - (\chi_H^2/df_H)}{(\chi_I^2/df_I) - 1}$$

$$SRMR = \sqrt{\frac{S}{p(p-1)/2}}$$

Where *S* is

$$S = \sum_{j=1}^{p}\sum_{k-1}^{j-1}\left(\frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}, \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}\right)^2$$

Max indicates the maximum values of the values given in brackets. χ^2 is the noncentral chi-square value, *df* is its degrees of freedom, *N* is the total sample size, χ_H^2 is the chi-square value of the hypothesized model, df_H is its degrees of freedom, χ_I^2 is the chi-square value of the independent model, df_I is its degrees of freedom, *p* is the number of variables in the model, s_{jk} and σ_{jk} are the sample and the model-estimated covariance between the j^{th} and k^{th} variables.

For the purpose of obtaining the ideal number of factors, RMSEA, CFI, TLI, SRMR, and the overall model-fit chi-squared test are usually compared to the commonly recommended cutoff values by the fit indexes; that is, RMSEA ≤ .05 (Jöreskog, 1993);

CFI ≥ .95 (Bentler, 1990; Hu & Bentler, 1999); TLI ≥ .95 (Hu & Bentler, 1999; Tucker & Lewis, 1973); SRMR ≤ .05 (Jöreskog, 1993); and for the overall model-fit chi-squared test *p* value ≥ .05 (Jöreskog, 1993). Chen (2007) proposed a change/difference of goodness-of-fit indexes (ΔGFI) between nested models, so that when ΔGFI between the models is small (e.g., smaller than .015 for RMSEA, .01 for CFI, and .01 for SRMR), one can stop the comparison and select the parsimonious model with fewer parameters, given that a small ΔGFI indicates no substantial differences between the compared models in terms of overall model fit. In the present study, we examined the effectiveness of ΔGFI along with other fit indexes in choosing the optimal number.

In addition to model-fit statistics, information criteria (IC) are also commonly used for model comparison. AIC, BIC, and SBIC/SABIC are all maximum likelihood-based information criteria that take model complexity (number of unknown parameters) into account (Chen, Luo, Palardy, Glaman, & McEnturff, 2017; Tein, Coxe, & Cham, 2013). Moreover, AIC and BIC are the commonly used ICs for model selection (Tein et al., 2013). The formulas for these ICs are presented below.

$$AIC = -2LL + 2k$$

$$BIC = -2LL + k * \ln{(n)}$$

$$SBIC = -2LL + k * \ln{(\frac{n+2}{24})}$$

LL is the log likelihood of the fitted model, *k* is the number of estimated parameters in the model, and *n* is the sample size. The main difference across these three ICs is found in the penalty terms that can lead to quite different solutions (Chen et al., 2017). The general guideline is to select models with a smaller IC value. Based on the above formulas, when a model contains fewer parameters (meaning the model is less complex) based on a large sample size, the likelihood value will increase and the *-2LL* will decrease (Chen et al., 2017). Thus, smaller values of AIC, BIC, and SBIC/SABIC indicate better model fit among models (Preacher et al., 2013). Cutoff values for the difference in AIC and BIC between two models have been proposed. To be more specific, a model with a smaller AIC value is sufficiently better than others when ΔAIC is larger than 4 (Burnham & Anderson, 1998), or when ΔBIC is larger than 2 (Raftery, 1995). The IC family has been widely used for model selection purposes in other SEM-based analyses such as mixture models. Meanwhile, Preacher and colleagues (2013) believed that these ICs can be adopted when searching for the optimal number of latent factors from a mode-selection perspective.

Performance of the Model-Selection-Based Approaches in EFA

These previous indexes are generally reported within a CFA/SEM framework, and their performance can be specific to the nature of the parameters (loadings vs. covariance). Hu and Bentler (1998, 1999) noted that RMSEA, CFI, and TLI are quite sensitive to detecting models with misspecified factor loadings, and SRMR is the most sensitive to detecting models with misspecified latent structure or factor covariance. Hu and Bentler (1999) suggested combining two indexes to evaluate model goodness-of-fit when the sample size is large, and noted that the results from combining CFI/TLI and SRMR or RMSEA and SRMR provide the lowest Type I and Type II error rate (Hu & Bentler, 1998). However, some authors have claimed that SRMR is not the most sensitive to misspecified factor covariance and that CFI, TLI, and RMSEA are not more sensitive to misspecified factor loadings (Fan & Sivo, 2005; Fan, Thompson, & Wang, 1999). Other methodologists furthermore have pointed out that even when models have quite low factor loadings (equal to or less than .5), RMSEA and SRMR can still perform better than CFI and TLI (Mahler, 2011; Sharma, Mukherjee, Kumar, & Dillon, 2005).

Many methodologists have further examined the performance of these fit indexes in EFA and found that RMSEA performs better than others when the sample size is larger than 250 (Clark & Bowles, 2018; Hu & Bentler, 1999) or when the target is to minimize the discrepancy of an approximation (representing verisimilitude) (Preacher et al., 2013), given that RMSEA is usually considered the degree of sample discrepancy from population (Cudeck & Henly, 2003; Garrido et al., 2016; MacCallum, 2003). However, when the sample size is large, SRMR is less accurate due to the tendency to choose fewer latent factors (Barendse, Oort, & Timmerman, 2015). Performing an EFA simulation study for ordered categorical data, Yang and Xia (2015) uncovered that CFI performed worse than RMSEA because the average of CFI is still greater than .95 when the number of factors is smaller than the true number. Garrido and colleagues (2016) examined these indexes' precision with categorical data via Monte Carlo simulation. Their results indicated that the accuracy of SRMR was lower than that of RMSEA, and that CFI and TFI showed the highest performance. The authors also investigated the performance of the change/difference of CFI/TLI, RMSEA, and SRMR, and discovered that the change/difference can significantly improve the latent number estimation (Garrido et al., 2016). Clark and Bowles (2018) studied the ability of common model-fit indexes through Monte Carlo simulation for continuous data; their results showed that the chi-squared test, RMSEA, and CFI/TLI can all reject under-factor models when choosing right threshold values (such as $p < .05$, RMSEA < .08, or CFI/TLI > .90), and these indexes often accept over-factor models even when choosing restricted thresholds (such as $p < .01$, RMSEA < .05, or CFI/TLI > .95).

Other studies also have demonstrated that the overall model-fit chi-squared test is good at rejecting the underspecified models, but bad at refusing the over-specified models, both for continuous and ordered categorical observed data (Hayashi, Bentler, & Yuan, 2007; Yang & Xia, 2015). What is more, the chi-squared test has been criticized for a long time, since it is likely to reject reasonable models because of (1) trivial misspecification in big samples, and (2) its inability to reject the wrong models with important misspecifications in small samples (Gerbing & Anderson, 1992). In summary, most of the fit indexes are less likely to pick the under-factor solution and more likely to select the over-factor solution.

Meanwhile, the IC family performs well when the goal is to choose a hypothesized model that minimizes overall discrepancy while representing generalizability (Preacher et al., 2013). When the sample size is large, AIC performs better in choosing the true number of factors than BIC (Song & Belin, 2008), and BIC usually prefers simpler models than AIC (Bozdogan, 2000). However, when the sample size is small, BIC and SBIC perform better in extracting the correct number of factors than AIC (Lopes & West, 2004; Tein et al., 2013).

In sum, many researchers have examined the effectiveness of different model fit indexes (the selection of either the largest or the smallest values) and information criteria for retaining the optimal number of factors in traditional EFA. However, most of these studies showed that relying solely on either the information criteria or model-fit indexes likely results in over-extraction (i.e., over-fitting the model with too many factors) when choosing restricted thresholds. Thus, one should not select the factor model solely based on the fit indexes or information criteria. An ideal EFA model would have a balance of goodness-of-fit and parsimony (Preacher et al., 2013).

The Model-Selection-Based Approaches for MEFA

Model selection using fit indexes becomes a more complex issue in multilevel SEM given that it involves more than one model (with multiple models at different levels). There are several disadvantages when only using the overall test statistics for detecting the entire model fit (i.e., estimating the within-cluster and the between-cluster variances/covariance(s) simultaneously). First, the goodness-of-fit of the between-cluster/high-level model might not be as sensitive as the within-cluster/low-level model due to the calculation based on the dissimilar gap weights from the within- and between-sample sizes, and the traditional methods are especially limited in detecting the goodness-of-fit for high-level models (Hsu, Kwok, Lin, & Acosta, 2015; Ryu & West, 2009; Yuan & Bentler, 2007). Second, when the traditional indexes (RMSEA, CFI, TLI, SRMR, and the chi-squared test) show a model has inadequate fit, it is difficult to tell whether the between-cluster model or the within-cluster model is incorrect, or whether both of them are misspecified (Yuan & Bentler, 2007). Third, misspecification at a single level is likely to influence the parameter estimations at the other levels; that is, even when the misspecification is tiny at the single level, the indexes can still reveal that the model fits the data poorly (Yuan & Bentler, 2007).

Level-specific SRMR, CFI and RMSEA have been proposed to resolve the above-mentioned problems and provide more details about the model goodness-of-fit at a specific level (Ryu & West, 2009; Yuan & Bentler, 2007). Level-specific SRMR, CFI, and RMSEA are obtained based on a partially saturated model. For example, a saturated between-cluster model is specified to measure the goodness-of-fit of a within-cluster model, and vice versa. In the previous simulation studies, the level-specific methods were found to have a greater ability to detect misspecification for between-cluster models than the single-level standard methods (Hsu, Lin, Kwok, Acosta, & Willson, 2017; Yuan & Bentler, 2007).

There is no research on the evaluating the effectiveness of these fit indexes and selection criteria in MEFA, so the purpose of this study was to examine the effectiveness of the traditional model-selection-based approaches and level-specific SRMR, CFI and RMSEA on extracting the optimal number of latent factors in MEFA. To that end, the study used Monte Carlo simulations for data generation.

Methods

Data Generation

In this study, Monte Carlo simulations were conducted using M*plus* version 8 (Muthén, & Muthén, 2017). Two-level data with eight items (y_1-y_8) were generated based on two scenarios through M*plus*. These eight items were all the standard normal distribution with mean = 0 and standard deviation (SD) = 1. Thus, multivariate normality was assumed for both scenarios, and all factor loadings were the standardized loadings.

Scenario 1

In this scenario (see Figure 1), Level-1 and Level-2 had different factor structures with two factors in Level-1 and one factor in Level-2. Specifically, in Level-1, y_1-y_4 were under Factor 1 and y_5-y_8 were under Factor 2. Within each factor, the first three items were good (main factor loadings = .6 and cross-loadings = .1), and the fourth item was a poor item (main factor loadings and cross-loadings were all .25). The correlation between the two within factors was .5. In Level-2, y_1-y_8 were all under one factor with the first six items being good items (factor loadings = .6) and the last two items being poor items (factor loadings = .25).

Scenario 2

In this scenario (see Figure 2), Level-1 and Level-2 had the same two-factor structures. Specifically, the Level-1 and Level-2 factor structures and factor loadings were the same as Level-1 in Scenario 1. The correlations between the two factors (within and between) were both .5. Because previous reviews have shown that when the correlation between factors is high ($r \geq .5$), the model fit statistics commonly extract fewer than the true latent factors. We wanted to find the best way of identifying the optimal number of factors when the correlation is high.

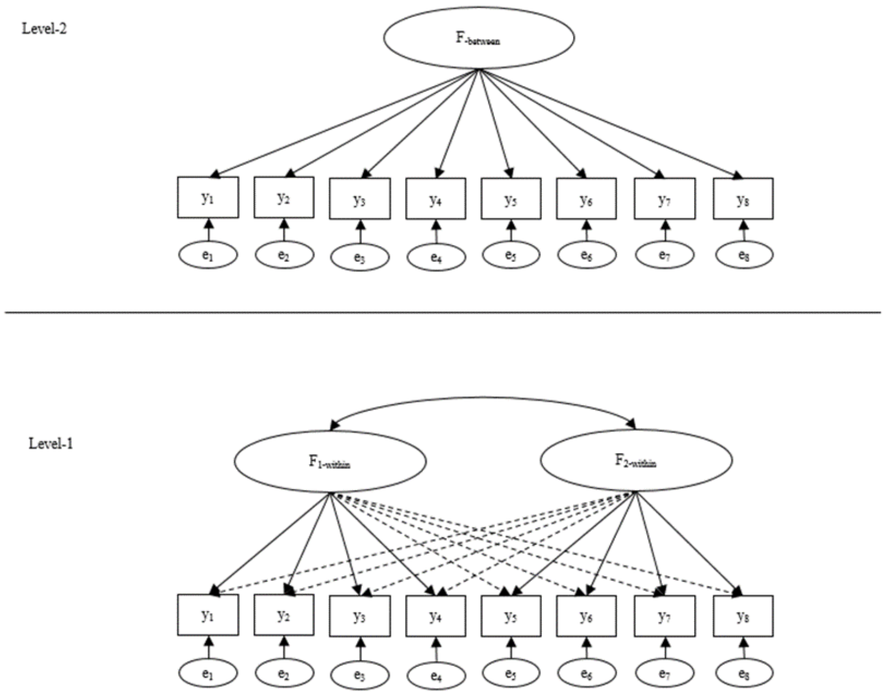

Figure 1 Scenario 1: Level-1 and Level-2 have different structures.

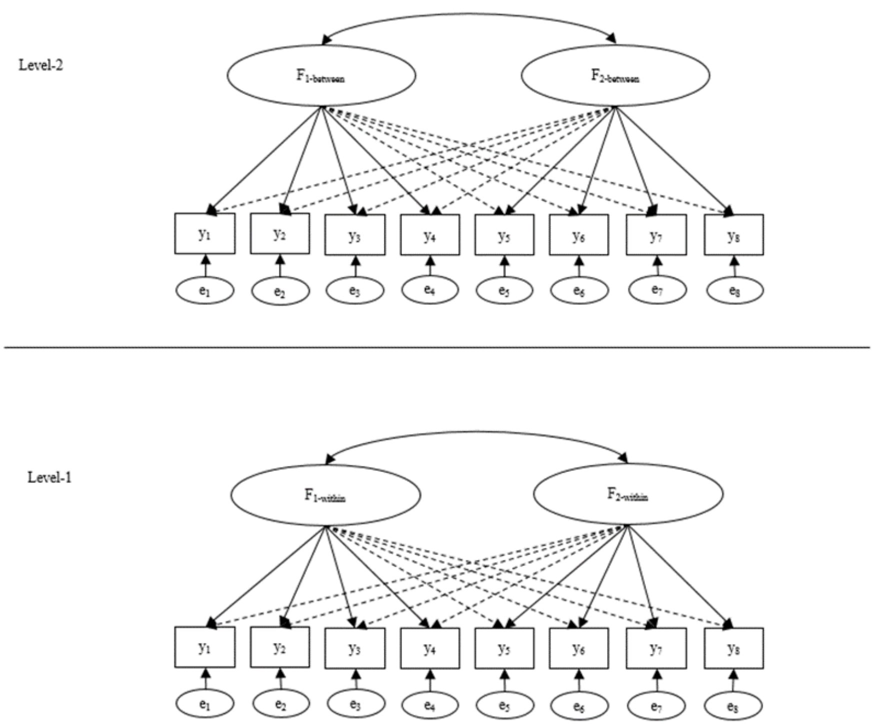

Figure 2 Scenario 2: Level-1 and Level-2 have the same structure.

Design Factors

The data generation also involved different sample size and intraclass correlation coefficient (ICC) conditions. ICC conditions are commonly a consideration in MSEM simulation studies; for example, Hsu et al. (2017) considered the generalization of their findings and simulated six different ICC conditions (.1, .2, .3, .4, .5, and 1.00). In the present study, the design factors included cluster numbers (CN = 100, 200, and 500), cluster sizes (CS = 10, 20, and 50), and ICCs (.1, .2). Meanwhile, 500 replications were produced for each combination of design factors in the two scenarios. As a result, a total of 2 (scenarios) × 3 (cluster numbers) × 3 (cluster sizes) × 2 (ICC) × 500 = 36 (Conditions) × 500 = 18,000 replications were generated.

Data Analysis

After each dataset was generated, M*plus* was used to analyze data with its commands. M*plus* provides two approaches to MEFA: the model-based approach and the design-based approach. In the first approach, a level-specific model is specified for each level (Wu & Kwok, 2012). In M*plus*, the syntax is TYPE = TWOLEVEL EFA. When estimating the Level-1 factor structure, the Level-2 model is saturated and vice versa. On the other hand, the design-based approach assumes factorial invariance across levels and accounts for the multilevel structure by correcting the standard errors according to the sampling design (Wu & Kwok, 2012). In M*plus*, the syntax is TYPE = COMPLEX EFA.

Compared to the model-based approach, the design-based approach is relatively simpler, as it only analyzes one global model and estimates parameters based on this model, which means it assumes the same structure across different levels; but sometimes this assumption may be wrong (Wu & Kwok, 2012). Both approaches select a default maximum likelihood estimator and a robust standard error estimator. Thus, in Scenario 1, we only used the model-based approach because the assumption of factorial invariance across levels was violated and different levels were assumed to have different factor, while in Scenario 2, we applied both the model-based and the design-based approaches for the data analysis.

Model Selection

When comparing different models with different numbers of factors, we used the model-selection-based approaches, including RMSEA (overall vs. level-specific), CFI (overall vs. level-specific), TLI, SRMR (overall vs. level-specific), ΔRMSEA, ΔCFI, ΔSRMR, AIC, BIC, SBIC, ΔAIC, and ΔBIC. Table 1 illustrates the specific values or principles of the acceptable models for these indexes and the criteria used in this study. In addition, as a rule of thumb, we chose the simple model once the indexes reached the criteria and cutoff values. For example, if RMSEAs of the one-factor and two-factor models were all less than or equal to .05, we chose the simpler model (one-factor model) as the final model only based on RMSEA.

Table 1 The Specific Values or Principles of Indexes and Criteria for the Acceptable Model Using the Model-Selection-Based Approaches.

Name	Explanation	Specific values or principles
Indexes		
RMSEA	The overall RMSEA	≤ .05
ΔRMSEA	The difference of two overall RMSEAs	≥ .015
$RMSEA_{PS\_W}$	The within-level specific RMSEA when the between-level model is saturated	≤ .05
$\Delta RMSEA_{PS\_W}$	The difference of two within-level specific RMSEAs	≥ .015
$RMSEA_{PS\_B}$	The between-level specific RMSEA when the within-level model is saturated	≤ .05
$\Delta RMSEA_{PS\_B}$	The difference of two between-level specific RMSEAs	≥ .015
CFI	The overall CFI	≥ .95
ΔCFI	The difference of two overall CFIs	≥ .01
CFI_{PS_W}	The within-level specific CFI when the between-level model is saturated	≥ .95
ΔCFI_{PS_W}	The difference of two within-level specific CFIs	≥ .01
CFI_{PS_B}	The between-level specific CFI when the within-level model is saturated	≥ .95
ΔCFI_{PS_B}	The difference of two between-level specific CFIs	≥ .01
TLI	The overall TLI	≥ .95
SRMR	The overall SRMR	≤ .05
$SRMR_{PS\_W}$	The within-level specific SRMR when the between-level model is saturated	≤ .05
$\Delta SRMR_{PS\_W}$	The difference of two within-level specific SRMRs	≥ .01
$SRMR_{PS\_B}$	The between-level specific SRMR when the within-level model is saturated	≤ .05
$\Delta SRMR_{PS\_B}$	The difference of two between-level specific SRMRs	≥ .01
p value	The overall model-fit chi-squared test *p* value	≥ .05
Criteria		
AIC	-	Smaller value
BIC	-	Smaller value
SBIC	-	Smaller value
ΔAIC	The difference of two AICs	≥ 4
ΔBIC	The difference of two BICs	≥ 2

Results

The Model-Based Approach

Level-1 Factor Extraction

Table 2 shows the percentages of iterations that retained one, two, and three factors, respectively, based on the various model-fit indexes and criteria. The results were aggregated over the two scenarios and the different sample size conditions because there was no difference in the Level-1 factor structure between the two scenarios. Also, little variation was found in the percentages across sample sizes.

Table 2 Level-1 Factor Extraction Results Using the Model-Based Approach.

Index or Criteria	ICC = 0.1				ICC = 0.2			
	*N**	Percentage			*N**	Percentage		
		1-factor	**2-factor**	3-factor		1-factor	**2-factor**	3-factor
Scenario 1								
RMSEA	500.0	0.3%	**99.7%**	0.0%	500.0	1.6%	**98.4%**	0.0%
$RMSEA_{PS\_W}$	500.0	0.2%	**99.8%**	0.0%	500.0	1.2%	**98.8%**	0.0%
CFI	500.0	5.3%	**94.7%**	0.0%	500.0	57.3%	**42.7%**	0.0%
CFI_{PS_W}	500.0	3.0%	**97.0%**	0.0%	500.0	41.5%	**58.5%**	0.0%
TLI	500.0	0.0%	**100.0%**	0.0%	500.0	0.0%	**99.9%**	0.0%
$SRMR_{PS\_W}$	500.0	99.8%	**0.2%**	0.0%	500.0	100.0%	**0.0%**	0.0%
p value	492.0	0.0%	**96.8%**	3.2%	492.2	0.0%	**96.8%**	3.2%
AIC	500.0	0.0%	**89.1%**	10.9%	500.0	0.0%	**88.4%**	11.6%
BIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.2%	**99.8%**	0.0%
SBIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.0%	**99.9%**	0.0%
ΔRMSEA	455.8	0.0%	**97.3%**	2.7%	465.2	0.0%	**96.7%**	3.3%
$\Delta RMSEA_{PS\_W}$	455.8	0.0%	**98.7%**	1.3%	464.7	0.0%	**98.4%**	1.6%
ΔCFI	462.2	0.0%	**100.0%**	0.0%	487.3	0.0%	**99.9%**	0.0%
ΔCFI_{PS_W}	459.2	0.0%	**100.0%**	0.0%	482.1	0.0%	**99.9%**	0.1%
$\Delta SRMR_{PS\_W}$	500.0	0.0%	**99.9%**	0.1%	500.0	0.0%	**99.8%**	0.2%
ΔAIC	500.0	0.0%	**97.3%**	2.7%	500.0	0.0%	**97.0%**	3.0%
ΔBIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.3%	**99.7%**	0.0%

Table 2 Continued.

Index or Criteria	ICC = 0.1				ICC = 0.2			
	*N**	Percentage			*N**	Percentage		
		1-factor	**2-factor**	3-factor		1-factor	**2-factor**	3-factor
Scenario 2								
RMSEA	500.0	0.1%	**99.9%**	0.0%	500.0	1.6%	**98.3%**	0.0%
$RMSEA_{PS\_W}$	500.0	0.1%	**99.9%**	0.0%	500.0	1.3%	**98.7%**	0.0%
CFI	500.0	6.0%	**94.0%**	0.0%	500.0	56.7%	**43.3%**	0.0%
CFI_{PS_W}	500.0	3.0%	**97.0%**	0.0%	500.0	40.5%	**59.5%**	0.0%
TLI	500.0	0.0%	**99.9%**	0.0%	500.0	0.0%	**99.9%**	0.0%
$SRMR_{PS\_W}$	500.0	99.6%	**0.4%**	0.0%	500.0	100.0%	**0.0%**	0.0%
p value	492.4	0.0%	**96.7%**	3.3%	493.1	0.0%	**97.1%**	2.9%
AIC	499.8	0.0%	**88.3%**	11.7%	500.0	0.0%	**89.0%**	11.0%
BIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.2%	**99.8%**	0.0%
SBIC	500.0	0.0%	**99.9%**	0.1%	500.0	0.0%	**99.9%**	0.1%
ΔRMSEA	456.1	0.0%	**96.2%**	3.8%	464.2	0.0%	**96.6%**	3.4%
$\Delta RMSEA_{PS\_W}$	456.1	0.0%	**98.2%**	1.8%	463.7	0.0%	**98.4%**	1.6%
ΔCFI	464.8	0.0%	**99.9%**	0.1%	486.6	0.0%	**99.9%**	0.0%
ΔCFI_{PS_W}	460.4	0.0%	**99.9%**	0.1%	481.0	0.0%	**99.9%**	0.1%
$\Delta SRMR_{PS\_W}$	499.1	0.0%	**99.8%**	0.2%	500.0	0.0%	**99.7%**	0.3%
ΔAIC	500.0	0.0%	**96.8%**	3.2%	500.0	0.0%	**96.8%**	3.2%
ΔBIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.3%	**99.7%**	0.0%

* *N* is the average of replications among nine sample size (3 cluster numbers × 3 cluster sizes), some of *N* were not equal to 500 due to convergency problems.

In general, most of the fit indexes had more than a 95% chance of correctly identifying the two-factor model as the best solution. Specifically, the AIC index performed slightly worse, with around a 90% chance of finding the two-factor solution. The $SRMR_{PS\_W}$ was the worst index, as it missed the two-factor solution and chose the one-factor model as the best solution more than 99% of the time. Interestingly, the performance of CFI and CFI_{PS_W} was good when ICC was .1, but deteriorated sharply when ICC increased to .2. On the other hand, when changes in the fit indexes were considered, such as ΔCFI, ΔCFI_{PS_W}, $\Delta SRMR_{PS\_W}$, and ΔAIC, their performance significantly improved to over 95%.

Level-2 Factor Extraction

Table 3 shows the percentages of iterations that retained one, two, and three factors, respectively, at the between-level under Scenario 1.

Table 3 Level-2 Factor Extraction Results Under Scenario 1 Using the Model-Based Approach.

Index or Criteria	ICC = 0.1				ICC = 0.2			
	*N**	Percentage			*N**	Percentage		
		1-factor	2-factor	3-factor		**1-factor**	2-factor	3-factor
RMSEA	499.8	**100.0%**	0.0%	0.0%	500.0	**100.0%**	0.0%	0.0%
$RMSEA_{PS\_B}$	474.4	**93.0%**	6.6%	0.4%	479.6	**94.1%**	5.7%	0.2%
CFI	499.8	**100.0%**	0.0%	0.0%	500.0	**100.0%**	0.0%	0.0%
CFI_{PS_B}	435.9	**82.4%**	16.1%	1.5%	465.9	**90.4%**	9.3%	0.3%
TLI	499.8	**100.0%**	0.0%	0.0%	500.0	**100.0%**	0.0%	0.0%
$SRMR_{PS\_B}$	348.4	**45.5%**	37.2%	17.2%	416.8	**64.5%**	31.1%	4.4%
p value	488.6	**96.0%**	3.8%	0.1%	491.6	**97.0%**	2.9%	0.1%
AIC	398.7	**81.3%**	16.9%	1.9%	427.9	**84.6%**	14.6%	0.8%
BIC	398.7	**100.0%**	0.0%	0.0%	427.9	**100.0%**	0.0%	0.0%
SBIC	398.7	**100.0%**	0.0%	0.0%	427.9	**100.0%**	0.0%	0.0%
ΔRMSEA	398.7	**96.5%**	2.2%	1.3%	427.9	**96.8%**	1.9%	1.3%
$\Delta RMSEA_{PS\_B}$	374.0	**66.0%**	25.8%	8.1%	408.3	**68.5%**	24.5%	7.0%
ΔCFI	398.7	**99.8%**	0.0%	0.2%	427.9	**99.9%**	0.1%	0.0%
ΔCFI_{PS_B}	335.4	**70.3%**	24.6%	5.1%	397.4	**72.6%**	21.9%	5.6%
$\Delta SRMR_{PS\_B}$	252.2	**3.4%**	22.2%	74.4%	345.0	**6.9%**	26.3%	66.8%
ΔAIC	398.7	**94.3%**	5.1%	0.6%	427.9	**95.7%**	4.2%	0.1%
ΔBIC	398.7	**100.0%**	0.0%	0.0%	427.9	**100.0%**	0.0%	0.0%

* *N* is the average of replications among nine sample size (3 cluster numbers × 3 cluster sizes), some of *N* were not equal to 500 due to convergency problems.

In this scenario, where there was only one Level-2 factor in the data generation, most of commonly used fit indexes (RMSEA, CFI, TLI, *p* value, BIC, SBIC, ΔRMSEA,

ΔCFI, ΔAIC, and ΔBIC) could correctly identify the one-factor model as the best solution (near or above 95% of the iterations), except for AIC, whose percentage for identifying the correct number of factors was around 80%. However, level-specific indexes such as CFI_{PS_B}, $SRMR_{PS\_B}$, $\Delta RMSEA_{PS\_B}$, and ΔCFI_{PS_B} performed significantly worse than the overall version of these indexes, except for $RMSEA_{PS\_B}$, which performed adequately. Similar to its performance in Level-1 factor extraction, level-specific SRMR and ΔSRMR were found to be of little use.

Table 4 shows the percentages of iterations that retained one, two, and three factors, respectively, at the between-level under Scenario 2. In this scenario, in which there were two Level-2 factors, none of the fit indexes demonstrated high accuracy in identifying the two-factor model as the best solution. However, a few indexes were promising (over 80% of the iterations) when ICC was relatively large (i.e., .2), including $RMSEA_{ps\_B}$, CFI_{ps_B}, $SRMR_{PS\_B}$, AIC, and ΔAIC.

Table 3 and 4 show higher ICC can help to increase the chance of identifying the correct number of factors at the between-level. Furthermore, this study discovered that higher cluster numbers or more cluster sizes have a similar influence on these indexes and criteria, especially for $SRMR_{PS\_B}$, AIC, and ΔAIC, that is, the Pearson correlations between the correct chance of identifying the factors from them and the sample size are statistically significant ($p < .05$).

Table 4 Level-2 Factor Extraction Results Under Scenario 2 Using the Model-Based Approach.

Index or Criteria	ICC = 0.1				ICC = 0.2			
	*N**	Percentage			*N**	Percentage		
		1-factor	**2-factor**	3-factor		1-factor	**2-factor**	3-factor
RMSEA	499.4	100.0%	**0.0%**	0.0%	500.0	100.0%	**0.0%**	0.0%
$RMSEA_{PS\_B}$	475.0	64.2%	**35.1%**	0.7%	476.6	15.0%	**83.2%**	1.8%
CFI	499.4	100.0%	**0.0%**	0.0%	500.0	100.0%	**0.0%**	0.0%
CFI_{PS_B}	444.7	39.7%	**57.9%**	2.4%	463.8	9.0%	**86.7%**	4.2%
TLI	499.4	100.0%	**0.0%**	0.0%	499.8	99.8%	**0.2%**	0.0%
$SRMR_{PS\_B}$	386.4	20.7%	**61.1%**	18.2%	446.0	1.5%	**88.1%**	10.4%
p value	483.2	61.0%	**38.3%**	0.8%	483.4	23.1%	**75.0%**	1.9%
AIC	447.0	35.7%	**58.8%**	5.5%	483.2	9.0%	**82.1%**	9.0%
BIC	447.0	99.9%	**0.1%**	0.0%	483.2	79.0%	**21.0%**	0.0%
SBIC	447.0	95.0%	**5.0%**	0.0%	483.2	54.4%	**45.6%**	0.0%
ΔRMSEA	447.0	92.3%	**4.9%**	2.9%	483.2	51.4%	**39.1%**	9.5%
$\Delta RMSEA_{PS\_B}$	379.6	26.8%	**60.7%**	12.6%	362.3	6.9%	**72.4%**	20.7%
ΔCFI	447.0	99.7%	**0.2%**	0.1%	483.2	95.7%	**3.6%**	0.8%
ΔCFI_{PS_B}	320.6	29.9%	**59.3%**	10.8%	338.8	7.8%	**75.1%**	17.1%
$\Delta SRMR_{PS\_B}$	252.8	0.1%	**23.2%**	76.7%	305.1	0.0%	**28.6%**	71.4%
ΔAIC	447.0	54.4%	**44.1%**	1.4%	483.2	17.5%	**80.5%**	2.0%
ΔBIC	447.0	99.9%	**0.1%**	0.0%	483.2	80.3%	**19.7%**	0.0%

* *N* is the average of replications among nine sample size (3 cluster numbers × 3 cluster sizes), some of *N* were not equal to 500 due to convergency problems.

The Design-Based Approach

Under Scenario 2, in which the factor structure was invariant across levels, the design-based approach was used. As shown in Table 5, most of the fit indexes were able to identify the two-factor model as the best solution, except for SRMR, AIC, and ΔAIC. Interestingly, when ICC changed from .1 to .2, the accuracy of RMSEA dropped from 99.7% to 61.1%.

Table 5 Factor Extraction Results Under Scenario 2 Using the Design-Based Approach.

Index or Criteria	ICC = 0.1				ICC = 0.2			
	*N**	Percentage			*N**	Percentage		
		1-factor	**2-factor**	3-factor		1-factor	**2-factor**	3-factor
RMSEA	500.0	0.3%	**99.7%**	0.0%	500.0	38.9%	**61.1%**	0.0%
CFI	500.0	2.7%	**97.3%**	0.0%	500.0	4.3%	**95.7%**	0.0%
TLI	500.0	0.0%	**100.0%**	0.0%	500.0	0.3%	**99.7%**	0.0%
SRMR	500.0	100.0%	**0.0%**	0.0%	500.0	99.8%	**0.2%**	0.0%
p value	483.4	0.0%	**97.8%**	2.2%	483.0	0.0%	**96.8%**	3.2%
AIC	500.0	9.8%	**70.9%**	19.4%	500.0	0.0%	**59.5%**	40.5%
BIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.0%	**99.8%**	0.2%
SBIC	500.0	0.0%	**99.9%**	0.1%	500.0	0.0%	**96.3%**	3.7%
ΔRMSEA	312.4	0.0%	**96.6%**	3.4%	381.7	0.0%	**96.5%**	3.5%
ΔCFI	320.4	0.0%	**100.0%**	0.0%	317.8	0.0%	**99.6%**	0.4%
ΔSRMR	500.0	0.0%	**99.9%**	0.1%	499.8	0.0%	**99.7%**	0.3%
ΔAIC	500.0	0.0%	**91.6%**	8.4%	500.0	0.0%	**71.7%**	28.3%
ΔBIC	500.0	0.0%	**100.0%**	0.0%	500.0	0.0%	**99.8%**	0.2%

* *N* is the average of replications among nine sample size (3 cluster numbers × 3 cluster sizes), some of *N* were not equal to 500 due to convergency problems.

Discussion

This study is the first to examine the accuracy of commonly used model-fit indexes and selection criteria in identifying the correct number of factors in MEFA through simulation methods. Furthermore, the study combined two commonly used approaches (the model-based approach and the design-based approach) for the MEFA model.

Based on the model-based approach, most of the commonly used fit indexes were effective at identifying the correct number of factors at the lower level, except for $SRMR_{PS\_W}$ and AIC. In addition, sample size had little effect on accuracy. The SRMR finding was consistent with those of previous studies. Thus, Barendse and colleagues

(2015) proposed that SRMR was less correct and had a tendency to choose fewer latent factors for EFA. Similarly, Garrido et al. (2016) suggested that the performance of SRMR was worse than RMSEA and that CFI and TFI showed the greatest accuracy for categorical Monte Carlo simulation data. Other studies have pointed out that the performance of the IC family is related to sample size (Bozdogan, 2000; Lopes & West, 2004; Song & Belin, 2008). However, the present study revealed that BIC and SBIC performed better in extracting low-level numbers of factors than AIC and that they were not influenced by sample size. Moreover, when ICC increased from .1 to .2, the performance of both overall CFI and CFI_{PS_W} was no longer reliable for the lower-level factor extraction.

Alternatively, still based on the model-based approach, when extracting higher-level factors, most of the fit indexes and information criteria tended to extract fewer factors. That is, when there was only one Level-2 factor, most of the fit indexes performed well, but when there were two Level-2 factors, none of the fit indexes performed adequately. Furthermore, although the performance of the examined fit indexes and criteria was not satisfactory at the high level, when ICC was higher, the performance of most of them improved, except for RMSEA, CFI, and TLI. This is due to the fact that the between-model was easier to identify when ICC was higher. On the other hand, when the factor structure was invariant across levels and the design-based approach was adopted, most of the fit indexes were effective, except for SRMR, AIC, and ΔAIC. However, the assumption of invariant factor structure may not always be tenable in practice. Thus, application of this approach will face a challenge.

In general, the results of this study suggest that the model-selection-based approaches have the ability to identify the number of factors for the within-level. However, these approaches were not acceptable for the between-level, and they were directly related to ICC.

Overall, when a researcher has multilevel data without having an a priori hypothesis of a preconceived structure, MEFA is advised to apply over the subsequent steps. First, ICC requires to be calculated to decide whether the clustering effect should be considered or not. Second, researcher should apply the model-based approach to conduct the EFA at different levels and obtain the corresponding model indexes and IC values. Third, most of model indexes and IC values (except for $SRMR_{PS\_W}$, AIC, CFI, and CFI_{PS_W}) work well for deciding the number of factors in the within-level. On the other hand, most of the fit indexes and information criteria tended to extract fewer factors at the between-level.

Despite these promising findings, there are still some limitations of the study. First, we considered very few numbers of factors for both levels, whereas the latent factors were likely to exceed more than two. The study revealed that most of the fit indexes and criteria tended to underestimate the number of factors for the between level. Thus, when a study has a complicated between-level structure, it is hard to extract the true factors. Second, the performance of the indexes and criteria for the between-level was related to the ICC value, which made the results uncertain, so future studies are needed to address that. Third, sample sizes of generation data in this study were large, while in practical studies, sample sizes are not as large as the generation data. In addition, the cluster sizes were balanced, also in practice, they are often unbalanced and unequal from different clusters. Fourth, in addition to these general model-comparison methods, eigenvalue >1 and parallel analysis are also commonly used techniques in EFA (Ledesma & Valero-Mora, 2007; Yang & Xia, 2015). Thus, in future studies, we would like to use the eigenvalue-based approach to determine the number of factors in MEFA.

CHAPTER III

EIGENVALUES-BASED APPROACHES TO IDENTIFY THE OPTIMAL NUMBER OF FACTORS IN MULTILEVEL EXPLORATORY FACTOR ANALYSIS

Use of the Eigenvalues-Based Approaches in Exploratory Factor Analysis

The eigenvalue-based approaches are commonly used in selecting the optimal number of factors in traditional exploratory factor analysis (EFA). They include the following three criteria: (a) Kaiser (1960) recommended retaining the number of factors that is equal to the number when eigenvalues > 1; (b) Cattell (1966) proposed use of scree plot test to extract the quantity; more specifically, to find out the harshest fall-off among the eigenvalues' scree plot; finally (c) Horn (1965) proposed use of parallel analysis (PA), which compares two distinctive sources of eigenvalues, one from the observed correlation matrix and the other from an uncorrelated normally distributed variables matrix.

"Eigenvalues > 1" Criterion

It is well documented that the "eigenvalues > 1" criterion can result in overestimation of the number of factors when sample size is small or moderate; thus, this criterion is neither reliable nor accurate (Conway & Huffcutt, 2003; Crawford et al., 2010; Ford, MacCallum, & Tait, 1986; Gorsuch, 1997; Henson & Roberts, 2006; Horn, 1965; Schmitt, 2011; Zwick & Velicer, 1986). Additionally, when sample size is large and very close to the population level, the "eigenvalues > 0" criterion is more accurate than the "eigenvalues > 1" criterion in extracting factors because eigenvalues are obtained from a correlation matrix with squared multiple correlations (SMCs) between each variable and all other variables substituting the 1s along the diagonal (Crawford et al., 2010; Guttman, 1954). Overestimation of factors is due to the fact that the Kaiser criterion is based on an assumed population correlation matrix, while when we apply the "eigenvalues > 1" to a sample, sampling error influences the extraction results (Hayton, Allen, & Scarpello, 2004).

Despite these shortcomings, the Kaiser criterion remains a popular technique, and is still the default method of deciding the number of factors in many statistical software and packages (Henson & Roberts, 2006). Two papers have reviewed the percentage of studies using the "eigenvalues > 1" criterion to determine the factor number in the field of organizational research. Results showed that its use decreased from 21.7% in 1986 (Ford et al., 1986) to 15.4% in 2003 (Conway & Huffcutt, 2003). Another research article (Henson & Roberts, 2006) evaluating 60 published papers across four psychological journals found that 56.7% of the papers still adopted the Kaiser criterion for identifying the factor number. Further, examining the application of EFA in second-language research, Plonsky and Gonulal (2015) discovered that the "eigenvalues > 1" was the most common method (31.4%) of deciding the number of factors. Similarly, Sakaluk and Short (2017) summarized that 51.4% of sexuality research articles extracted their factors though the Kaiser criterion.

Scree Plot

The scree plot is a two-dimensional chart consisting of an x-axis and a y-axis, the former representing the components and the latter, the eigenvalues. Researchers try to determine the point at which the curve changes significantly, which, in turn, indicates the maximum number of factors to retain (Ledesma, Valero-Mora, & Macbeth, 2015).

The scree plot is also a widely applied eigenvalues-based criterion. For example, Henson and Roberts (2006) revealed that 35.0% of psychological papers used this test to decide factor numbers. Similarly, Sakaluk and Short (2017) noted that 32.9% of sexuality research articles chose the scree plot method for factor extraction. However, this plot test is a visual test; thus, its nature is subjective and ambiguous, especially when there is (a) no certain breaks or (b) more than one break (Hayton et al., 2004; Yang & Bozdogan, 2011; Zwick & Velicer, 1982). To illustrate, Ledesma, Valero-Mora, and Macbeth (2015) executed a test among six knowledgeable experts, who were given the same scree plot for telling the factor number, and found that the experts offered varying answers changing from one to four factors. Thus, the scree plot has low inter-rater reliability values (Crawford, & Koopman, 1979). Yet it has been demonstrated to be much more precise than the "eigenvalues > 1" criterion; in addition, it has a tendency to retain surplus factors (Henson & Roberts, 2006; Zwick & Velicer, 1986).

Parallel Analysis

Among these three criteria of the eigenvalues-based approaches, and even compared with the model-selection-based approaches, PA is the most recommended criterion when deciding the quantity of latent factors for both continuous and categorical data (Cho, Li, & Bandalos, 2009; Dinno, 2009; Glorfeld, 1995; Henson & Roberts, 2006; Ledesma & Valero-Mora, 2007; Schmitt, 2011; Weng & Cheng, 2005; Yang & Xia, 2015). PA is a sample-based modification of the population-based "eigenvalues >

1" criterion, because eigenvalues of the correlation matrix of uncorrelated variables are all 1.0 at the population level (Horn, 1965).

The assumption underlying PA is that nontrivial factors from measured data with a valid factor structure have bigger eigenvalues than parallel factors from random data with the same number of variables and sample size (Hayton et al., 2004; Lautenschlager, 1989). To be more specific, PA generates a series of random datasets that have the same sample size and the same number of variables as the measured data set, and PA variables are uncorrelated to each other (Horn, 1965). If the nontrivial factors exist, then eigenvalues of the measured data should be larger than the randomly generated eigenvalues (Schmitt, 2011). Thus, PA combines the "eigenvalues > 1" criterion with the effect of sample size (Zwick & Velicer, 1986). Of special importance, PA is not sensitive to distributional form when data are under the assumptions of independent and identical distribution and is robust to non-normal distributions of a random sample (Dinno, 2009; Glorfeld, 1995). Moreover, many scholars have recommended the use of the Polychoric correlation instead of the Pearson correlation in PA when items are ordinal/noncontinuous data (Cho et al., 2009; Timmerman & Lorenzo-Seva, 2011). However, PA has a tendency to overestimate the numbers (Glorfeld, 1995; Zwick & Velicer, 1986), but when the latent factors are highly correlated (correlation ≥ .5), it tends to under-extract the latent numbers (Yang & Xia, 2015). Meanwhile, PA can be used to improve the scree plot test through adding PA results to the scree plot (Ledesma et al., 2015). Although PA has been recommended for a long time, unfortunately, it is rarely employed in published papers, with only 6.7% applications in psychological research (Henson & Roberts, 2006) and 5.1% in sexuality research (Sakaluk & Short, 2017).

In recent years, multilevel research studies have become more popular in many areas, such as the social and behavioral sciences, including educational studies (Mayberry, Espelage, & Koenig, 2009; Moore et al., 2016; Reise, Ventura, Nuechterlein, & Kim, 2005; D'Haenens, Van Damme, & Onghena, 2010). This seems fitting because, for instance, students are nested within classes, classes are nested within schools, and schools are further nested within school districts. In order to get accurate statistical results, we need to consider the structure of the multilevel data and the impact of clustering. After reviewing previous papers, we found that simulation studies of the eigenvalues-based approaches comparison have been developed in single-level EFA. Thus, the target of this study was to extend the eigenvalues-based approaches to multilevel EFA (MEFA) and compare the abilities of several eigenvalues' computation/extraction techniques. As the crucial step in the eigenvalues-based approaches, eigenvalue calculation contains many extraction techniques, such as Principal Component Analysis (PCA), Principal Axis Factoring (PAF), Maximum Likelihood (ML), and Unweighted Least Squares (ULS) (Costello & Osborne, 2005).

This paper first reviewed the eigenvalues extraction/calculation techniques and then explained the use of Monte Carlo simulation for MEFA data. It then provided the results of applying the eigenvalues-based approaches (i.e., the "eigenvalues > 1" criterion and PA method) to retaining the factor number for MEFA data.

Eigenvalues Extraction Techniques

The key issue in the eigenvalues-based approaches is to calculate eigenvalues. Many common statistical software applications, such as SPSS, STATA, and SAS, provide several extraction techniques for eigenvalues computation, including principal component analysis (PCA), principal axis factoring (PAF), iterated principal axis factor (Iterated PAF), alpha factoring, maximum likelihood (ML), unweighted least squares (ULS), generalized least squares (GLS), image factoring, and Harris extraction (Costello & Osborne, 2005; Osborne & Banjanovic, 2016). Of these, PCA, PAF, and ML are the most commonly used (De Winter & Dodou, 2012; Osborne & Banjanovic, 2016; Schmitt, 2011).

PCA

Strictly speaking, PCA is not a latent-variable method of computing eigenvalues for deciding the optimal number of factors (Howard, 2016; Keith, Caemmerer, & Reynolds, 2016; Widaman, 2007). Thus, the purpose of PCA is merely to decrease the number of variables by generating linear combinations that extract as many of the original observed variances as possible, which is different from the purpose of the latent-variable method, which is to detect the latent factors that explain relationships among overserved variables (Conway & Huffcutt, 2003). The latent-variable method separates common variance and unique variance within observed variables, but PCA does not distinguish between these two forms of variances and only utilizes components to represent these two variances (Howard, 2016). Thus, PCA is a component model, not a common factor model. Nevertheless, it is one of the most commonly used factor-extraction techniques for EFA (Schmitt, 2011), and is the default factor-extraction techniques in many statistical software packages, such as SPSS (Costello & Osborne, 2005).

PAF

Many previous research studies have used PA based on PCA (Crawford et al., 2010; Keith, Caemmerer, & Reynolds, 2016; Timmerman & Lorenzo-Seva, 2011). One study noted that PA-PCA was not the "gold standard" approach for choosing the optimal factor numbers for the intelligence data (Keith, Caemmerer, & Reynolds, 2016). Yet another simulation study claimed that PA based on PCA can perform as well as or better than PAF for models with one factor or multiple minimally correlated ($r < .7$) factors (Crawford et al., 2010).

The other extraction techniques (PAF, Iterated PAF, ML, ULS, GLS, Image Factoring, and Harris Extraction) are all based on a common factor model, whose aim is to understand the latent factors that explain the relationships among measured variables (Conway & Huffcutt, 2003). As a variation of PCA, PAF substitutes the diagonal elements of the correlation or covariance matrix with the initial communality estimates, or initial estimates of the shared variances (Osborne & Banjanovic, 2016). PAF uses least-squares (Unweighted Least Squares [ULS] or Ordinary Least Squares [OLS]) to estimate the common factor model (De Winter & Dodou, 2012). Since the detailed mathematics involved are automated in modern statistical programs (Howard, 2016), the present study does not elaborate on that.

Empirical researchers believe that PA based on PAF can over-extract the latent factors (Buja & Eyuboglu, 1992; Steger, 2006), which has also been proven in a simulation study (Timmerman & Lorenzo-Seva, 2011). However, other simulation studies found that PAF yielded more accurate information about the number of factors than PCA (Crawford et al., 2010; Green, Levy, Thompson, Lu, & Lo, 2012; Keith et al., 2010).

Iterated PAF is similar to PAF, except that an iterative assessment procedure is utilized, which means individual sequential estimation of the communality is used to replace the diagonal of the matrix of associations (Costello & Osborne, 2005). The iterative process continues until the communality estimates become stable or their variations are less than a preset threshold (Keith et al., 2010).

ML

ML attempts to infer from a sample of individuals to an entire population because it is an asymptotically efficient estimator (De Winter & Dodou, 2012). Thus, it seeks to extract factors and estimate parameters that optimally produce the population correlation (or variance-covariance) matrix (Costello & Osborne, 2005). Additionally, ML determines which factor loading and unique variance estimations are most likely to replicate the observed data (De Winter & Dodou, 2012).

Two approaches based on ML may be used to calculate eigenvalues: the model-based approach (i.e., a level-specific model is specified for each level) and the design-based approach (i.e., it assumes factorial invariance across levels and considers the multilevel structure by correcting the standard errors taking the sampling design into account) (Wu & Kwok, 2012). (More details about these two approaches will be presented under Methods below.)

An important assumption behind ML is that variables are multivariate normal with linear interrelationships, whereas PCA or PAF does not need to have this assumption (De Winter & Dodou, 2012; Howard, 2016). ML is affected by sample size; that is, the accuracy of ML increases when sample size increases (Fabrigar, Wegener, MacCallum, & Strahan, 1999). When ML is used to decide the number of factors for EFA, PA with ML procedures tend to overestimate the factor number (Humphreys & Montanelli, 1975). However, De Winter and Dodou (2012) suggested that through simulation and empirical studies, ML performs better than PAF when unequal loadings exist in factors and for under-extraction of latent factors, because compared with PAF, ML is less able to extract the weaker factors because it puts less weight on the weaker correlations.

Alpha Extraction, ULS, Image Factoring, and Harris Extraction

Alpha Extraction attempts to maximize Cronbach's alpha assessment of the reliability of the latent factors (Costello & Osborne, 2005). ULS, in turn, does not make any requirements about normality and seeks to minimize the measurement errors and operationalizes as the sum of squared residuals. Finally, Image Factoring and Harris Extraction are derived from the image factor model, as opposed to the common factor model or the component model (Cattell & Vogelmann, 1977).

Discussions of the relative advantages and disadvantages of these eigenvalues' extraction techniques are rare based on multilevel data. Thus, the present study evaluated four different commonly used factor-extraction techniques (PCA, PAF, Iterated PAF, and ML) for deciding the number of factors in MEFA.

Combination of Criteria

In addition to the above, the present study also considered a combination of three eigenvalues-based criteria (eigenvalues > 1, scree plot test, and PA) and four eigenvalues extraction techniques (PCA, PAF, Iterated PAF, and ML). In order to simplify the results and make the conclusions more practical, the study first excluded the scree plot test, since its results are naturally subjective and uncertain, and have a low inter-rater reliability (Crawford & Koopman, 1979). Next, because the performance of the eigenvalues > 1 criterion is not satisfactory in EFA (Crawford et al., 2010; Henson & Roberts, 2006), the study only examined the commonly used ML and PCA extraction techniques under this criterion for MEFA. Finally, the study mainly focused on the PA criterion. Although a previous study showed that PA based on PCA performed well for EFA (Crawford et al., 2010), the present study also excluded PA because PCA is not a latent method (Howard, 2016). Consequently, this study presents the results of PAF, Iterated PAF, and ML extraction techniques under PA for MEFA.

Methods

In this study, I used the same simulation design as Study 1 and the details of the data generation and design factors are restated here.

Multilevel Data Generation

In this study, Monte Carlo simulations were conducted using M*plus* version 8 (Muthén, & Muthén, 2017). Two-level data with eight items (y_1-y_8) were generated based on two scenarios through M*plus*. These eight items were all the standard normal distribution with mean = 0 and standard deviation (SD) = 1. Thus, multivariate normality was assumed for both scenarios, and all factor loadings were the standardized loadings.

Scenario 1

In this scenario (see Figure 1), Level-1 and Level-2 had different factor structures, with two factors in Level-1 and one factor in Level-2. Specifically, in Level-1, y_1-y_4 were under Factor 1 and y_5-y_8 were under Factor 2. Within each factor, the first three items were good (main factor loadings = .6 and cross-loadings = .1), and the fourth item was a poor item (main factor loadings and cross-loadings were all .25). The correlation between the two within factors was .5. In Level-2, y_1-y_8 were all under one factor with the first six items being good items (factor loadings = .6) and the last two items being poor items (factor loadings = .25).

Scenario 2

In this scenario (see Figure 2), Level-1 and Level-2 had the same two-factor structures. Specifically, the Level-1 and Level-2 factor structures and factor loadings were the same as Level-1 in Scenario 1. The correlations between the two factors (within and between) were both .5. Because previous reviews have shown that when the correlation between factors is high ($r \geq .5$), the model fit statistics commonly extract fewer than the true latent factors. We wanted to find the best way of identifying the optimal number of factors when the correlation is high.

Design Factors

The data generation also involved different sample size and intraclass correlation coefficient (ICC) conditions. ICC conditions are commonly a consideration in multilevel structural equation modeling (MSEM) simulation studies. For example, Hsu, Lin, Kwok, Acosta, and Willson (2017) considered the generalization of study finding and simulated six different ICC conditions (.1, .2, .3, .4, .5, and 1.00). Specifically, in the present study, the design factors included cluster numbers (CN = 100, 200, and 500), cluster sizes (CS =10, 20, and 50), and ICCs (.1, .2). Meanwhile, 500 replications were produced for each combination of design factors in the two scenarios. As a result, a total of 2 (scenarios) × 3 (cluster numbers) × 3 (cluster sizes) × 2 (ICC) × 500 = 36 (Conditions) × 500 = 18,000 replications were generated.

Parallel Analysis Random Data Generation

M*plus* (Muthén & Muthén, 2017) was also used to generate PA random datasets of standard normal variables for each condition with 1,000 replications, for a total of 36 (conditions) × 1000 = 36,000 replications. For each condition, 1,000 PA random datasets were generated based on the specific sample sizes and ICCs, y_1-y_8 are uncorrelated normally distributed variables for within- and between-levels. Thus, PA random datasets had the same three cluster numbers (CN = 100, 200, and 500), three cluster sizes (CS =10, 20, and 50), and two ICCs (.1, .2) as previous generated multilevel datasets. The within-level PA datasets were used to identify the number of factors in the within-level, while the between-level PA datasets were utilized to identify in the between-level.

Eigenvalues Calculation

M*plus* and SAS 9.4 (SAS Institute Inc., 2002) were used to analyze the data and calculate eigenvalues for different eigenvalues-based criteria.

First, the study examined the eigenvalue > 1 criterion in M*plus*. There are two approaches to calculate eigenvalues based on ML extraction technique for MEFA analysis: the model-based approach and the design-based approach. In the former, a level-specific model is specified for each level (Wu & Kwok, 2012). In M*plus*, the syntax is TYPE = TWOLEVEL. When estimating the Level-1 factor structure, the Level-2 model is saturated and vice versa. In addition, the default setting for the estimators in M*plus* was ML (maximum likelihood parameter estimates with conventional standard errors and chi-square test statistic), and for the rotation it was GEOMIN oblique rotation (Muthén & Muthén, 2017).

On the other hand, the design-based approach assumes factorial invariance across levels and accounts for the multilevel structure by correcting the standard errors according to the sampling design (Wu & Kwok, 2012). In M*plus*, the syntax is TYPE = COMPLEX. Other settings were also the default; the estimator was MLR (maximum likelihood parameter estimates with robust standard errors and a chi-square test statistic, MLR is robust to nonnormality and nonindependence of observations), and the rotation was GEOMIN oblique rotation (Muthén & Muthén, 2010). MLR was used because MLR standard errors are calculated using a sandwich estimator. In addition, the MLR chi-square test statistic is asymptotically equivalent to T2*test statistic proposed by Yuan and Bentler (2000). Compared to the model-based approach, the design-based approach is relatively simple. That is, it only analyzes one global model and estimates parameters based on that model, which means it assumes the same structure across different levels; however, sometimes this assumption is wrong (Wu & Kwok, 2012). Thus, in Scenario 1, we only used the model-based approach because the assumption of factorial invariance across levels was violated. In Scenario 2, we applied both the model-based and the design-based approaches for the data analysis. In terms of the results of these two approaches, the model-based approach (M*plus* using ML) provided eigenvalues of between- and within- correlation matrices for sample separately, whereas the design-based approach (M*plus* using MLR) only offered eigenvalues of one combined matrix for both levels. The study used these above extracted eigenvalues to evaluate the "eigenvalues > 1" criterion for multilevel exploratory analysis.

In the following steps, the study changed analysis software from M*plus* to SAS in order to compute eigenvalues using different extraction techniques, because SAS provides all of the above-mentioned eigenvalues extraction techniques (PCA, PAF, Iterated PAF, ML, etc.; Osborne & Banjanovic, 2016). M*plus* cannot provide all of them. However, because SAS does not provide a direct syntax to analyze MEFA, M*plus* was first applied to obtain the within- and between-correlation matrices separately for the multilevel datasets. Again, the syntax in M*plus* is TYPE = TWOLEVEL, and other settings also served as the default. After getting these matrices, SAS was manipulated to analyze and calculate eigenvalues through different extraction techniques (PCA, PAF, Iterated PAF, and ML). In SAS, the default critical eigenvalue is 1, so under PROC FACTOR, SAS retains and rotates any component whose eigenvalue is 1 or larger. In addition, ML estimation in M*plus* and SAS provides similar eigenvalues, making it possible to compare the results of these two packages.

Parallel Analysis

The study first obtained the within- and between-correlation matrices separately for two levels through M*plus* from generated PA random data, and then used PAF, Iterated PAF, and ML extraction techniques through SAS to calculate PA eigenvalues. In each condition of the PA random data, we then calculated average values of eigenvalues of the 1,000 replications for each factor. Thus, under different extraction techniques, this study compared each factor's eigenvalues from both measured and random data at different levels; specifically, the first factor's eigenvalue from simulated observed data was compared with the first factor's average eigenvalue from random data; 0 was retained if the latter value was larger; otherwise, 1 was retained. We did a similar comparison and retained the number for the remaining seven factors until the latter value was larger, and thus finally found the factor numbers at both the within- and between-levels.

Results

"Eigenvalues > 1" Criterion

"Eigenvalues > 1" Criterion Based on the ML/MLR Extraction Technique

Model-based approach (Level-1 factor extraction). Table 6 shows the percentages of 500 iterations that retained the factor number for the Level-1 model of the model-based approach using the "eigenvalues > 1" criterion based on ML extraction technique. The true model was a two-factor model at the within-level for the two scenarios. In general, the "eigenvalues > 1" criterion based on ML extracted a one-factor model at Level-1, and the chances of erroneously extracting underestimated factor number (only one factor) were more than 80%. Furthermore, there was a tendency for the under-factor situation to become worse when ICC was .2 compared to .1, holding other settings the same. Another finding, as shown in Table 6, was that the consequences were very similar for both scenarios when holding the ICC and sample size equal.

Table 6 Level-1 Factor Extraction Results for the Model-Based Approach Using the "Eigenvalues > 1" Criterion Based on the ML Extraction Technique.

ICC	Sample Size*	Percentage			
		Scenario 1		Scenario 2	
		1-factor	**2-factor**	1-factor	**2-factor**
0.1	100(10)	83.0%	**17.0%**	81.6%	**18.4%**
0.2	100(10)	98.2%	**1.8%**	98.2%	**1.8%**
0.1	100(20)	94.4%	**5.6%**	94.4%	**5.6%**
0.2	100(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	100(50)	99.4%	**0.6%**	99.8%	**0.2%**
0.2	100(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(10)	93.8%	**6.2%**	95.0%	**5.0%**
0.2	200(10)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(20)	98.6%	**1.4%**	99.6%	**0.4%**
0.2	200(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.2	200(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(10)	99.8%	**0.2%**	100.0%	**0.0%**
0.2	500(10)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(20)	99.8%	**0.2%**	100.0%	**0.0%**
0.2	500(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.2	500(50)	100.0%	**0.0%**	100.0%	**0.0%**

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Model-based approach (Level-2 factor extraction). Tables 7 and 8 present the percentages of 500 iterations that retained different factors for the Level-2 model for the model-based approach in Scenario 1 and Scenario 2, respectively, using the "eigenvalues > 1" criterion based on the ML extraction technique. The true factor numbers in the between-level for the two scenarios were different; that is, it was a one-factor model in Scenario 1 and a two-factor model in Scenario 2.

Table 7 Level-2 Factor Extraction Results for the Model-Based Approach in Scenario 1 Using the "Eigenvalues > 1" Criterion Based on the ML Extraction Technique.

ICC	Sample Size*	Percentage				
		1-factor	2-factor	3-factor	4-factor	5-factor
0.1	100(10)	**0.8%**	31.6%	59.8%	7.6%	0.2%
0.2	100(10)	**3.2%**	50.8%	43.4%	2.6%	0.0%
0.1	100(20)	**2.0%**	51.4%	43.6%	3.0%	0.0%
0.2	100(20)	**5.4%**	61.2%	33.2%	0.2%	0.0%
0.1	100(50)	**4.8%**	65.4%	29.4%	0.4%	0.0%
0.2	100(50)	**7.8%**	72.4%	19.6%	0.2%	0.0%
0.1	200(10)	**2.2%**	58.8%	37.2%	1.8%	0.0%
0.2	200(10)	**9.0%**	76.2%	14.8%	0.0%	0.0%
0.1	200(20)	**7.0%**	72.4%	20.6%	0.0%	0.0%
0.2	200(20)	**14.4%**	78.0%	7.6%	0.0%	0.0%
0.1	200(50)	**15.4%**	80.0%	4.6%	0.0%	0.0%
0.2	200(50)	**21.6%**	75.4%	3.0%	0.0%	0.0%
0.1	500(10)	**14.0%**	76.2%	9.8%	0.0%	0.0%
0.2	500(10)	**35.0%**	64.2%	0.8%	0.0%	0.0%
0.1	500(20)	**31.4%**	66.2%	2.4%	0.0%	0.0%
0.2	500(20)	**51.8%**	48.0%	0.2%	0.0%	0.0%
0.1	500(50)	**51.0%**	48.8%	0.2%	0.0%	0.0%
0.2	500(50)	**61.2%**	38.8%	0.0%	0.0%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

In Scenario 1 (see Table 7), under most simulation conditions, the Kaiser criterion based on ML was not able to identify the correct latent factor numbers. Moreover, this criterion had a large probability (over 75%) of overestimating the between-level factor number except under three simulation conditions: one condition with cluster number = 500, cluster size = 20, and ICC = .2, and two conditions with cluster number = 500 and cluster size = 50. When sample sizes were the same, the larger ICC helped to increase the chance of extracting the true latent numbers.

In Scenario 2 (see Table 8), in the majority of cases, the "eigenvalues > 1" based on ML had a probability of more than 50% of extracting the correct factor numbers except for three cases, two cases with cluster number = 100 and cluster size = 10, and one case with cluster number = 500, cluster size = 10, and ICC = .1. The results were similar to Scenario 1 in that the larger ICC can help to increase the chance of identifying the correct numbers at the between-level.

In general, the "eigenvalues > 1" criterion based on the ML extraction technique performed better in detecting the factor numbers at the between-level than at the within-level. Higher cluster numbers or more cluster size or larger ICC can help to increase the chance of identifying the correct factor numbers at the between-level.

Table 8 Level-2 Factor Extraction Results for the Model-Based Approach in Scenario 2 Using the "Eigenvalues > 1" Criterion Based on the ML Extraction Technique.

ICC	Sample Size*	Percentage			
		1-factor	**2-factor**	3-factor	4-factor
0.1	100(10)	1.8%	**41.0%**	50.0%	7.2%
0.2	100(10)	1.4%	**45.2%**	50.2%	3.2%
0.1	100(20)	6.6%	**61.6%**	30.6%	1.2%
0.2	100(20)	1.2%	**61.2%**	35.8%	1.8%
0.1	100(50)	14.4%	**71.2%**	14.4%	0.0%
0.2	100(50)	2.4%	**71.6%**	25.8%	0.2%
0.1	200(10)	9.6%	**61.4%**	28.6%	0.4%
0.2	200(10)	2.6%	**71.8%**	25.2%	0.4%
0.1	200(20)	19.2%	**70.4%**	10.4%	0.0%
0.2	200(20)	1.2%	**85.8%**	13.0%	0.0%
0.1	200(50)	32.8%	**65.0%**	2.2%	0.0%
0.2	200(50)	1.4%	**91.4%**	7.2%	0.0%
0.1	500(10)	31.0%	**65.0%**	4.0%	0.0%
0.2	500(10)	0.8%	**96.8%**	2.4%	0.0%
0.1	500(20)	47.8%	**51.8%**	0.4%	0.0%
0.2	500(20)	0.4%	**99.0%**	0.6%	0.0%
0.1	500(50)	63.4%	**36.6%**	0.0%	0.0%
0.2	500(50)	0.0%	**99.6%**	0.4%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Design-based approach (Scenario 2). Table 9 exhibits the percentages of 500 iterations that retained different factors for the design-based approach in Scenario 2 using the "eigenvalues > 1" criterion based on the MLR extraction technique. The true model was a two-factor model for both levels. In general, the Kaiser criterion had a great chance (more than 75%) to underestimate the factor number for the multilevel data. It was not appropriate for identifying the true factor number when using the design-based approach through the "eigenvalues > 1" criterion based on the MLR extraction technique.

Table 9 Factor Extraction Results for the Design-Based Approach in Scenario 2 Using the "Eigenvalues > 1" Criterion Based on the MLR Extraction Technique.

ICC	Sample Size*	Percentage	
		1-factor	**2-factor**
0.1	100(10)	83.4%	**16.6%**
0.2	100(10)	76.4%	**23.6%**
0.1	100(20)	93.6%	**6.4%**
0.2	100(20)	86.0%	**14.0%**
0.1	100(50)	98.8%	**1.2%**
0.2	100(50)	91.0%	**9.0%**
0.1	200(10)	94.2%	**5.8%**
0.2	200(10)	88.4%	**11.6%**
0.1	200(20)	98.4%	**1.6%**
0.2	200(20)	94.6%	**5.4%**
0.1	200(50)	99.8%	**0.2%**
0.2	200(50)	97.2%	**2.8%**
0.1	500(10)	99.6%	**0.4%**
0.2	500(10)	98.4%	**1.6%**
0.1	500(20)	100.0%	**0.0%**
0.2	500(20)	99.4%	**0.6%**
0.1	500(50)	100.0%	**0.0%**
0.2	500(50)	99.8%	**0.2%**

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Altogether, the "eigenvalues > 1" criterion based on the ML/MLR eigenvalues extraction technique was not accurate in retaining the latent factor number for multilevel exploratory factor analysis. This criterion resulted in an underestimation of the factor numbers at the within-level, regardless of whether the model-based approach or the design-based approach was used. Although this criterion can extract the true factor numbers at the between-level in some conditions, we need to avoid using it, because in practical studies, determining the Level-1 factor numbers is usually the first vital step; hence, inaccurate low-level factor numbers can lead to poor models and further incorrect interpretation even if the higher levels have the correct factor numbers.

In sum, this study demonstrated that the "eigenvalues > 1" criterion based on the ML/MLR eigenvalues extraction technique cannot be used to retain the factor numbers for multilevel data. In addition, under the model-based approach, ICC had different influences on this criterion; that is, a larger ICC had a negative effect on determining the true numbers at the within-level but a positive effect at the between-level, while larger sample sizes always had a positive influence in the Level-2 model for this criterion.

"Eigenvalues > 1" Criterion Based on the PCA Extraction Technique

PCA (Level-1 factor extraction). Table 10 shows the percentages of 500 iterations that retained the factor number for the Level-1 model using the "eigenvalues > 1" criterion based on the PCA extraction technique. The results were similar to the findings shown in Table 6, which also used the "eigenvalues > 1" criterion to extract the within-level factor. In general, the "eigenvalues > 1" criterion based on PCA also underestimated the within-level factor numbers, and higher ICC led to worse results.

Table 10 Level-1 Factor Extraction Results Using the "Eigenvalues > 1" Criterion Based on the PCA Extraction Technique.

ICC	Sample Size*	Percentage			
		Scenario 1		Scenario 2	
		1-factor	**2-factor**	1-factor	**2-factor**
0.1	100(10)	83.8%	**16.2%**	82.4%	**17.6%**
0.2	100(10)	98.4%	**1.6%**	98.4%	**1.6%**
0.1	100(20)	94.4%	**5.6%**	94.8%	**5.2%**
0.2	100(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	100(50)	99.6%	**0.4%**	99.8%	**0.2%**
0.2	100(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(10)	93.8%	**6.2%**	95.0%	**5.0%**
0.2	200(10)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(20)	98.8%	**1.2%**	99.6%	**0.4%**
0.2	200(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	200(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.2	200(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(10)	99.8%	**0.2%**	100.0%	**0.0%**
0.2	500(10)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(20)	99.8%	**0.2%**	100.0%	**0.0%**
0.2	500(20)	100.0%	**0.0%**	100.0%	**0.0%**
0.1	500(50)	100.0%	**0.0%**	100.0%	**0.0%**
0.2	500(50)	100.0%	**0.0%**	100.0%	**0.0%**

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

PCA (Level-2 factor extraction). Tables 11 and 12 present the percentages of 500 iterations that retained the different factors for the Level-2 model in Scenario 1 and Scenario 2, respectively, through the "eigenvalues > 1" criterion based on the PCA extraction technique. Results shown in these two tables are analogous to those shown in Tables 7 and 8, respectively. Briefly, the "eigenvalues > 1" criterion based on PCA behaved better in identifying the latent factor in Level-2 than in Level-1. In addition, a larger sample size or greater ICC can help to increase the likelihood of identifying the correct latent factor(s) in Level-2.

Table 11 Level-2 Factor Extraction Results in Scenario 1 Using the "Eigenvalues > 1" Criterion Based on the PCA Extraction Technique.

ICC	Sample Size*	Percentage				
		1-factor	2-factor	3-factor	4-factor	5-factor
0.1	100(10)	**0.6%**	31.6%	60.0%	7.6%	0.2%
0.2	100(10)	**3.2%**	50.4%	43.8%	2.6%	0.0%
0.1	100(20)	**1.8%**	52.0%	43.2%	3.0%	0.0%
0.2	100(20)	**5.4%**	61.6%	32.8%	0.2%	0.0%
0.1	100(50)	**4.6%**	65.4%	29.6%	0.4%	0.0%
0.2	100(50)	**7.8%**	73.4%	18.6%	0.2%	0.0%
0.1	200(10)	**2.4%**	59.2%	36.8%	1.6%	0.0%
0.2	200(10)	**9.6%**	76.0%	14.4%	0.0%	0.0%
0.1	200(20)	**7.2%**	72.4%	20.4%	0.0%	0.0%
0.2	200(20)	**14.8%**	77.6%	7.6%	0.0%	0.0%
0.1	200(50)	**15.0%**	80.8%	4.2%	0.0%	0.0%
0.2	200(50)	**21.8%**	75.0%	3.2%	0.0%	0.0%
0.1	500(10)	**13.6%**	77.0%	9.4%	0.0%	0.0%
0.2	500(10)	**36.0%**	63.0%	1.0%	0.0%	0.0%
0.1	500(20)	**32.2%**	65.4%	2.4%	0.0%	0.0%
0.2	500(20)	**52.2%**	47.8%	0.0%	0.0%	0.0%
0.1	500(50)	**51.8%**	48.0%	0.2%	0.0%	0.0%
0.2	500(50)	**62.2%**	37.8%	0.0%	0.0%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 12 Level-2 Factor Extraction Results in Scenario 2 Using the "Eigenvalues > 1" Criterion Based on the PCA Extraction Technique.

ICC	Sample Size*	Percentage			
		1-factor	**2-factor**	3-factor	4-factor
0.1	100(10)	1.8%	**40.8%**	50.8%	6.6%
0.2	100(10)	1.4%	**45.0%**	50.8%	2.8%
0.1	100(20)	6.6%	**61.6%**	30.6%	1.2%
0.2	100(20)	1.2%	**60.8%**	36.2%	1.8%
0.1	100(50)	14.2%	**71.0%**	14.8%	0.0%
0.2	100(50)	2.4%	**72.0%**	25.4%	0.2%
0.1	200(10)	9.0%	**62.0%**	28.6%	0.4%
0.2	200(10)	2.2%	**72.6%**	24.8%	0.4%
0.1	200(20)	19.2%	**70.0%**	10.8%	0.0%
0.2	200(20)	1.2%	**86.2%**	12.6%	0.0%
0.1	200(50)	33.4%	**64.8%**	1.8%	0.0%
0.2	200(50)	1.4%	**91.8%**	6.8%	0.0%
0.1	500(10)	31.2%	**65.0%**	3.8%	0.0%
0.2	500(10)	0.8%	**96.8%**	2.4%	0.0%
0.1	500(20)	48.4%	**51.2%**	0.4%	0.0%
0.2	500(20)	0.4%	**99.0%**	0.6%	0.0%
0.1	500(50)	64.6%	**35.4%**	0.0%	0.0%
0.2	500(50)	0.0%	**99.6%**	0.4%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Overall, the "eigenvalues > 1" criterion using the ML or PCA eigenvalues extraction technique did not correctly extract the latent factors for MEFA.

Parallel Analysis

Parallel Analysis Based on the PAF Extraction Technique

PA based on PAF (Level-1 factor extraction). Table 13 displays the percentages of 500 iterations that retained the latent factors for the Level-1 model through PA based on the PAF extraction technique. As mentioned previously, the true Level-1 factor number was two, regardless of the scenario. PA based on the PAF method extracted the true factor model in Level-1; that is, there was more than a 98% chance of accurately extracting within-level factor numbers.

Table 13 Level-1 Factor Extraction Results Using PA Based on the PAF Extraction Technique.

ICC	Sample Size*	Percentage					
		Scenario 1			Scenario 2		
		1-factor	**2-factor**	3-factor	1-factor	**2-factor**	3-factor
0.1	100(10)	0.0%	**99.6%**	0.4%	0.2%	**98.2%**	1.6%
0.2	100(10)	0.2%	**98.2%**	1.6%	0.2%	**98.4%**	1.4%
0.1	100(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	100(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	100(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	100(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	200(10)	0.0%	**100.0%**	0.0%	0.0%	**99.8%**	0.2%
0.2	200(10)	0.0%	**100.0%**	0.0%	0.0%	**99.4%**	0.6%
0.1	200(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	200(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	200(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	200(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	500(10)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	500(10)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	500(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	500(20)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.1	500(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%
0.2	500(50)	0.0%	**100.0%**	0.0%	0.0%	**100.0%**	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

PA based on PAF (Level-2 factor extraction). Table 14 presents the percentages of 500 iterations that retained different factors for the Level-2 model in Scenario 1 using PA based on the PAF eigenvalues extraction method. As mentioned previously, the true Level-2 factor number was one for this scenario. PA based on the PAF method identified the accurate between-level factor model in Level-2 with more than a 65% probability in every simulation condition for this study. Larger ICC or greater cluster number or bigger cluster size helped to increase the probability of extracting the correct latent numbers for Level-2 when controlling for other conditions.

Table 14 Level-2 Factor Extraction Results in Scenario 1 Using PA Based on the PAF Extraction Technique.

ICC	Sample Size*	Percentage								
		0-factor	**1-factor**	2-factor	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	0.8%	**65.2%**	12.6%	4.0%	3.4%	1.2%	0.6%	0.0%	12.2%
0.2	100(10)	0.0%	**73.2%**	14.8%	5.8%	2.8%	2.0%	0.2%	0.2%	1.0%
0.1	100(20)	0.0%	**67.0%**	19.2%	7.4%	2.6%	1.0%	1.0%	0.6%	1.2%
0.2	100(20)	0.0%	**76.8%**	16.6%	4.0%	1.4%	0.8%	0.2%	0.2%	0.0%
0.1	100(50)	0.0%	**73.0%**	18.8%	5.4%	1.4%	0.6%	0.6%	0.2%	0.0%
0.2	100(50)	0.0%	**80.4%**	16.4%	2.6%	0.2%	0.4%	0.0%	0.0%	0.0%
0.1	200(10)	0.0%	**70.4%**	17.8%	6.0%	1.4%	1.6%	0.2%	1.0%	1.6%
0.2	200(10)	0.0%	**81.4%**	13.6%	3.8%	1.2%	0.0%	0.0%	0.0%	0.0%
0.1	200(20)	0.0%	**76.2%**	17.2%	4.8%	1.4%	0.2%	0.2%	0.0%	0.0%
0.2	200(20)	0.0%	**87.8%**	10.8%	1.2%	0.2%	0.0%	0.0%	0.0%	0.0%
0.1	200(50)	0.0%	**82.8%**	15.2%	1.8%	0.2%	0.0%	0.0%	0.0%	0.0%
0.2	200(50)	0.0%	**85.2%**	13.0%	1.6%	0.2%	0.0%	0.0%	0.0%	0.0%
0.1	500(10)	0.0%	**80.6%**	16.0%	2.4%	0.8%	0.2%	0.0%	0.0%	0.0%
0.2	500(10)	0.0%	**90.4%**	9.2%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%
0.1	500(20)	0.0%	**88.2%**	10.8%	0.8%	0.2%	0.0%	0.0%	0.0%	0.0%
0.2	500(20)	0.0%	**92.2%**	7.6%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
0.1	500(50)	0.0%	**91.8%**	8.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%
0.2	500(50)	0.0%	**93.8%**	6.0%	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 15 presents the percentages of 500 iterations that retained different factors for the Level-2 model in Scenario 2 using PA based on the PAF eigenvalues extraction method. Under this scenario, the true between-level model was a two-factor model. However, this time it performed worse than in Scenario 1. That is, although PA based on

PAF still had the capability to extract the correct factor number in most simulation conditions, the possibility was lower than 65% when the cluster number =100. Moreover, it was similar to previous results, in that larger ICC or greater cluster numbers or bigger cluster size improved performance.

In general, the performance of the PA criterion based on the PAF extraction technique was almost perfect at detecting the between-level factors. If the structure became complicated, as going from one factor in Scenario 1 to two factors in Scenario 2, more cluster numbers were needed to extract the correct latent number. Under Scenario 2, if the sample size was large enough (cluster number = 500), the performance was acceptable (the percentage of correctly identifying the number was large than 75%). In other words, bigger cluster numbers or cluster sizes or larger ICC can help to increase the chance of identifying the correct factor numbers in the between-level through PA based on the PAF eigenvalues extraction method.

Table 15 Level-2 Factor Extraction Results in Scenario 2 Using PA Based on the PAF Extraction Technique.

ICC	Sample Size*	Percentage								
		0-factor	1-factor	**2-factor**	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	0.6%	63.2%	**16.4%**	5.6%	3.2%	1.4%	0.2%	0.0%	9.4%
0.2	100(10)	0.0%	27.2%	**47.0%**	13.6%	6.6%	3.0%	1.4%	1.2%	0.0%
0.1	100(20)	0.0%	58.0%	**27.2%**	7.0%	3.4%	2.6%	0.6%	0.2%	1.0%
0.2	100(20)	0.0%	16.6%	**56.6%**	17.0%	7.2%	2.2%	0.2%	0.2%	0.0%
0.1	100(50)	0.0%	52.2%	**33.8%**	11.6%	1.4%	1.0%	0.0%	0.0%	0.0%
0.2	100(50)	0.0%	12.4%	**64.2%**	17.4%	4.4%	1.6%	0.0%	0.0%	0.0%
0.1	200(10)	0.0%	57.4%	**28.6%**	10.2%	2.8%	0.2%	0.2%	0.0%	0.6%
0.2	200(10)	0.0%	7.6%	**66.0%**	19.2%	5.4%	1.4%	0.2%	0.2%	0.0%
0.1	200(20)	0.0%	44.0%	**44.6%**	8.6%	1.6%	0.6%	0.4%	0.2%	0.0%
0.2	200(20)	0.0%	2.4%	**76.0%**	19.2%	2.2%	0.2%	0.0%	0.0%	0.0%
0.1	200(50)	0.0%	32.0%	**55.6%**	10.8%	1.4%	0.2%	0.0%	0.0%	0.0%
0.2	200(50)	0.0%	0.8%	**83.8%**	14.8%	0.6%	0.0%	0.0%	0.0%	0.0%
0.1	500(10)	0.0%	30.4%	**56.0%**	11.6%	2.0%	0.0%	0.0%	0.0%	0.0%
0.2	500(10)	0.0%	0.0%	**87.4%**	11.8%	0.8%	0.0%	0.0%	0.0%	0.0%
0.1	500(20)	0.0%	12.8%	**77.8%**	8.8%	0.6%	0.0%	0.0%	0.0%	0.0%
0.2	500(20)	0.0%	0.0%	**91.0%**	9.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0.1	500(50)	0.0%	4.0%	**89.8%**	6.0%	0.2%	0.0%	0.0%	0.0%	0.0%
0.2	500(50)	0.0%	0.0%	**94.0%**	6.0%	0.0%	0.0%	0.0%	0.0%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Parallel Analysis Based on the Iterated PAF Extraction Technique

PA based on Iterated PAF (Level-1 factor extraction). Table 16 displays the percentages of 500 iterations that retained different factors for the Level-1 model using PA based on the Iterated PAF extraction technique. PA based on Iterated PAF extracted the true factor model at Level-1 with a more than 85% chance under different simulation conditions. Compared with PA based on PAF, PA based on Iterated PAF performed worse in every simulation condition.

Table 16 Level-1 Factor Extraction Results Using PA Based on the Iterated PAF Extraction Technique.

ICC	Sample Size*	Percentage			
		Scenario 1		Scenario 2	
		2-factor	3-factor	**2-factor**	3-factor
0.1	100(10)	**91.0%**	9.0%	**88.2%**	11.8%
0.2	100(10)	**87.0%**	13.0%	**87.6%**	12.4%
0.1	100(20)	**95.2%**	4.8%	**92.6%**	7.4%
0.2	100(20)	**93.4%**	6.6%	**92.2%**	7.8%
0.1	100(50)	**94.0%**	6.0%	**94.0%**	6.0%
0.2	100(50)	**94.6%**	5.4%	**94.2%**	5.8%
0.1	200(10)	**94.6%**	5.4%	**91.8%**	8.2%
0.2	200(10)	**91.2%**	8.8%	**91.8%**	8.2%
0.1	200(20)	**95.8%**	4.2%	**95.6%**	4.4%
0.2	200(20)	**95.8%**	4.2%	**95.8%**	4.2%
0.1	200(50)	**95.4%**	4.6%	**94.2%**	5.8%
0.2	200(50)	**95.8%**	4.2%	**95.8%**	4.2%
0.1	500(10)	**95.2%**	4.8%	**95.0%**	5.0%
0.2	500(10)	**94.0%**	6.0%	**96.6%**	3.4%
0.1	500(20)	**94.2%**	5.8%	**95.6%**	4.4%
0.2	500(20)	**95.8%**	4.2%	**95.8%**	4.2%
0.1	500(50)	**91.6%**	8.4%	**90.4%**	9.6%
0.2	500(50)	**93.0%**	7.0%	**93.4%**	6.6%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

PA based on Iterated PAF (Level-2 factor extraction). Tables 17-18 show the percentages of 500 iterations that retained different factors for the Level-2 model in Scenario 1 and Scenario 2, respectively, using PA based on the Iterated PAF extraction technique. When using PA based on Iterated PAF for the simulated Level-2 data, this study faced a convergence problem, which showed as an "ERROR: Communality greater than 1.0" in the SAS log. Although SAS support provided a syntax = HEYWOOD to the PROC FACTOR statement to help model convergence or a syntax = ULTRAHEYWOOD to avoid communality greater than 1.0, the results were inaccurate or misleading under the above modifications (Steinberg, 2010). Therefore, this study did not apply these modification syntaxes and reports the convergence rate for Tables 17-19.

Table 17 offers the percentages of 500 iterations that retained different factors for the Level-2 model in Scenario 1 using the PA criterion based on the Iterated PAF extraction technique. In Scenario 1, when ICC was higher, the performance of PA based on Iterated PAF became poorer at identifying the between-level factor in most simulation situations controlling for other settings. When ICC = .2, the lowest probability of extracting the correct factors using PA based on Iterated PAF was 22.4%, whereas when ICC = .1, the lowest chance was 67.3%.

Table 18 presents the percentages of 500 iterations that retained different factors for the Level-2 model in Scenario 2 using the PA criterion based on the Iterated PAF extraction technique. In Scenario 2, when ICC was higher, the performance was better at identifying the Level-2 factor in most simulation circumstances controlling for other settings. When ICC = .1, the lowest probability of obtaining the correct latent factors using PA based on Iterated PAF was 10.7%, whereas when ICC =.2, the lowest probability was 35.8%.

Table 17 Level-2 Factor Extraction Results in Scenario 1 Using the PA Criterion Based on the Iterated PAF Extraction Technique.

ICC	Sample Size*	Convergence Rate	Percentage								
			0-factor	**1-factor**	2-factor	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	35.0%	0.6%	**72.0%**	12.6%	6.9%	3.4%	1.7%	0.6%	0.0%	2.3%
0.2	100(10)	80.6%	0.0%	**68.5%**	20.8%	5.0%	2.2%	0.5%	0.2%	0.2%	2.5%
0.1	100(20)	80.2%	0.0%	**69.8%**	21.4%	4.5%	4.2%	0.0%	0.0%	0.0%	0.0%
0.2	100(20)	95.2%	0.0%	**75.0%**	16.4%	5.3%	2.5%	0.4%	0.2%	0.2%	0.0%
0.1	100(50)	93.8%	0.0%	**70.6%**	18.3%	6.6%	4.5%	0.0%	0.0%	0.0%	0.0%
0.2	100(50)	98.4%	0.0%	**88.6%**	5.7%	3.0%	1.0%	1.6%	0.0%	0.0%	0.0%
0.1	200(10)	83.8%	0.0%	**67.3%**	24.3%	3.8%	1.9%	0.2%	0.2%	0.2%	1.9%
0.2	200(10)	99.6%	0.0%	**93.2%**	1.6%	2.4%	0.8%	0.6%	0.8%	0.6%	0.0%
0.1	200(20)	98.8%	0.0%	**85.8%**	4.3%	4.3%	2.0%	2.2%	0.4%	1.0%	0.0%
0.2	200(20)	100.0%	0.0%	**97.6%**	1.8%	0.4%	0.0%	0.2%	0.0%	0.0%	0.0%
0.1	200(50)	100.0%	0.0%	**95.0%**	3.8%	0.8%	0.4%	0.0%	0.0%	0.0%	0.0%
0.2	200(50)	100.0%	0.0%	**89.8%**	8.6%	1.2%	0.4%	0.0%	0.0%	0.0%	0.0%
0.1	500(10)	100.0%	0.0%	**98.2%**	1.0%	0.0%	0.2%	0.2%	0.0%	0.4%	0.0%
0.2	500(10)	100.0%	0.0%	**82.6%**	13.4%	3.2%	0.6%	0.2%	0.0%	0.0%	0.0%
0.1	500(20)	100.0%	0.0%	**80.2%**	15.6%	2.6%	1.0%	0.6%	0.0%	0.0%	0.0%
0.2	500(20)	100.0%	0.0%	**77.6%**	18.2%	3.0%	1.0%	0.2%	0.0%	0.0%	0.0%
0.1	500(50)	100.0%	0.0%	**76.2%**	17.4%	3.4%	2.6%	0.2%	0.2%	0.0%	0.0%
0.2	500(50)	100.0%	0.0%	**76.6%**	17.4%	4.6%	0.6%	0.2%	0.4%	0.2%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 18 Level-2 Factor Extraction Results in Scenario 2 Using the PA Criterion Based on the Iterated PAF Extraction Technique.

ICC	Sample Size*	Convergence Rate	Percentage								
			0-factor	1-factor	**2-factor**	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	33.6%	0.6%	72.0%	**10.7%**	7.7%	3.6%	0.6%	0.0%	0.6%	4.2%
0.2	100(10)	83.2%	0.0%	30.0%	**51.9%**	11.3%	2.2%	1.0%	0.5%	0.7%	2.4%
0.1	100(20)	82.8%	0.0%	61.6%	**29.5%**	3.4%	2.2%	0.2%	0.0%	0.5%	2.7%
0.2	100(20)	95.6%	0.0%	22.2%	**46.4%**	18.8%	7.7%	3.8%	0.8%	0.2%	0.0%
0.1	100(50)	95.2%	0.0%	57.4%	**25.8%**	13.0%	2.5%	0.2%	0.6%	0.4%	0.0%
0.2	100(50)	99.4%	0.0%	28.0%	**37.8%**	21.3%	6.6%	3.6%	1.8%	0.6%	0.2%
0.1	200(10)	84.8%	0.0%	58.3%	**30.0%**	7.1%	1.9%	0.7%	0.2%	0.2%	1.7%
0.2	200(10)	97.8%	0.0%	27.4%	**35.8%**	17.2%	9.2%	7.0%	1.6%	1.0%	0.8%
0.1	200(20)	98.0%	0.0%	67.8%	**15.5%**	10.4%	3.3%	2.0%	0.6%	0.4%	0.0%
0.2	200(20)	100.0%	0.0%	6.2%	**58.8%**	19.2%	5.4%	2.2%	1.8%	1.8%	4.6%
0.1	200(50)	100.0%	0.0%	43.6%	**44.8%**	9.4%	0.6%	0.4%	0.0%	0.2%	1.0%
0.2	200(50)	100.0%	0.0%	0.8%	**73.0%**	12.6%	2.4%	1.6%	0.2%	1.4%	8.0%
0.1	500(10)	100.0%	0.0%	51.4%	**36.6%**	6.0%	1.6%	0.4%	1.2%	1.8%	1.0%
0.2	500(10)	100.0%	0.0%	0.0%	**80.4%**	14.0%	1.4%	0.4%	0.6%	1.0%	2.2%
0.1	500(20)	100.0%	0.0%	3.6%	**78.4%**	16.4%	1.4%	0.2%	0.0%	0.0%	0.0%
0.2	500(20)	100.0%	0.0%	0.0%	**84.2%**	14.8%	0.8%	0.0%	0.0%	0.0%	0.2%
0.1	500(50)	100.0%	0.0%	0.2%	**84.4%**	13.8%	1.6%	0.0%	0.0%	0.0%	0.0%
0.2	500(50)	100.0%	0.0%	0.0%	**86.0%**	12.8%	0.0%	0.2%	0.2%	0.2%	0.6%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 19 Convergence Rate of 1,000 Replications in a Level-2 Model for Random Data Using the PA Criterion Based on the Iterated PAF Extraction Technique.

ICC	Sample Size*	Convergence Rate	
		Scenario 1	Scenario 2
0.1	100(10)	27.7%	26.7%
0.2	100(10)	15.6%	74.2%
0.1	100(20)	71.2%	71.8%
0.2	100(20)	23.4%	92.8%
0.1	100(50)	92.6%	91.8%
0.2	100(50)	33.6%	96.5%
0.1	200(10)	73.7%	73.6%
0.2	200(10)	35.5%	97.5%
0.1	200(20)	96.2%	95.7%
0.2	200(20)	47.8%	99.9%
0.1	200(50)	99.7%	99.7%
0.2	200(50)	53.8%	100.0%
0.1	500(10)	100.0%	99.1%
0.2	500(10)	57.8%	100.0%
0.1	500(20)	100.0%	99.9%
0.2	500(20)	58.4%	100.0%
0.1	500(50)	100.0%	82.3%
0.2	500(50)	62.6%	100.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

In general, PA based on Iterated PAF performed worse than PA based on PAF for Level-2 factor extraction in these two scenarios, especially when ICC was higher.

Parallel Analysis Based on the ML Extraction Technique

PA based on ML (Level-1 factor extraction). Table 20 presents the percentages of 500 iterations that obtained latent factors for the Level-1 model using the PA based on the ML eigenvalues extraction approach. PA based on ML also worked well to identify the true factor in the Level-1 model in most of simulation conditions; however, the probability was the lowest in every simulation situation compared to PA based on PAF or Iterated PAF. In addition, the simulation random data also had a convergence problem, even for the Level-1 model, as illustrated in Table 21.

Table 20 Level-1 Factor Extraction Results Using PA Based on the ML Extraction Technique.

ICC	Sample Size*	Percentage					
		Scenario 1			Scenario 2		
		2-factor	3-factor	4-factor	**2-factor**	3-factor	4-factor
0.1	100(10)	**69.0%**	28.6%	2.4%	**66.2%**	33.0%	0.8%
0.2	100(10)	**61.2%**	36.2%	2.6%	**62.2%**	35.2%	2.6%
0.1	100(20)	**81.8%**	17.6%	0.6%	**74.0%**	25.6%	0.4%
0.2	100(20)	**74.6%**	24.6%	0.8%	**72.8%**	26.4%	0.8%
0.1	100(50)	**82.8%**	17.2%	0.0%	**82.4%**	17.6%	0.0%
0.2	100(50)	**82.8%**	17.2%	0.0%	**82.6%**	17.4%	0.0%
0.1	200(10)	**77.0%**	22.2%	0.8%	**75.0%**	24.6%	0.4%
0.2	200(10)	**73.2%**	26.2%	0.6%	**75.6%**	23.8%	0.6%
0.1	200(20)	**84.0%**	16.0%	0.0%	**85.2%**	14.8%	0.0%
0.2	200(20)	**83.0%**	17.0%	0.0%	**84.4%**	15.4%	0.2%
0.1	200(50)	**85.2%**	14.8%	0.0%	**83.4%**	16.6%	0.0%
0.2	200(50)	**84.6%**	15.4%	0.0%	**85.6%**	14.4%	0.0%
0.1	500(10)	**84.4%**	15.6%	0.0%	**84.0%**	16.0%	0.0%
0.2	500(10)	**81.6%**	18.4%	0.0%	**84.6%**	15.4%	0.0%
0.1	500(20)	**86.2%**	13.8%	0.0%	**85.4%**	14.6%	0.0%
0.2	500(20)	**85.4%**	14.6%	0.0%	**87.8%**	12.2%	0.0%
0.1	500(50)	**78.8%**	21.2%	0.0%	**76.4%**	23.6%	0.0%
0.2	500(50)	**78.8%**	21.2%	0.0%	**82.6%**	17.4%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 21 Convergence Rate of 1,000 Replications in a Level-1 Model for Parallel Random Data Using PA Based on the ML Extraction Technique.

ICC	Sample Size*	Convergence Rate	
		Scenario 1	Scenario 2
0.1	100(10)	65.8%	64.9%
0.2	100(10)	67.4%	66.0%
0.1	100(20)	69.7%	66.6%
0.2	100(20)	100.0%	70.0%
0.1	100(50)	68.2%	68.6%
0.2	100(50)	100.0%	67.5%
0.1	200(10)	67.0%	65.5%
0.2	200(10)	67.4%	66.4%
0.1	200(20)	71.3%	70.8%
0.2	200(20)	71.7%	70.4%
0.1	200(50)	69.5%	69.7%
0.2	200(50)	69.6%	70.8%
0.1	500(10)	68.2%	66.5%
0.2	500(10)	68.2%	68.0%
0.1	500(20)	71.2%	71.6%
0.2	500(20)	70.6%	71.3%
0.1	500(50)	70.8%	70.8%
0.2	500(50)	71.4%	73.6%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

PA based on ML (Level-2 factor extraction). Tables 22-23 present the percentages of 500 iterations that maintained different factors for the Level-2 model in Scenario 1 and Scenario 2, respectively, using PA based on the ML extraction technique. Most of the conditions had a less than 50% probability of identifying the true Level-2 model in both scenarios. This performance of the PA based the ML was also the worst among the three PA methods for Level-2 data. Furthermore, as can be seen from Tables 22-24, the PA based on the ML method had the lowest convergence rate for simulation data. This situation might also lead to bad performance results.

Table 22 Level-2 Factor Extraction Results for Scenario 1 Using PA Based on the ML Extraction Technique.

ICC	Sample Size*	Convergence Rate	Percentage								
			0-factor	**1-factor**	2-factor	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	9.2%	76.1%	**6.5%**	17.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0.2	100(10)	38.4%	44.3%	**22.4%**	3.6%	8.9%	8.9%	5.2%	6.3%	0.5%	0.0%
0.1	100(20)	34.6%	85.5%	**0.6%**	2.3%	1.2%	6.4%	1.7%	1.7%	0.6%	0.0%
0.2	100(20)	74.8%	11.2%	**75.7%**	0.3%	0.8%	4.8%	2.4%	2.9%	1.9%	0.0%
0.1	100(50)	72.0%	40.0%	**47.2%**	0.6%	0.8%	1.7%	4.7%	3.9%	1.1%	0.0%
0.2	100(50)	94.6%	1.1%	**90.3%**	3.0%	1.5%	1.3%	1.7%	1.1%	0.2%	0.0%
0.1	200(10)	41.0%	79.5%	**4.4%**	1.0%	4.9%	2.0%	4.9%	2.0%	1.5%	0.0%
0.2	200(10)	96.8%	53.3%	**35.3%**	3.1%	4.5%	2.5%	0.8%	0.2%	0.2%	0.0%
0.1	200(20)	91.8%	1.1%	**86.7%**	3.3%	3.9%	1.1%	0.9%	2.4%	0.7%	0.0%
0.2	200(20)	100.0%	0.0%	**40.4%**	17.4%	19.8%	14.0%	6.6%	1.8%	0.0%	0.0%
0.1	200(50)	100.0%	10.4%	**24.8%**	19.6%	19.0%	15.0%	8.6%	2.4%	0.2%	0.0%
0.2	200(50)	100.0%	0.0%	**32.0%**	22.2%	16.8%	18.6%	9.2%	1.2%	0.0%	0.0%
0.1	500(10)	99.4%	33.4%	**27.8%**	10.7%	14.3%	8.5%	4.2%	1.2%	0.0%	0.0%
0.2	500(10)	100.0%	0.0%	**28.0%**	22.2%	18.8%	19.0%	8.8%	3.2%	0.0%	0.0%
0.1	500(20)	100.0%	0.0%	**19.0%**	21.8%	22.0%	22.8%	11.0%	3.2%	0.2%	0.0%
0.2	500(20)	100.0%	0.0%	**24.0%**	26.0%	22.8%	14.4%	10.6%	2.2%	0.0%	0.0%
0.1	500(50)	100.0%	0.0%	**20.6%**	27.4%	20.2%	18.6%	10.4%	2.6%	0.2%	0.0%
0.2	500(50)	100.0%	0.0%	**23.4%**	22.0%	22.0%	18.4%	11.8%	2.0%	0.4%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 23 Level-2 Factor Extraction Results for Scenario 2 Using PA Based on the ML Extraction Technique.

ICC	Sample Size*	Convergence Rate	Percentage								
			0-factor	1-factor	**2-factor**	3-factor	4-factor	5-factor	6-factor	7-factor	8-factor
0.1	100(10)	10.2%	31.4%	5.9%	**43.1%**	5.9%	0.0%	2.0%	3.9%	3.9%	3.9%
0.2	100(10)	49.8%	39.0%	2.8%	**4.8%**	11.6%	19.7%	11.2%	9.6%	1.2%	0.0%
0.1	100(20)	44.2%	17.2%	20.8%	**15.4%**	6.8%	16.7%	15.4%	7.2%	0.5%	0.0%
0.2	100(20)	74.6%	71.8%	0.8%	**1.9%**	2.9%	6.4%	9.4%	4.6%	2.1%	0.0%
0.1	100(50)	79.0%	63.8%	7.1%	**1.0%**	2.3%	5.3%	10.1%	8.1%	2.3%	0.0%
0.2	100(50)	86.8%	99.3%	0.0%	**0.0%**	0.0%	0.2%	0.2%	0.2%	0.0%	0.0%
0.1	200(10)	51.2%	62.1%	4.7%	**2.0%**	7.0%	7.4%	7.4%	7.4%	2.0%	0.0%
0.2	200(10)	90.0%	62.4%	0.2%	**2.9%**	3.3%	6.0%	9.8%	8.0%	6.4%	0.9%
0.1	200(20)	91.0%	0.7%	53.8%	**15.2%**	8.4%	4.6%	5.5%	7.5%	3.5%	0.9%
0.2	200(20)	98.2%	0.0%	0.0%	**30.3%**	27.1%	10.6%	9.6%	12.8%	5.9%	3.7%
0.1	200(50)	99.6%	0.0%	2.0%	**36.1%**	38.8%	15.7%	4.4%	1.6%	1.2%	0.2%
0.2	200(50)	99.6%	10.8%	0.0%	**38.2%**	23.1%	6.4%	4.8%	6.6%	5.2%	4.8%
0.1	500(10)	98.0%	0.0%	2.4%	**35.9%**	32.9%	17.8%	4.7%	3.7%	2.0%	0.6%
0.2	500(10)	100.0%	0.0%	0.0%	**57.6%**	32.4%	4.6%	2.8%	1.0%	1.2%	0.4%
0.1	500(20)	100.0%	0.0%	0.0%	**47.6%**	37.2%	12.8%	2.2%	0.0%	0.2%	0.0%
0.2	500(20)	100.0%	0.0%	0.0%	**61.2%**	33.0%	5.0%	0.0%	0.2%	0.6%	0.0%
0.1	500(50)	100.0%	0.0%	0.0%	**52.8%**	37.8%	9.0%	0.4%	0.0%	0.0%	0.0%
0.2	500(50)	100.0%	0.0%	0.0%	**69.0%**	28.4%	1.0%	0.2%	1.0%	0.4%	0.0%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Table 24 Convergence Rate of 1,000 Replications in a Level-2 Model for Parallel Random Data Using PA Based on the ML Extraction Technique.

ICC	Sample Size*	Convergence Rate	
		Scenario 1	Scenario 2
0.1	100(10)	4.7%	4.6%
0.2	100(10)	15.6%	17.3%
0.1	100(20)	13.6%	14.2%
0.2	100(20)	23.4%	26.8%
0.1	100(50)	27.0%	27.3%
0.2	100(50)	33.6%	31.1%
0.1	200(10)	12.3%	13.7%
0.2	200(10)	35.5%	34.3%
0.1	200(20)	32.0%	32.7%
0.2	200(20)	47.8%	47.9%
0.1	200(50)	51.9%	49.7%
0.2	200(50)	53.8%	58.3%
0.1	500(10)	39.3%	42.3%
0.2	500(10)	57.8%	57.1%
0.1	500(20)	53.7%	56.2%
0.2	500(20)	58.4%	60.6%
0.1	500(50)	62.3%	62.3%
0.2	500(50)	62.6%	60.4%

*Sample size combined cluster number and cluster size; thus, 100(10) means the cluster number is 100 and the cluster size is 10.

Discussion

This study applied two different eigenvalues-based criteria (the “eigenvalues > 1” and PA criterion) to identify the numbers of factors for multilevel exploratory analysis. In general, the “eigenvalues > 1” criterion using the ML/MLR-based eigenvalues extraction and the “eigenvalues > 1” criterion using the PCA eigenvalue were not reliable or sufficiently accurate at estimating the numbers of factor for MEFA; specifically, the “eigenvalues > 1” criterion was likely to underestimate the factor numbers for the within-level model and overestimate the number for the between-level model, resulting in a bad model and wrong interpretations. Even though this criterion performed better at detecting the factor numbers at the between-level than the within-level, it serves no practical purpose to apply the “eigenvalues > 1” criterion, because the first essential phase of MEFA in practical studies is to correctly identify the lower level latent factors. Thus, the results of the “eigenvalues > 1” criterion in this study were consistent with those of previous studies (Crawford et al., 2010; Schmitt, 2011; Zwick & Velicer, 1986), showing that this criterion was not reliable. In addition, the present study extended this finding to multilevel data.

This study also focused on the PA criterion based on different extraction techniques (PAF, Iterated PAF, and ML). Overall, the PA approach performed well in extracting the Level-1 model factor number, and the correct rate of identifying the two-factor model was more than 60% for PA based on ML, more than 85% for PA based on Iterated PAF, and more than 95% for PA based on PAF. Many previous researchers have already approved PA as the most suggested standard for choosing the number of latent factors (Cho et al., 2009; Henson & Roberts, 2006; Ledesma & Valero-Mora, 2007; Yang & Xia, 2015). For this study, we treated the Level-1 model as a traditional one-level EFA model, so the results are consistent with that.

However, when the PA approach was used to identify the Level-2 model factor number, results were not as good as at Level-1. This is not surprising as the sample size was less when investigating the Level-2 model than the Level-1 model. In addition, higher cluster numbers or more cluster size or larger ICC can help to increase the chance of identifying the correct factor number at the between-level. The Level-2 factor extraction became complicated and was influenced by ICC, sample size, and the true model itself.

This study demonstrated that PA based on the PAF extraction technique performed best among the three extraction techniques examined. However, in some conditions, the correct rate of the PA based on PAF can be lower than 50% when the sample size is small. Thus, the accuracy of the high-level factor numbers using the PA criterion is still uncertain when the cluster number or cluster size is not large enough. In addition, PA based on Iterated PAF or ML both had convergence issues. Therefore, in a practical application, a study would fail to get the correct number of factors due to the convergence issues.

As a result, we strongly recommend the use of PA based on the PAF extraction technique to identify the factor numbers in MEFA. A simulation study asserted that the PA based on PAF can perform as well in EFA models (Crawford et al., 2010). However, evaluation articles regarding PA based on Iterated PAF or ML for EFA have not been found so far, not to mention for MEFA. In addition, MEFA needs a large sample size for both cluster number and cluster size.

Despite these important findings, this study suffers from several limitations. First, the simulation data were normally distributed whereas many practical studies have non-normal or categorical data. Second, the structures of the simulated multilevel data were simple; that is, only two factors at the within-level and one or two factors at the between-level and only two-level EFA were analyzed. Results from the PA criterion revealed that higher cluster numbers or cluster size were needed to identify higher levels when structures became complicated. Third, sample sizes of generation data in this study was large, while in practical study the sample size is often smaller. In addition, the cluster sizes were balanced in this study, also in practice, they are often unbalanced and unequal from different clusters. Fourth, the study only had two ICC situations; higher ICC can help to extract the correct number for the Level-2 data; thus, in the future, we can study more situations of higher ICC, which may help decide the latent factors even with small sample sizes. Fifth, there still are other unstudied methods that determine the number of factors in single-level EFA, such as minimum average partial (MAP, Velicer, 1976).

CHAPTER IV

CONCLUSIONS

Multilevel data are common in social sciences and exploratory factor analysis (EFA) is an important tool that allows researchers to explore the underlying factor structure of the variables in which they are interested without any presumable structure. Nevertheless, determining the number of the underlying factors is the first and most important step of EFA. This dissertation was the first one to explore the accuracy of two commonly applied approaches (i.e., the model-selection-based approaches and the eigenvalues-based approaches) in detecting the appropriate number of factors in Multilevel EFA through simulation data. This study also considered and examined the impact of several potential factors including sample size (i.e. cluster number and cluster size), ICC, and model specification.

Study 1 evaluated the performance of commonly used model-fit indexes and selection criteria, including RMSEA, CFI, TLI, SRMR, the overall model-fit chi-squared test, AIC, BIC, and SBIC. The results showed that most of the commonly applied fit indexes and selection criteria were successful in discovering the correct number of factors from the within level except for $SRMR_{PS\_W}$ and AIC through the model-based approach. Meanwhile, the performance of CFI and CFI_{PS_W} became doubtful when ICC is equal to .2, and the performance of these fit indexes and criteria became unsatisfactory at the between-level through the model-based approach.

Study 2 revealed that, similar to the previous findings based on the single level data, the "eigenvalues > 1" criterion was not effective at searching for the optimal number of factors regardless of the data level. On the other hand, parallel analysis (PA) approach performed well in recovering the correct number of factors at the within level. When PA was used along with PAF to identify a between-level model, the correct rates were still acceptable in Scenario 1, while in Scenario 2, the performance of PA based on PAF was related to sample size and ICC. In addition, PAF performed the best followed by Iterated PAF and then ML

In both studies, larger number of clusters, larger cluster size, and larger ICC generally led to great chances of obtaining the correct number of factors at the between-level. In addition, if the number of Level-1 sample size was fixed, more cluster sizes would have larger chances of deciding the right number of factors at the level-2.

There are some limitations for the current studies, and future research can extend this line of studies by considering the following factors. First, more complicated model structures for both within- and between-level should be considered for data generation, such as more numbers of factors in each level, non-linear factor-loadings, and different correlation between factors. Second, different level of ICC can be simulated and studied how it influences the performance of difference approaches in identifying the between-level factors. Third, simulation data can extend from normal distribution to non-normal distribution, from numerical type to categorical type.

In sum, when a researcher has multilevel data without having an a priori hypothesis of a preconceived structure, MEFA is recommended to use through the following steps. First, ICC needs to be computed to determine whether the clustering effect should be considered. Second, the model-based approach should be used to conduct the EFA at different levels and get the corresponding model indexes, IC values and other related information. Third, the model-selection-based approaches (except for $SRMR_{PS\_W}$, AIC, CFI, and CFI_{PS_W}) and PA can work well for finalizing the number of factors in the within-level. On the other hand, only PA based on PAF is recommended to extract the factors at the between-level.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332.

Barendse, M. T., Oort, F. J., & Timmerman, M. E. (2015). Using exploratory factor analysis to determine the dimensionality of discrete responses. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 87-101.

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

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*(1), 62-91.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36*(1), 111-150.

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In *Handbook of statistical modeling for the social and behavioral sciences* (pp. 185-249). Boston, MA: Springer.

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

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: a practical information-theoretic approach.* New York, NY: Springer-Verlag.

Burton, L. J., & Mazerolle, S. M. (2011). Survey instrument validity part I: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal*, *6*(1), 27-35.

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

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*(3), 289-325.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, *84*(4), 610-619.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464-504.

Chen, Q., Luo, W., Palardy, G. J., Glaman, R., & McEnturff, A. (2017). The efficacy of common fit indices for enumerating classes in growth mixture models when nested data structure is ignored: a monte carlo study. *SAGE Open, 7*(1), 1-19.

Cheung, M. W. L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology, 37*(5), 522-541.

Child, D. (1990). *The essentials of factor analysis* (2nd Ed.). London, UK: Cassel Educational Limited.

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

Clark, D. A., & Bowles, R. P. (2018). Model fit and item factor analysis: overfactoring, underfactoring, and a program to guide interpretation. *Multivariate Behavioral Research*, *53*(4), 544-558.

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

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*, 1-9.

Crawford, A. V., Green, S. B., Levy, R., Lo, W. J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*(6), 885-901.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*(1), 223-226.

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods, 8*(3), 378-383.

De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*(4), 695-710.

D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement, 21*(2), 209-235.

Dinno, A. (2009). Exploring the sensitivity of Horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, *44*(3), 362-388.

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.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*(3), 343-367.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56-83.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291-314.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*(1), 93.

Gerbing, D. W., & Anderson, J. C. (1992). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods & Research, 21*(2), 132-160.

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.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*(4), 455-467.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*(3), 532-560.

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W. J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement*, *72*(3), 357-374.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149-161.

Hayashi, K., Bentler, P. M., & Yuan, K. H. (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.

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

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

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51-62.

Hsu, H. Y., Kwok, O. M., Lin, J. H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, *50*(2), 197-215.

Hsu, H. Y., Lin, J. H., Kwok, O. M., Acosta, S., & Willson, V. (2017). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, *77*(1), 5-31.

Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424-453.

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

Huang, F. L., & Cornell, D. G. (2016). Using multilevel factor analysis with clustered data: Investigating the factor structure of the Positive Values Scale. *Journal of Psychoeducational Assessment, 34*(1), 3-14.

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.

Humphreys, L. G., & Montanelli Jr, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research*, *10*(2), 193-205.

Hyman, M. R., & Sierra, J. J. (2012). Adjusting self-reported attitudinal data for mischievous respondents. *International Journal of Market Research*, *54*(1), 129-145.

Jöreskog, K. G. (1993). Testing structural equation models. *Sage Focus Editions*, *154*, 294-306.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141-151.

Keith, T. Z., Caemmerer, J. M., & Reynolds, M. R. (2016). Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor? *Intelligence*, *54*, 37-54.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, *51*(6), 881-898.

Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research*, *24*(3), 365-395.

Ledesma, R. D., & Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation, 12*(2), 1-11.

Ledesma, R. D., Valero-Mora, P., & Macbeth, G. (2015). The scree test and the number of factors: A dynamic graphics approach. *The Spanish Journal of Psychology*, *18*(11), 1-11.

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica, 14*(1), 41-68.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113-139.

MacCallum, R. C., Browne, M. W., & Cai, L. (2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: 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.

Mahler, C. (2011). *The effects of misspecification type and nuisance variables on the behaviors of population fit indices used in structural equation modeling* (Doctoral dissertation). University of British Columbia, Vancouver, Canada.

Mayberry, M. L., Espelage, D. L., & Koenig, B. (2009). Multilevel modeling of direct effects and interactions of peers, parents, school, and community influences on adolescent substance use. *Journal of Youth and Adolescence*, *38*(8), 1038-1049.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108-141.

Moore et al. (2016). Characterizing social environment's association with neurocognition using census and crime data linked to the Philadelphia Neurodevelopmental Cohort. *Psychological Medicine*, *46*(3), 599-610.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*(1), 267-316.

Muthén, L., & Muthén, B. (2017). *Mplus* (Version 8) [computer software]. (1998-2017). Los Angeles, CA: Authors.

Muthén, L., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS*. Cary, NC: SAS Institute.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*(S1), 9-36.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate Behavioral Research, 48*(1), 28-56.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25,* 111-163.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Thousand Oaks, CA: Sage.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment, 84*(2), 126-136.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*(4), 583-601.

Sakaluk, J. K., & Short, S. D. (2017). A methodological review of exploratory factor analysis in sexuality research: Used practices, best practices, and data analysis resources. *The Journal of Sex Research*, *54*(1), 1-9.

SAS Institute Inc. (2002). SAS/STAT® Version 9 [software]. Cary, NC: Author.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321.

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*(4), 345-362.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333-343.

Sharma, S., Mukherjee, S., Kumar, A., & Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *Journal of Business Research, 58*(7), 935-943.

Song, J., & Belin, T. R. (2008). Choosing an appropriate number of factors in factor analysis with incomplete data. *Computational Statistics & Data Analysis*, *52*(7), 3560-3569.

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.

Steinberg, J. (2010). Exploring the dimensionality of large-scale standardized educational assessments using PROC FACTOR. Presentation at NESUG 2010 Statistics and Analysis, Burlington, VT.

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 640-657.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Thrupp, M. (2002). Why ''meddling'' is necessary: A response to Teddlie, Reynolds, Townsend, Scheerens, Bosker & Creemers. *School Effectiveness and School Improvement, 13*(1), 1-14.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10.

Velicer, W. F. (1976). The relation between factor score estimates, image scores and principal component scores. *Educational and Psychological Measurement, 36,* 149-159.

Weng, L. J., & Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, *65*(5), 697-716.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In *Factor Analysis at 100* (pp. 191-218). Routledge.

Wu, J. Y., & Kwok, O. M. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(1), 16-35.

Yang, H. Y., & Bozdogan, H. (2011). Learning factor patterns in exploratory factor analysis using the genetic algorithm and information complexity as the fitness function. *Journal of Pattern Recognition Research*, *2*, 307-326.

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*(3), 756-772.

Yuan, K. H., & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*(1), 53-82.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30,* 165-200.

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.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12*(2), 127-140.