EVALUATION OF THE SPLIT-DATA STRATEGY IN FACTOR ANALYSIS

BY

XINCHANG ZHOU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Educational Psychology
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Master's Committee:

Assistant Professor Yan Xia, Chair
Assistant Professor Ge Jiang, Co-Chair
Professor Jinming Zhang

Abstract

When evaluating the psychometric properties of an assessment, researchers can perform an exploratory factor analysis (EFA), followed by a confirmatory factor analysis (CFA) on the same dataset (the whole-sample strategy) to evaluate the model structure. However, the model structure obtained by the whole-sample strategy is based on only one dataset and is, therefore, subject to capitalization on chance. To strengthen the generalizability of models, researchers suggest conducting cross-validation and applying different datasets in practice. Nevertheless, because collecting multiple datasets are not always feasible in practice, researchers commonly conduct the split-data strategy by randomly splitting the dataset into two halves, performing EFA on the first half, and conducting CFA on the second half to validate the structure obtained from EFA. Despite the popularity of the split-data strategy, evidence supporting this strategy is not sufficient in the literature. To examine the utility of the split-data strategy, this thesis research includes two studies using Monte Carlo simulations to explore whether the split-data strategy has advantages over the whole-sample strategy in correctly identifying two critical aspects of model structures in psychological assessments: the number of latent factors and the existence of cross-loadings. Results show that the split-data strategy is less effective than the whole-sample strategy in evaluating the number of factors and cross-loadings in all simulation conditions. Using the split-data strategy is only acceptable, though not necessary, under conditions with large samples (greater than 1,000 for the investigated models) and good model quality (i.e., large primary loadings, no cross-loading, and small factor correlations).

Keywords: confirmatory factor analysis, exploratory factor analysis, cross-validation, parallel analysis, model-data fit

To Aurora.

Acknowledgments

This work was made possible through the support of many people. I am deeply indebted to my advisor Dr. Yan Xia, for his intelligence, steadfast support, and guidance. Also, I would like to thank my co-advisor, Professor Ge Jiang, and committee member Professor Jinming Zhang for their valuable suggestions. I would also like to thank my supportive family members for their selfless assistance in helping me complete this thesis research. I also want to extend my gratitude to Ruibing Song for his continuous support and encouragement.

Table of Contents

List of Abbreviations

EFA	Exploratory Factor Analysis.
CFA	Confirmatory Factor Analysis.
PA	Parallel Analysis.
TPA	Traditional Parallel Analysis.
RMSEA	Root Means Square Error of Approximation.
CFI	Comparative Fit Indices.
TLI	Tucker-Lewis Index.
SRMR	Standardized Root Mean Square Residual.

Chapter 1

Introduction

Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are the core methods in test construction and are widely used for evaluating the psychometric properties of an assessment and uncovering the relationships between latent factors and items (Goretzko et al., 2019; Hurley et al., 1997; Osborne, 2014; Orcan, 2018; Schmitt, 2011). Researchers can choose between the EFA method and the CFA method depending on the research purpose. When researchers do not have a priori knowledge about the model structure but want to understand the potential relationship among variables, EFA can uncover what model structures fit observed data (Child, 2006; Fabrigar et al., 1999). EFA is based on the common-factor model, assuming that there are common latent factors, and the correlation between any two variables can be explained by their links to the underlying factors (Floyd & Widaman, 1995; Schreiber, 2021). Any measured variable can be associated with any factor in EFA without imposing any preconceived structure. EFA aims to reproduce the correlation matrix and find the most suitable model, including determining the number of common factors influencing variables and the relationships between factors and items (McDonald, 1985; Browne, 2001; Osborne, 2014). Figure 1 shows a typical EFA model with two common factors, and each factor is measured by three items.

In psychological and educational research, CFA is probably the most widely used method for examining test structures and measurement models within the structural equation modeling framework and an indispensable analytical tool for building validation (Cohen, 1988; Jöreskog, 1969; Stapleton, 1997). Difference from EFA, CFA is a method based on a *priori* knowledge of the models. In CFA, because researchers need to specify the factor loading patterns and other freely estimated parameters (e.g., residual covariances) in advance, this method requires an empirical or theoretical foundation underlying the measurement model

to guide the model specification and evaluation (e.g., Brown & Moore, 2012; Stevens, 1995). Researchers propose the hypothetical structure of the factor models, impose specific constraints on the model, and then conduct the CFA to evaluate how well the model fits the data. If the model-implied covariance matrix deviates from the sample covariance matrix (i.e., poor model-data fit), researchers should reject the model and test alternative models. CFA allows researchers to test the relationship between the observed variables and the underlying latent constructs in the hypothesized model, testing whether the model is acceptable and whether items have good quality (Gorsuch, 1983; Williams, 1995). Figure 2 shows a typical CFA model with two observed variables, and each factor is measured by three items.

A common strategy is to apply both EFA and CFA to the same dataset when conducting factor analysis, called the whole-sample strategy. The whole-sample strategy is beneficial because of its convenience and ability to perform both EFA and CFA analyses while preserving the original sample size. However, there is a potential lack of cross-validation if the whole-sample strategy is adopted (e.g., Fokkema & Greiff, 2017). If researchers perform both EFA and CFA on the same dataset, the researchers may end up with an overfitted model that can only be applied in this particular dataset. Specifically, the model derived from the whole-sample strategy tends to fit well on that dataset, but the model may not perform effectively when fitted to a new random sample from the same population (e.g., Osborne & Fitzpatrick, 2012). When the model is obtained from only one dataset, capitalization on chance can be a serious issue that can lead to a final model lacking generalizability (MacCallum et al., 1992). To strengthen the model's generalizability and avoid capitalization by chance, researchers suggest cross-validation (Mosier, 1951). Cross-validation is widely used in many fields such as model validation, scale development, and machine learning (e.g., de Rooij & Weeda, 2020; Zhang & Stout, 1999). This thesis research focused on cross-validation in the application of factor analysis.

In order to perform cross-validation, researchers can apply EFA and CFA analysis as two consecutive steps on different datasets, performing EFA first to establish a model based on the first sample and then re-evaluate this model using CFA using a separate dataset (Cabrera-Nguyen, 2010; Hinkin, 1995; Ssebuggwawo et al., 2010; Worthington & Whittaker, 2006). The two (or more) separate samples collected at different times, in different places, or by different methods represent samples with varying sources of noise. If the model established in EFA based on one sample is further supported by the CFA results based on a different sample, the model has strong generalizability. Researchers have shown that this combined EFA-CFA method on different datasets provides richer insights regarding the factor structures (Costello & Osborne, 2005; Gerbing & Hamilton, 1996; Henson & Roberts, 2006). However, conducting EFA and CFA on different samples is not always feasible in practice because researchers need to spend more financial and human resources to collect multiple datasets.

An alternative approach for cross-validation is to randomly divide a large sample into two halves (Cudeck & Brown, 1983). As suggested by Cudeck and Brown (1983), splitting data for cross-validation can be an aid in model structure evaluation. Researchers divided the data into two random halves, using one half as a calibration sample to perform EFA and the other half as a validation sample to perform CFA. If the model obtained from EFA using one half of the dataset is validated by CFA using the other half of the dataset, the researchers can accept the model as the final model.

Despite the potential benefit of cross-validation, splitting data may also bring potential risks. Dividing the original sample into two halves leads to the loss of model fit and parameter evaluation precision. If the sample size is small, the split datasets can be too small to provide accurate parameter estimates and lead to high non-convergence rates (Cudeck & Brown, 1983). Therefore, the premise of conducting the split-data strategy is a sufficient sample size. Another disadvantage is that the split-data strategy may lead to conflicting results after data segmentation (Cudeck & Brown, 1983). To be specific, cross-validation would be affected by random fluctuation, such that, after splitting the data, the results in EFA and CFA may not be consistent with each other, which further brings difficulty in interpretation.

The split-data strategy is widely employed in evaluating the psychometric properties of measures across different areas of behavioral research (e.g., Caleon & King, 2021; Crasta et al., 2021; Schmitt, 2011). I conducted a review of the articles published from January 2021 to February 2022 in *Psychological Assessment*. A total of 34 publications with "exploratory factor analysis" or "confirmatory factor analysis" as a keyword were found (Table 1). Of these articles, 13 used only CFA or EFA (e.g., Rogoza et al., 2021; Stanton et al., 2021). Six articles conducted EFA first on one sample, followed by CFA on different samples (e.g., Burke et al., 2021; Colledani et al., 2021). Eight articles used CFA and EFA on the same datasets (i.e., using the whole-data strategy; e.g., Doherty et al., 2021; Li et al., 2021). Seven articles applied the split-data strategy, using the first half of data for EFA and the second half for CFA (e.g., Brown et al., 2022; Nordgren et al., 2021).

Despite the popularity of the split-data strategy in behavioral research, this strategy has not been thoroughly reviewed or systematically evaluated. Of the seven articles that employed the split-data strategy, one mentioned that the authors of this article used the split-data strategy for convenience as a substitute for conducting EFA and CFA in different samples. Another two articles stated that the application of the split-data strategy is to explore the model structure and test the hypothesized model. However, none of the seven articles mentioned their rationales for using the split-data strategy or compared the effectiveness of the split-data strategy and the whole-sample strategy. Of the eight articles using the whole-sample strategy, the researchers applied CFA for confirmatory and applied EFA (or ESEM) to check if there are any unexamined factor structures in the model.

Given the controversial pros and cons of the split-data strategy, the primary purpose of this research is to examine whether the split-data strategy can help researchers correctly discover the model structure. This research includes two studies using Monte Carlo simulations to compare the split-data strategy with the whole-sample strategy and explore whether the split-data strategy has advantages in correctly identifying two critical aspects of model structures: the number of model factors and the existence of cross-loadings.

Chapter 2 first reviewed the methods for determining the number of factors and evaluating cross-loadings in CFA and EFA. Thereafter, the thesis's research questions and hypotheses were elaborated.

Chapter 3 evaluated Study 1 and described the data simulation and analysis process in detecting the correct number of factors. In sum, 144 datasets were simulated based on four models by varying the sample sizes, loadings, nonnormality, factor numbers, factor correlations, and item numbers. The results from the two strategies were compared to examine which strategy performed more effectively.

Chapter 4 described the data simulation and analysis process for Study 2, which compares the split-data strategy and whole-data strategy when evaluating the existence of cross-loadings. In sum, 144 datasets with one or two cross-loadings in four population models were generated.

Chapter 5 summarized the results of two simulation studies. After that, I discussed the limitations of this thesis and possible directions for future research.

Chapter 2

Method Review

2.1 Detecting the Number of Factors in Exploratory Factor Analysis

Considering the nature of EFA is to explore the underlying model structure, one of the most critical stages of EFA is determining the number of common factors underlying a set of variables. Models that include more factors will produce better data-fit statistics, but researchers prefer parsimonious models with relatively fewer factors that fit the data well. To obtain acceptable parsimonious models, psychometricians have proposed many factor retention procedures in determining the number of model factors. Procedures include the Kaiser's rule (K1 rule; Kaiser, 1960), the scree plot, the minimum average partial (Velicer, 1976), the very simple structure criterion (VSS; Revelle & Rocklin, 1979), the optimal coordinate and acceleration factor (Cattell, 1966; Raiche et al., 2006), the parallel analysis (PA; Horn, 1965), and the fit indices. Among all the approaches in factor number detection, researchers recommend parallel analysis because of its accuracy in most conditions and fit indices because of their widespread adoption (e.g., Courtney, 2013).

2.1.1 Parallel Analysis

Among all various procedures, numerous studies have recommended parallel analysis (PA) and its variants because of its accuracy in providing accurate conclusions (e.g., Çokluk & Koçak, 2016; Green et al., 2015; Lim & Jahng, 2019). Horn (1965)'s traditional parallel analysis (TPA) generates multiple random datasets for comparison to determine the number of factors. TPA is known to be a relatively accurate method (e.g., Çokluk & Koçak, 2016; Xia, 2021). Since TPA requires the random generation of parallel datasets from an identity correlation matrix, with the number of columns the same as the number of variables in

the dataset, this approach has grown in popularity with software programming development. In TPA, p eigenvalues based on the $p \times p$ sample correlation matrix are calculated first. Then, parallel datasets are generated based on a $p \times p$ identity correlation matrix. The newly generated parallel datasets have the same number of items and the same sample size as the observed data, but no common factors are assumed, and the variables are multivariate normally distributed (Horn, 1965). Thereafter, eigenvalues from the sample correlation matrix are compared with the eigenvalues of the generated multiple random correlation matrices. When looking at the k^{th} factor, TPA compares the mean of k^{th} eigenvalues from the parallel datasets with the k^{th} eigenvalue from the observed data (TPA50; Horn, 1965; Humphreys & Montanelli Jr., 1975), and if the former is larger than the latter, the k^{th} factor is retained. The rationale is caused by sampling variability. The k^{th} eigenvalue of a k factor model will be bigger than the mean of the k^{th} eigenvalue of a random dataset with no factor (Franklin et al.,1995; Xia, 2021). Some researchers also advocate using the 95^{th} percentiles of the distributions of eigenvalues as a criterion (TPA95; Buja & Eyuboglu, 1992; Glorfeld, 1995). The performance of TPAM and TPA95 are both acceptable, but TPAM appears to be a more reliable method than TPA95 for providing a more accurate number of factors (Crawford et al., 2010; Lim & Jahng, 2019).

2.1.2 Model Fit Statistics

Model fit indices and χ^2 can be applied in both EFA and CFA to determine if a model with a specific number of factors is acceptable. These model fit statistics include χ^2 goodness of fit, and the root means square error of approximation (RMSEA; Steiger, 1990), the comparative fit indices (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), and the standardized root mean square residual (SRMR; Jöreskog & Sörbom, 1988).

χ^2 is used to determine whether there is a statistically significant difference between the covariance matrix of the obtained dataset and the model-implied covariance matrix. If the χ^2 value is 0, the model-implied covariance matrix perfectly reproduces the data covariance matrix (null hypothesis, H0). The χ^2 significance test value p represents whether there is a significant difference between the hypothesized model and the observed dataset, and a nonsignificant p ($p>.05$) value indicates an acceptable fit (e.g., Gatignon, 2009; McHugh, 2013).

To address the above limitation of χ^2, a series of model fit indices are commonly applied, including RMSEA, CFI, TLI, and SRMR. RMSEA is an absolute fit index, which assesses how far a hypothesized model is from a perfect fit in the population. An RMSEA value smaller than .06 indicates a relatively good model-data fit (Hu & Bentler, 1999). CFI and TLI are both comparative fit indices that compare the hypothesized model with the baseline model (i.e., a model with all observed variables uncorrelated). The two

indexes are similar, but TLI includes the degrees of freedom in the baseline and analytical model (Bentler, 1990; Finch, 2019; Tucker & Lewis, 1973). In both CFI and TLI, a cut-off of .95 or greater indicates a good model fit (Hu & Bentler, 1999). SRMR is also a comparative fit index that shows the average of standardized residuals between the sample and model-implied covariance matrices (Cangur & Ercan, 2015; Jöreskog & Sörbom, 1988). SRMR indicates a good fit when its values are lower than .06 (Hu & Bentler, 1999).

Some studies have explored the use of the χ^2 test and fit indices in determining the number of factors in EFA (e.g., Asparouhov & Muthén, 2009; Finch, 2019; Yang & Xia, 2014). The applicability of various fitting indices is still under debate because the performance of fit indices is sensitive to many design factors, such as sample size, loadings, and the type of variables (Finch, 2020). Some researchers propose that SRMR does not perform well in correctly identifying factor numbers in the model (Yang & Xia, 2014; Finch, 2019). Yang and Xia (2014) suggested that RMSEA is preferred in providing accurate estimates of factor numbers on large samples and low factor correlation conditions. Clark and Bowles (2018) studied indicator variables in a normal distribution and argued that CFI and TLI are more accurate fit indicators than RMSEA. Finch (2019) suggested that RMSEA can be used as a supplementary method to PA when the loading is small, while the performance of CFI and TLI is relatively poor.

2.2 Detecting Cross-loadings in Exploratory Factor Analysis

Cross-loading is another critical aspect of the model structure. In factor analysis, a variable may measure multiple factors simultaneously, which will result in cross-loadings. If researchers fail to identify large cross-loadings and ignore trivial cross-loadings in the data correctly, the stability in the factor structure will be severely affected and ultimately lead to flawed theoretical conclusions. Cross-loadings should be appropriately captured and reflected in the EFA model.

When determining if a model with specific cross-loadings is acceptable, the χ^2 test, fit indices, and size of the cross-loadings are all commonly used procedures. Many researchers have discussed applying the χ^2 test statistics and fit indices to evaluate the existence of cross-loadings (e.g., Cooper et al., 2010; Li et al., 2020; Mai et al., 2018).

Another way to determine if cross-loadings should be kept in EFA is to use cut-off criteria to determine whether the cross-loading should be removed. The cut-off value in the EFA model does not follow a fixed criterion, but researchers recommend .3 as an acceptable threshold for detecting cross-loadings (e.g., Costello & Osborne, 2005; Howard, 2015; Matsunaga, 2010). If a cross-loading is greater than .3, the cross-loading is sufficient and can be preserved. If a cross-loading is smaller than .3, researchers can omit this small cross-loading from the final structure model.

2.3 Detecting the Number of Factors in Confirmatory Factor Analysis

Hypothesized models are specified based on theoretical or practical evidence before conducting CFA. Researchers evaluate a hypothesized model structure with a pre-specified number of factors to see how well the model can reproduce the sample covariance matrix.

In determining the number of common factors in CFA, fit indices and the χ^2 test are widely applied to determine whether a model is acceptable. In addition to χ^2, researchers suggest RMSEA, SRMR, CFI, and TLI to evaluate model structure in applied CFA research because these indices are not sensitive to slight model-misspecification under large sample sizes (e.g., Marsh et al., 1996; Sun, 2005).

2.4 Detecting Cross-loadings in Confirmatory Factor Analysis

In practice, most (if not all) cross-loadings should be fixed to zero in the typical CFA procedure because CFA models usually assume that the underlying factors have a concise structure where each variable only measures one factor (e.g., Fu et al., 2021). However, restricting all cross-loadings to zero often leads to an unsatisfactory model fit, distorted factor structure, and overestimated factor correlations (Schmitt, 2011; Yang & Green, 2010). When cross-loadings are relatively high, researchers propose that people should not avoid the inclusion of cross-loadings into their hypothesized models because even if cross-loaded items increase the complexity of the model, researchers can still obtain well-fitting solutions (e.g., Lucas, 2004; Ozkok et al., 2019). Conversely, properly including the cross-loadings in hypothesized models can benefit a lot in improving model fit.

Similarly, researchers can use fit indices and χ^2 to measure how well the model fits the data when when cross-loadings exist in CFA. If a model with cross-loadings results in a better fit than a model without cross-loadings, the researchers should consider accepting the model with cross-loadings. In addition, researchers can also determine the existence of cross-loadings using the size of the cross-loadings. In CFA, cross-loadings of .3 or higher can be kept in the final model (e.g., Costello & Osborne, 2005; Howard, 2015; Matsunaga, 2010). Cross-loadings smaller than .3 are viewed as insufficient evidence to support cross-load items and thus should be ignored.

2.5 Research Questions

The current simulation research composes two simulation studies. The first simulation study aims to evaluate if the split-data strategy provides more accuracy when determining the number of factors by addressing the following research questions:

1. Does the split-data strategy provide more accurate estimates of the number of factors than the whole-sample strategy?

   Hypothesis: The split-data strategy will perform as accurately as the whole-sample strategy only in large samples. In small samples, the split-data strategy will provide less accurate results.

2. How do the split-data strategy and the whole-sample strategy perform in determining the factor number in different simulation conditions (loadings, number of items per factor, factor correlations, nonnormality)?

   Hypothesis: With a sufficient sample size, both strategies will perform well in detecting the number of factors under the conditions of normal distribution, big loadings, small factor correlations, and many items per factor. The whole-sample strategy will consistently provide more accurate results than the split-data strategy.

The second simulation study evaluates if the split-data strategy leads to a more appropriate decision regarding the existence of cross-loadings. The research questions are as follows:

1. Does the split-data strategy provide more accurate evaluations of the existence of cross-loadings than the whole-sample strategy (measured by EFA and CFA on the whole sample)?

   Hypothesis: A larger sample size facilitates the evaluations of complex model structure (a model with cross-loadings), so the whole-sample strategy will provide more accurate results than the split-data strategy. In the split-data strategy, a reduced sample size will lead to higher error rates.

2. How do the split-data strategy and the whole-sample strategy perform in evaluating the cross-loadings in different simulation conditions?

   Hypothesis: Both strategies will result in more accurate evaluations under big primary loadings and large samples. Increasing the number of cross-loadings in the population will lead to low accuracy rates in evaluating the existence of cross-loadings in both strategies. When the values of cross-loadings are similar to primary loadings, both strategies will perform less accurately.

Chapter 3

Study 1: Determining the Number of Factors

3.1 Method

3.1.1 Data Generation

Study 1 aimed to evaluate the effectiveness of the split-data strategy for determining the number of factors. Figure 3 shows data simulated according to four population models (M1–M4). M1–M4 varied in the number of factors and the number of items measuring each factor. M1 was a 1-factor 4-item model, while M2 had twice as many items as M1. M3 was a 3-factor CFA model with four items measured for each factor. M4 was also a 3-factor model, but eight items measured each factor. For M1–M4, the loadings (λ) were varied at .4 and .7, representing items that were just acceptable and items that had high quality, respectively. For M3–M4, factor correlations (ρ) varied at .3 and .6. When ρ were relatively high (i.e., .6), correctly detecting the number of factors became more difficult.

For a given population model, the equation for generating the data was

$$x = \Lambda\xi + \delta, \tag{1}$$

where x was a $p \times 1$ vector of p observed variables, ξ was a $k \times 1$ vector of k factors such that $p<k$, Λ was a $p \times k$ matrix containing the factor loadings (λ_{ij} for loading in i row, j column, $0 \leq i \leq p, 0 \leq j \leq k$), and δ was a $p \times 1$ vector representing the residuals (the error). It is assumed that $E_{(x)} = E_{(\xi)} = E_{(\delta)} = 0$ and $Cov_{(\xi\delta)} = 0$.

The covariance matrix of x was

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \tag{2}$$

where Φ was the $k \times k$ covariance matrix of ξ and Θ was the $p \times p$ covariance matrix of δ. Because the unique factors were uncorrelated, Θ was a diagonal matrix. The diagonal elements were set to be 1 in Φ such that Φ is a correlation matrix. Additionally, the diagonal elements in Θ were $1 - \lambda^2$ such that the standardized and unstandardized solutions were the same.

For example, for M3, Φ was

$$\begin{bmatrix} 1 & \sigma & \sigma \\ \sigma & 1 & \sigma \\ \sigma & \sigma & 1 \end{bmatrix}$$

the Λ was

$$\begin{bmatrix} \lambda & \lambda & \lambda & \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda & \lambda & \lambda & \lambda & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda & \lambda & \lambda & \lambda \end{bmatrix}$$

and Θ was a diagonal matrix with the diagonal elements being $1 - \lambda^2$.

For each variance-covariance matrix for a given model specification, this study simulated datasets following multivariate normal or nonnormal distributions. The data followed a multivariate normal distribution when skewness and kurtosis were (0, 0). Data with (0, 1) were considered slightly nonnormal. When the skewness and kurtosis were (1, 3), the data were considered moderately nonnormal. When skewness and kurtosis were (2, 7), data were considered highly nonnormal. Normally distributed data were generated using the *mvrnorm* function from the *MASS* package (Venables & Ripley, 2002) in R (R Core Team, 2020). The R function obtained from Foldnes and Olsson (2016) was applied to generate nonnormally distributed data.

The sample sizes (n) were varied at 200, 400, and 1,000, which are commonly used in behavioral and educational research. 200 is generally considered a small sample, and 1,000 is a relatively large sample. For each simulation condition, 1,000 replications were implemented.

In sum, the total number of conditions in Study 1 was 144. Specifically, M1 and M2 yielded $2(\lambda = .4, .7) \times 4$(combinations of skewness and kurtosis) $\times 3(n = 200, 400, 1,000) = 24$ simulation conditions. M3 and M4 yielded $2(\lambda) \times 2(\rho) \times 4$ (skewness and kurtosis) $\times 3(n) = 48$ simulation conditions.

3.1.2 Data Analyses

The study conducted the traditional parallel analysis (TPA; Horn, 1965) to evaluate the number of latent factors in EFA accurately. TPA95 and TPA50 were used to determine the number of factors. In addition to the parallel analysis, the χ^2 test fit indices were also employed to evaluate the number of factors in EFA. In CFA, the χ^2 test and fit indices were applied. The χ^2 goodness of fit statistic estimated absolute model fit, and nonsignificant p (p>.05) values indicate a good fit. Considering that the χ^2 test is sensitive to sample size and tends to reject models in large samples, the root means square error of approximation (RMSEA; Steiger, 1990), the comparative fit indices (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), and standardized root mean square residual (SRMR; Jöreskog & Sörbom, 1988) were also employed. According to Hu and Bentler's (1999) suggestion, an acceptable fit was indicated by RMSEA <.06, CFI and TLI >.95, and SRMR <.08.

Each simulated dataset was analyzed using two strategies. The first strategy randomly divided the dataset into two halves. The first half of the data was used for both parallel analysis and EFA analysis with the correct number of factors. Then a CFA analysis with the correctly specified models was performed on the rest of the data. The second strategy conducted the EFA, parallel analysis, and CFA analyses with the correctly specified models separately on the entire sample, without data splitting. The normal-theory maximum likelihood estimation was employed for all the data analyses. In EFA, the *oblimin* rotation from the *Psych* package was applied to obtain oblique rotations because M3 and M4 had correlated factors (Revelle, 2022). In CFA, the *cfa* function from the *lavaan* package was used (Rosseel, 2012).

In brevity, results based on selected model-fit indices were reported. Specifically, for EFA, results based on TPA50, χ^2, and RMSEA were presented. For CFA, results obtained from χ^2 and RMSEA were reported. For example, in the split-data strategy, TPA50, χ^2, and RMSEA were used in the first half of the data, and χ^2 and RMSEA were applied in the second half as a validation. When all measures support the correct number of factors, a method (EFA, TPA, or CFA) would be seen as correctly measuring the factor numbers.

Table 2 shows the four patterns of results (pattern $a - d$) obtained from the split-data or the whole-sample strategy. For example, if the split-data strategy resulted in pattern a, results suggested that the split-data strategy was supported because both EFA and CFA consistently led to the correct number of factors. If a strategy resulted in pattern $b - d$, results from EFA and CFA were inconsistent, which could potentially lead to problematic interpretation. However, if pattern b occurs, the final number of factors was still correct if researchers decide to keep the number of factors obtained from the confirmatory analysis (CFA). Despite the correct number of factors, the inconsistent results in pattern b led to difficulty in interpretation.

This study calculated the proportions of pattern a and pattern b in each strategy were calculated to compare the split-data and whole-sample strategies. A high proportion of pattern a showed the priority of a strategy in estimating the correct number of factors. The results also included the proportions of pattern b. A high percentage of pattern b meant a high proportion of conflict results (although the estimations of factor number were correct in CFA, the EFA/TPA results conflicted with CFA results).

3.2 Result

The percentage of replications for a given model that did not reach convergence was deficient (up to 9.9% in M2, under a small n of 200, skewness of two, and kurtosis of seven, with λ of .4 and ρ of .6). Therefore, results were only based on replications that converged.

Figure 4 and Figure 5 present the proportions of pattern a and pattern b for whole-sample and split-data strategies in different sample sizes (n), number of factors, number of items, loadings (λ), skewness and kurtosis, and factor correlation (ρ). The effectiveness of the two strategies was measured by the proportions of pattern a and pattern b (mostly pattern a) in detecting the factor number of the model.

3.2.1 M1 and M2

Figure 4 represents the proportions of pattern a and pattern b rates of different analysis strategies in M1 and M2. Both M1 and M2 are one-factor models but vary in item numbers.

Impact of sample size (n). Figure 4 demonstrated that the whole-sample strategy outperformed the split-data strategy in all sample sizes. When n was large enough (1,000), the rates of correctly detecting the factor numbers (pattern a) in both strategies were high. In the whole-sample strategy, pattern a's proportions were between 86.8% and 95.7%, while the split-data strategy was less accurate (85.7% to 95.3% for pattern a). With a large n = 1,000, both strategies estimated the number of factors well. There was a slight advantage in the whole-sample strategy because the split-data strategy led to more mistakes even in large samples (see Figure 4).

When n = 400, both strategies had declined accuracy, but the whole-sample strategy outperformed the split-data strategy in detecting the number of factors. The whole-sample strategy yielded higher accuracy rates (82.6% to 95.4% in pattern a, 0% to 0.2% in pattern b), while the split-data strategy provided less accurate results (70.2% to 94.7% in pattern a, 0% to 3.7% in pattern b).

When n = 200, both strategies resulted in higher error rates. In small samples with 200 observations, the whole-sample strategy provided more accurate results (71.4% to 94.6% in pattern a, 0% to 1.6% in pattern b) than the split-data strategy, which yielded proportions of pattern a between 55.1% and 84.1%,

and 0% to 13.2% pattern b. In small samples (200), more contradictory results (pattern b) emerged in the split-data strategy.

Overall, results showed that the whole-sample strategy was recommended in small n (200) because of the higher proportion of accuracy estimates (pattern a) and less conflicting results (pattern b) than the split-data strategy. However, the results showed that researchers should scrutinize the final models when $n = 200$ because any strategy could lead to a wrong estimation, even in the more accurate whole-sample strategy. When n is large enough (1,000), both the whole-sample strategy and the split-data strategy can accurately estimate the number of factors. It is worth noting that when n is medium to large (400 to 1,000), the proportions of pattern b were small in both strategies, which means that conflicting results hardly appear in medium to large samples.

Impact of nonnormality. The violation of the multivariate normality (MVN) assumption can impact the effectiveness of factor analysis strategies. Results from Figure 4 showed only slight changes in estimating the number of factors in both strategies when data distributions changed from normality to severe nonnormality.In the whole-sample strategy, pattern a ranged from 80.7% to 95.4% in normally distributed data, 78.5% to 95.7% when slightly nonnormal, 77.4% to 95.1% when moderately nonnormal, and dropped to 71.4% to 95.2% when skewness and kurtosis were (2, 7). In the split-data strategy, pattern a fell between 59.6% and 94.4% in normally distributed data, 55.1% to 95% when slightly nonnormal, 57% to 94.8% when moderately nonnormal, and ranged between 62.1% and 95.3% when skewness and kurtosis were (2, 7). Regardless of the skewness and kurtosis, the whole-data strategy was more effective than the split-data strategy in general.

Impact of λ. As a representation of the connection between latent factors and items, loading was a vital index when considering the performance of factor analysis methods. Accuracy for both strategies showed an increasing tendency when the λ became higher because a large λ represented that the items can effectively measure latent factor, which in turn led to more accurate verification of factor number. For λ of .4, the whole-sample strategy exhibited higher accuracy rates in detecting the number of factors with values between 81.8% and 95.5% in pattern a, compared with 55.1% to 95% in the split-data strategy. Under λ valued .4, the whole-sample strategy also had fewer conflict results (pattern b) ranged from 0% to 1.6% than the split-data strategy (0% to 13.2%). For λ of .7, the whole-sample strategy was still superior with 71.4% to 95.7% proportions of pattern a, compared with the split-data strategy (62.4% to 95.3%). Both strategies had nearly 0% proportions of pattern b (conflict results) when $\lambda = .7$.

In sum, the whole-sample strategy produced more accurate results (pattern a) in estimating the number of factors than the split-data strategy for both λ values. When λ was small, the split-data strategy led to more conflicting estimations in factor numbers (pattern b), while the whole-sample strategy rarely

produced conflict results. Both strategies had few conflict results when λ was .7.

Impact of the number of items per factor. The number of items per factor was another crucial factor in factor analysis. To measure the latent factors more accurately, researchers tend to keep more items for each factor. As was true for the λ and n, both strategies achieved more accurate results when the item number changed from 4 to 8. For a smaller item number (4), compared with the split-data strategy (55.1% to 94.3% for pattern a), the whole-sample strategy performed more accurately in detecting the factor numbers (71.4% and 95.1% for pattern a). Likewise, the whole-sample strategy (81.2% to 95.7% for pattern a) was also preferable when doubling the item number. In comparison, the split-data strategy ranged from 81.2% to 95.7% for pattern a. Comparing the two strategies, the whole-sample strategy performed more accurately in estimating the number of factors and resulted in fewer conflicts than the split-data strategy under both 4-item and 8-item conditions.

3.2.2 M3 and M4

Figure 5 represents the proportions of pattern a and pattern b of different analysis strategies in M3 and M4. In addition to n, λ, and nonnormality, the influence of ρ on the choice of analysis strategies was evaluated. One concern was that $\rho = .6$ in conjunction with $\lambda = .4$ resulted in low accuracy estimates of factor numbers in both strategies because of its poor quality, as shown in Figure 5. Therefore, this combination ($\rho = .6$ and $\lambda = .4$) were not included when discussing the impact of sample size (n), nonnormality, and item number per factor.

Impact of sample size (n). The increase in sample size resulted in more excellent performance for both strategies. Still, the whole-sample strategy outperformed the split-data strategy in all samples. Both strategies performed well for large samples (1,000) in detecting the number of factors. The whole-sample strategy performed more effectively, exhibiting 75.3% to 95.3% in pattern a, while the split-data strategy yielded 70.0% to 93.2% in pattern a. Both strategies yielded low proportions of pattern b (0% to 0.5% in the whole-sample strategy and 0% to 10.3% in the split-data strategy).

Under small samples (200), the whole-sample strategy outperformed the split-data strategy in accurately identifying the number of factors. The whole-sample strategy displayed higher rates of pattern a and lower rates of pattern b than the split-data strategy (37.1% to 93.5% for pattern a and 0% to 49.7% for pattern b). Inaccurate results occurred in the split-data strategy, yielding rates between 18% and 89.2% for pattern a and between 0.8% to 71.6% for pattern b.

Again, the reduction in n impaired the ability of the split-data strategy to identify the factor number. The split-data strategy led to more mistakes in factor number estimation and high proportions of

contradictory results in small to medium samples, so the whole-sample strategy was supported. For large samples (1,000 or more), although the whole-sample strategy appeared to be more effective, both strategies can be applied to estimate the number of factors in the model. However, researchers should take caution before accepting the results, as both strategies may lead to wrong estimates of factor numbers.

Impact of nonnormality. The whole-sample strategy outperformed the split-data strategy in all skewness and kurtosis combinations. When the data deviation from normal distribution became bigger, both strategies appeared to perform worse in detecting the correct number of factors. Although declines in accuracy rates occurred in both strategies when the nonnormality became bigger, the whole-sample strategy still provided more accurate predictions in factor numbers than the split-data strategy.

Impact of the number of items per factor. The impact of the number of items per factor is also detected in M3 and M4. In the three-factor models, unlike one-factor models, larger item numbers do not necessarily mean more accurate estimates of factor numbers because increased model complexity (more factors or more items) results in less useful fit indices.

For a smaller item number (4), the whole-sample strategy performed more accurately in detecting the factor numbers (37.1% to 95.3% for pattern a and 0% to 49.7% for pattern b) than the split-data strategy (18% to 94% for pattern a and 0% to 71.6% for pattern b). When the number of items for each factor doubled, the whole-sample strategy was also suggested in detecting the number of factors (60% to 94.8% for pattern a and 0% to 11.8% for pattern b), while the split-data strategy performed worse (29.5% to 92.7% for pattern a and 0% to 28.5% for pattern b). When the item number for each factor changed from 4 to 8, both strategies resulted in fewer conflict results (pattern b). The whole-sample strategy was superior to the split-data strategy under both 4-item and 8-item conditions.

Impact of loading λ and factor correlation ρ. In the three-factor models, the effects of ρ and λ were analyzed together. Based on the combinations of λ (.4 and .7) and ρ (.3 and .6), the result can be seen in Figure 5.

Both strategies performed well under small ρ (.3) combined with big ρ (.7). A minor ρ means low overlap between the latent factors, and a large ρ means each item can effectively measure the latent factor. The model is ideal, so both analysis strategies worked well in predicting the correct number of factors, resulting in more accurate estimations in factor numbers and fewer conflict results. The whole-sample strategy outperformed the split-data strategy in factor number detection, with 61.3% to 95% proportions of pattern a and 0% pattern b. The split-data strategy exhibited lower accuracy rates between 41.5% and 94% for pattern a and 0% to 1.6% for pattern b. The results showed that the whole-sample strategy was preferred.

When the ρ and λ were small (.3 and .4, respectively), the difference between latent factors was still evident, but the item quality was not good. The validity of the items is insufficient because of poor item

quality, which can further lead to the wrong estimation of factor numbers or conflict results, so the accuracy rates of both strategies became low. The whole-data strategy displayed higher and more stable estimations in factor number, resulting in 56.3% to 94.9% for pattern a and 0% to 37.6% for pattern b. The split-data strategy was not as effective as the whole-data strategy when detecting the factor numbers, with proportions of 23.7% to 92.5% in pattern a and 0.1% to 43.2% in pattern b. All results showed that the whole-sample strategy yielded more accurate results in detecting the number of factors.

When λ was .7 and ρ was .6, the factors were tightly correlated, and the items were of high quality. The whole-sample strategy was still suggested because of better performance in detecting the number of factors, yielding 37.1% to 95.3% for pattern a and 0% to 49.7% for pattern b. In comparison, the split-data strategy resulted in 18% to 92.7% in pattern a and 0% to 71.6% in pattern b. The whole-sample strategy almost consistently exhibited higher accuracy rates than the split-data strategy. Notably, both strategies yielded high proportions of pattern b (i.e., EFA/TPA did not yield the correct number of factors, despite CFA suggesting the correct number of factors), causing difficulties when interpreting the results.

When the ρ was large (.6) but λ was small (.4), the latent factors would be obscure, and the quality of items was too poor to measure the latent factors accurately. Both strategies resulted in the low pattern a proportions and high pattern b proportions. Still, the whole-sample strategy was recommended because of higher proportions of pattern a (8.1% to 93.8% in the whole-sample strategy and 6.5% to 78.5% in the split-data strategy).

Considering all combinations of λ and ρ conditions, the whole-sample strategy was superior to the split-data strategy in identifying the number of factors because of higher pattern a rates in estimating the number of factors and lower probabilities of conflicting results (pattern b). It is worth noting that both strategies led to problematic estimations under small loadings and small to medium samples. The results should be scrutinized before accepting.

3.2.3 Summary

Overall, the results supported the whole-sample strategy in both one-factor and three-factor models due to high accuracy rates (pattern $a \geq 90\%$) in determining the factor numbers under medium to large samples and high-quality models (big loadings, small factor correlations, normally distributed, many items in each factor). Also, the whole-sample strategy performed more effectively than the split-data strategy in correctly detecting the number of factors in all simulation conditions. The estimation of factor number in the split-data strategy was accurate only if the sample size (n) and loadings (λ) were big enough. All in all, the study does not provide evidence supporting the split-data strategy in identifying factor numbers.

Chapter 4

Study 2: Evaluating the Existence of Cross-loadings

4.1 Method

4.1.1 Data Generation

Study 2 aimed to evaluate the effectiveness of the split-data strategy for evaluating if the cross-loadings (c) should be kept or not. As shown in Figure 6, data were generated according to four population models (M5–M8). The four models varied in the number of cross-loadings and the number of items measuring each factor. M5 and M7 were both 3-factor 4-item models, and their difference was in the number of cross-loadings. In M6 and M8, each factor was measured by eight items. For M5–M8, the primary loadings (λ) were varied at .4 and .7, and the factor correlations (ρ) were varied at .3 and .6. For M5–M8, the cross-loadings (c) were set at .2, representing a small c that should be ignored, and .4, representing a big c that should be kept in the final model.

In Study 2, data with three sample sizes (n = 200, 400, and 1,000) were simulated from the multivariate normal distributions according to the specification of the models. For each condition, 1,000 replications were implemented. The total number of conditions in Study 2 was 144. Specifically, M5 and M6 yielded 2 (λ) $\times$ 2 (ρ) $\times$ 2 (c) $\times$ 3 (n) = 24 simulation conditions. M7 and M8 yielded 2 (λ) $\times$ 2 (ρ) $\times$ 2 (c_1) $\times$ 2 (c_2) $\times$ 3 (n) = 48 simulation conditions.

4.1.2 Data Analyses

Each simulated dataset was still analyzed using the split-data and the whole-sample strategy. EFA and CFA analyses on the entire sample were conducted for the whole-sample strategy. Specifically, the models with the correct specification of cross-loadings were used as the analytical models in CFA. In EFA, *oblimin* rotation was applied to obtain oblique rotations due to correlated factors. The normal-theory maximum likelihood estimation was employed for all the analyses. For the split-data strategy, EFA was conducted on the first half of the data, and the following CFA with the correct specification of cross-loadings was applied to the second half of the data. For both the EFA and CFA analyses, the fit indices, including the χ^2 goodness of fit statistic ($p > .05$ = good fit), CFI ($>.95$ = good fit), TLI ($>.95$ = good fit), RMSEA ($<.06$ = good fit) and SRMR ($<.08$ = good fit), were calculated.

The size of c was another criterion to evaluate the existence of cross-loadings. Researchers suggested that cross-loadings greater than .3 should be kept (Costello & Osborne, 2005; Howard, 2015; Matsunaga, 2010; Tabachnick & Fidell, 2001). When the population-level c was set at .2, cross-loading should not exist in the final model. When c was set at .4, cross-loading should exist in the final model.

Considering some criteria yielded low accuracy rates (for instance, the χ^2 goodness of fit was too sensitive to sample size, producing less than 10% accuracy rates in large samples), only several criteria were included in the results presented below to keep the article at a reasonable length. For EFA and CFA, the CFI, RMSEA, and size of cross-loadings were used to evaluate if the model with the correct number of cross-loadings was acceptable. Cross-loadings greater than .3 were kept in the model. When all measures show that the model with the correct number of cross-loadings is acceptable, a method (EFA or CFA) would be seen as correctly measuring the cross-loadings.

The four patterns of results (pattern $a - d$) of the whole-sample strategy and the split-data strategy are listed below (Table 3). The effectiveness of the split-data strategy was compared with the whole-sample strategy. To compare the two strategies, the proportions of pattern a and pattern b in each strategy were calculated. A high proportion of pattern a showed a strategy can correctly evaluate if the cross-loadings should be kept or not. A high percentage of pattern b meant a high proportion of conflict results (although CFA suggested the model with the correct number of cross-loadings was acceptable, the EFA results conflicted with CFA results).

4.2 Result

In Study 2, the convergence rates were relatively low in some conditions. For example, when n = 200, the lowest convergence rate was 51.9% in the 3-factor 4-item with c = .4. Additional replications were performed until there were 1,000 converged replications in each simulation condition. Figure 7 and 8 shows the proportions of pattern a and pattern b for the whole-sample strategy and the split-data strategy in different sample sizes, number of cross-loadings, number of items per factor, loadings, and factor correlation.

4.2.1 M5 and M6

Figure 7 presents the proportion of replications that support different analysis strategies in M5 and M6. It is worth noting that when λ = .4, low accuracy rates occurred regardless of the strategy used in the analysis because the value of c (.2 or .4) was quite close to λ. When analyzing the impact of sample size, size of cross-loadings, and the number of items per factor, only the conditions with λ = .7 were analyzed.

Impact of sample size (n). The whole-sample strategy was preferable to the split-data strategy in all sample sizes. The accuracy rates of both strategies were low when n = 200. The whole-sample strategy yielded more accurate results (36.8% to 64.5% for pattern a and 13.9% to 63.1% for pattern b). The split-data strategy exhibited 24.4% to 51.1% for pattern a and 15.4% to 44.1% for pattern b). Both strategies resulted in high error rates under n = 200, and even the more accurate strategy (whole-sample) could lead to the wrong conclusions regarding cross-loadings.

When n = 400, both strategies became slightly more effective. For the comparison of the two strategies, the whole-sample strategy still appeared to yield higher accuracy rates in correctly evaluating if the cross-loading should be kept or not (35.9% to 66.9% for pattern a and 8.4% to 64.1% for pattern b). The split-data strategy provided less accurate evaluations of the model, resulting in 28.5% to 63.7% for pattern a and 15.7% to 49.6% for pattern b.

The accuracy rates in both strategies were high when the sample was 1,000. The whole-sample strategy was still the outperformer, yielding 26.7% to 75.3% for pattern a and 3.4% to 73.3% for pattern b. The split-data strategy performed less accurately in evaluating the cross-loading, with 27.2% to 69.7% for pattern a and 7.3% to 57.9% for pattern b.

In sum, the whole-sample strategy led to more accurate results (pattern a) than the split-data strategy in all sample sizes. When n became large, the difference between the two strategies decreased, but the accuracy rates of the split-data strategy were still lower than those of the whole-sample strategy.

Impact of the number of items per factor. The number of items in each factor affected the accuracy rates in both strategies, as shown in Figure 7. In M5, each factor was measured by four items. Results

showed that the whole-sample strategy still outperformed the split-data strategy. The whole-sample strategy yielded more effective performance than the split-data strategy (26.7% to 65.2% for pattern a and 3.4% to 73.3% for pattern b), while the split-strategy yielded less accurate results (24.4% to 56.2% for pattern a and 7.3% to 57.9% for pattern b).

In M6, the item numbers were doubled. Likewise, the whole-sample strategy was still recommended, resulting in 39% to 75.3% for pattern a and 12.7% to 42% for pattern b, while the split-data strategy yielded more errors (31.4% to 69.7% for pattern a and 15.4% to 41.1% for pattern b). The whole-sample strategy was superior to the split-data strategy in correctly evaluating the existence of the cross-loading. When the item numbers were doubled, both strategies appeared to be more effective, resulting in higher proportions of pattern a.

Impact of cross-loading c. There were two kinds of cross-loadings introduced. One is a small c that should be ignored in the final model, and the other is a big c that should be kept. When $c = .4$, the whole-sample strategy was recommended for having higher proportions of accepting the correct model with cross-loading than the split-data strategy, with 26.7% to 75.3% for pattern a and 24.7% to 74.3% for pattern b, while the split-data strategy yielded 24.4% and 69.7% for pattern a and 30.3% to 57.9% for pattern b. When $c = .2$, both strategies performed worse. The whole-sample strategy consistently outperformed the split-data strategy in accepting the correct model without cross-loading, with 39% to 59.8% for pattern a and 3.4% to 19.9% for pattern b, compared with the split-data strategy (30.8% to 51% for pattern a and 7.3% to 19.1% for pattern b). In sum, the whole-sample strategy produced more accurate evaluations.

Impact of primary loading λ and factor correlation ρ. Results showed that the whole-sample strategy outperformed the split-data strategy in nearly all combinations of λ (.4 and .7) and ρ (.3 and .6). Both strategies displayed high accuracy under small ρ (.3) combined with big λ (.7). The whole-sample strategy exhibited higher accuracy rates, yielding 39% to 75.3% for pattern a and 15.9% to 47.9% for pattern b. The split-data strategy exhibited lower accuracy rates, resulting in 30.8% to 69.7% for pattern a and 16.4% to 49.4% for pattern b.

When the ρ and λ were small (.3 and .4), accuracy rates for both strategies were low because λ and c had similar values, which caused confusion. When $c = .2$ and $\lambda = .4$, correctly evaluating and rejecting the cross-loading was difficult because of the similarity values of c and λ, resulting in lower than 15% pattern a and lower than 15% pattern b proportions in both strategies. For $c = .4$, although the confusion in the model persisted, correctly accepting the model with one cross-loading was not as difficult. The whole-sample strategy appeared to be more effective than the split-data strategy, yielding 17.6% to 54.7% for pattern a and 34.8% to 70.3% for pattern b. The split-data strategy produced 5.8% to 45.3% for pattern a and 16.5% to 72.6% for pattern b, showing high error rates in evaluating if the cross-loading should be kept or not.

For a λ value of .7 and a ρ of .6, the whole-sample strategy is still suggested because of relatively stable performance in detecting the number of factors (26.7% to 60.9% for pattern a). The split-data strategy displayed less accurate results in evaluating if the cross-loading should be kept or not, with 24.4% to 56.8% for pattern a.

When the ρ were large (.6) but the λ were small (.4), both strategies performed terribly when c = .2 (pattern a lower than 25% and pattern b lower than 20%). In evaluating if a big c (.4) should be kept or not, the whole-sample strategy outperformed the split-data strategy, providing 7.2% to 35.3% for pattern a and 49% to 91.3% for pattern b. The split-data strategy yielded worse performance, with 4.5% to 27.1% of pattern a and 18.8% to 84.4% of pattern b. Both strategies had high rates of conflicting results (the EFA and CFA disagreed on the existence of cross-loading), but the whole-sample strategy was preferable.

4.2.2 M7 and M8

As discussed before in M5 and M6, when λ = .4, low accuracy rates occurred regardless of the strategy used in the analysis because the value of c_1 and c_2 (.2 or .4) was quite close to λ. Still, only the simulations with λ = .7 were used when analyzing the impact of n (Figure 8).

Figure 8 presents the proportion of replications that support different analysis strategies in M7 and M8. Figure 8 contains the three combinations of cross-loadings: a) c_1 and c_2 were both small (.2), the expected model contained no cross-loading; b) the expected model contained only one cross-loading (c_1 or c_2); c) c_1 and c_2 were both relatively big (.4), the expected model contained both c_1 and c_2.

Impact of sample size (n). The sample size (n) affected the effectiveness of the two strategies, as shown in Figure 8. It is worth noting that both strategies performed worse in evaluating if the cross-loading should be kept or not in two cross-loadings models (M7 and M8) compared to one cross-loading models (M5 and M6).

When n = 200, all strategies yielded low proportions of pattern a. The split-data strategy exhibited 2.8% to 21.6% for pattern a than the whole-sample strategy (3.1% to 35.9%). Both strategies resulted in high rates of conflict results (pattern b), yielding 19.8% to 96.5% in the whole-sample strategy and 18.2% to 82.6% in the split-data strategy. Model structures obtained under small samples need to be carefully examined before accepting.

When n = 1,000, both strategies yielded higher proportions of pattern a and pattern b. The whole-sample strategy yielded higher pattern a and pattern b proportions (0% to 40.9% for pattern a and 22.8% to 100% for pattern b) than the split-data strategy (0.3% to 38.5% for pattern a and 21.3% to 91.7% for pattern b). The whole-sample strategy was superior to the split-data strategy in all sample sizes.

Impact of the cross-loading c_1 and c_2. When both c_1 and c_2 were small (.2), the final model should contain neither c_1 nor c_2. However, when $\lambda = .4$, both strategies could hardly refuse c_1 and c_2 correctly and accept the correct model with no cross-loading because the value of cross-loadings (.2) was relatively close to λ, resulting in lower than 5% proportions of pattern a and pattern b in both strategies. For $\lambda = .7$, the whole-sample strategy performed similarly to the split-data strategy, yielding 1.4% to 13.8% of pattern a and 19.8% to 35.9% of pattern b, followed by the split-data strategy (4.6% to 13.7% of pattern a and 18.2% to 28.7% of pattern b). Both strategies performed poorly under $c_1 = .2$ and $c_2 = .2$. When c_1 valued .4, and c_2 valued .2 (or vice versa), both strategies still performed low accuracy in evaluating if the cross-loading should be kept or not when the $\lambda = .4$, resulting in lower than 10% pattern a rates and lower than 25% pattern b rates in both strategies. For $\lambda = .7$, the whole-sample strategy yielded more effective performance, exhibiting 17.4% to 40.9% in pattern a and 20.1% to 44.2% in pattern b. The split-data strategy was consistently less accurate, yielding 13.4% to 38.5% in pattern a and 19.5% to 35% in pattern b.

The final model should include two cross-loadings when both c_1 and c_2 were .4. Two strategies resulted in similar proportions of pattern a (0% to 23.8% for the whole-sample strategy and 0.3% to 23.9% for the split-data strategy). However, the whole-sample strategy had higher proportions of pattern b (57.7% to 100%, compared with 18.3% to 97.2% in the split-data strategy). It is worth noting that there were large proportions of pattern b in both strategies, implying that high percentages of conflict results occurred, which caused problems when explaining the results.

Overall, both strategies performed poorly in models with two cross-loadings, resulting in low proportions of pattern a. Suppose researchers decided to keep the number of cross-loadings obtained from the confirmatory analysis (CFA) and believed that the strategy is supported as long as the final result was correct. In that case, the final accuracy rates were the sum of the proportions of pattern a and pattern b, and the whole-sample strategy still outperformed the split-data strategy in all cross-loading conditions.

4.2.3 Summary

Overall, the whole-sample strategy was superior to the split-data strategy in correctly evaluating if the cross-loadings should be kept or not in both one and two cross-loading models. The applicability of the split-data strategy was not as effective as the whole-sample strategy and only performed relatively well under large samples (1,000). It is worth noting that both strategies showed higher error rates when analyzing models with cross-loadings (incorrectly retaining small cross-loadings or incorrectly rejecting large cross-loadings). Relatively speaking, the performance of both strategies became better under big primary loadings and cross-loadings in large samples.

Chapter 5

Discussion

This thesis research used simulated datasets to evaluate the effectiveness of the split-data strategy in determining the number of factors and cross-loadings. By comparing the split-data strategy and the whole-sample strategy, this thesis research provides an understanding of the consequences of applying the split-data strategy.

In determining the number of factors:

1. 1. Results showed little merit in the use of the split-data strategy because the whole-sample strategy consistently outperforms the split-data strategy in all simulation conditions, regardless of the sample size, loading, factor correlation, item number, and level of nonnormality. The only situation where researchers could employ the split-data strategy is when the sample size is large. However, the two strategies show almost identical results even with large sample sizes of 1,000. Although the application of the split-data strategy is used for cross-validation and reduces capitalization on chance, the reduction of sample size actually increases the capitalization on chance and makes the results less accurate.

2. 2. Both strategies performed well in detecting the number of factors under a sufficient sample size (1,000), with big loadings and small factor correlations. For conditions with small samples and low data quality (small loadings and big factor correlations), neither of these strategies works well.

In evaluating the existence of cross-loadings:

1. Both strategies result in low accuracy rates even in large samples based on the simulation results when evaluating the cross-loadings. Nevertheless, it is worth noting that the low accuracy is not

because of the strategies themselves but because the cross-loadings adopted in the simulation design are close to the criterion value of retaining the cross-loading (.3), which resulted in erroneously retaining small cross-loadings (.2) or mistakenly emitting large cross-loadings (.4). However, the whole-sample strategy still outperformed the split-data strategy, resulting in more accurate results when determining the existence of cross-loadings.

2. Both strategies resulted in more accurate evaluations under big primary loadings, large sample sizes, and big cross-loadings. When the number of cross-loadings changed from one to two, both strategies performed worse. In all simulation conditions, the whole-sample strategy is more accurate than the split-data strategy.

In summary, the whole-sample strategy provided more accurate results in determining the number of factors and evaluating the existence of cross-loadings. The split-data strategy, however, is only applicable under conditions with large samples (greater than 1,000 for the investigated models), large primary loadings, no cross-loading, and small factor correlations.

The split-data strategy has several disadvantages that can lead to inaccurate conclusions regarding the final model structure. The first disadvantage is that the sample size reduces substantially. People apply the split-data strategy for cross-validation to avoid capitalization by chance caused by one dataset. However, the sample size reduction in the split-data strategy increases capitalization by chance and leads to inappropriate models. When conducting the split-data strategy, the sample size for both EFA and CFA reduces to half, leading to less accurate evaluations of the factor number and cross-loadings. Although not listed in the result part, conducting the split-data strategy also led to higher standard errors, less accurate model-data fit evaluation, and higher nonconvergence rates. In order to deal with the sample size reduction problem, the split-data strategy requires at least 1,000 observations to provide stable evaluations for the models investigated in the present work. Given that behavioral research frequently has sample sizes within 1,000, using the split-data strategy is likely to result in less appropriate factor structures. In the whole-sample strategy, however, 400 appears to be a sufficient sample size to draw relatively accurate conclusions regarding the number of factors for the investigated models.

How large should the sample size be enough to justify the use of the split-data strategy? Considering that real datasets are more complicated than simulated datasets, missing values and nonnormally distributed data can frequently cause difficulty in the analysis when using the split-data strategy. As a result, researchers may need even bigger samples to have a stable evaluation of the model in the split-data strategy. As discussed in this thesis research, results based on the present work suggest that future research utilizes a sample size of more than 1,000 (based on the four models in this thesis), or even larger, if data and model complexities

(e.g., missingness, nonnormality, small loadings) exist.

The second disadvantage of the split-data strategy is that conflicting results are more likely to occur. When EFA and CFA results are inconsistent, difficulty will arise in interpretation. For instance, if the EFA fails to suggest the correct number of factors, but the following CFA supports a model with the correct number of factors (i.e., pattern b), researchers may support the model if they decide to keep the number of factors or cross-loadings obtained from the confirmatory analysis (CFA). When EFA suggests a correct model but CFA fails to cross-validate the results obtained from EFA (i.e., pattern c), the final model could be incorrect if researchers place more emphasis on the CFA results. If both EFA and CFA suggest the wrong models (i.e., pattern d), researchers cannot even get an appropriate model. Conflicting results are more likely to occur using the split-data strategy than the whole-sample strategy, leading to confusing and wrong models.

Lastly, using the split-data strategy can increase the replication crisis (e.g., Shrout & Rodgers, 2018) in behavioral research because different random seeds for splitting the data can lead to different results. In practice, researchers may test other seeds until they find two random datasets that are ideal enough to provide consistent model structures from both EFA and CFA. However, this practice is not acceptable because reliable results should not depend on the seed of the random dataset generator (Cavalloro et al., 2007). Given the same dataset, a different researcher may conclude different models because of different random seeds, making the results not replicable.

Despite the above disadvantages, one may question if the split-data strategy has any advantage of cross-validation over the whole-sample strategy in detecting the model structure. If the model obtained from the first half of the data can be cross-validated by the other half of the data, it appears that the model tends to be more trustworthy and can be generalized to other samples. However, whether the split-data strategy can be applied to cross-validation is still an open question because regardless of how researchers split the data, the two halves of the data are quite similar subsamples and share the same experiment flaws. For example, Podsakoff et al. (2003) discussed a series of common method biases in behavioral research, such as social desirability, item ambiguity, and reverse-coded items. Cross-validation is not meaningful when both halves of the data suffer from the same method biases because the split-data strategy essentially uses results from a flawed experiment to cross-validate the results from the experiment flawed in exactly the same manner. From this perspective, applying data from different sources to perform cross-validation is a more appropriate choice (Cabrera-Nguyen, 2010; Hinkin, 1995; Ssebuggwawo et al., 2010). However, if only one dataset is available, the whole-sample strategy provides more accurate conclusions regarding the model structure based on the simulation results. Split-data strategy leads to similar results to the whole-sample strategy only if the sample size is sufficient, which depends on the complexity of the model.

There remains room for improvement and refinement of the current study. First, the maximum

sample size applied in this thesis research is 1,000, a relatively large sample in behavioral research. However, the split-data strategy may require a higher sample size to achieve more accurate model estimations in practice. For example, based on the articles in the literature search, the commonly used sample sizes for the split-data strategy are 1,000, 2,000, or more, depending on the complexity of the model (Table 1). In addition, the existence of cross-loadings and small primary loadings requires even large samples to avoid inaccurate estimations and higher error rates. In this research, I only focus on the performance of the split-data strategy on sample sizes that are commonly seen in behavioral research. Future research should consider the performance of the split-data strategy on larger samples to draw broader conclusions. Furthermore, the present research conducted several standard factor analysis models with equal loadings and equal factor correlations. Future research can introduce more complex models (e.g., bi-factor model structures or higher-order CFA models) to examine the consequences of using the split-data strategy.

Second, the present thesis research simulated datasets with different skewness and kurtosis in Study 1. However, the present research only employed the normal-theory maximum likelihood estimation (ML) for all simulations. Further research can consider the use of robust estimation methods such as robust maximum likelihood (MLR; Muthén & Muthén, 2010; Li, 2015), Satorra-Bentler chi-square (MLM; Chou et al., 1991; Savalei, 2018), and Satorra-Bentler adjusted chi-square statistic (MLMV; Nevitt & Hancock, 2000; Yuan et al., 2005).

Third, in the present simulation studies, I fit the correct models to the data for both EFA and CFA. In practice, researchers typically apply EFA using half of the sample to obtain a model and then evaluate this model using the second half of the sample. Ideally, the models for CFA in the simulation should be derived from the EFA, but such a design would greatly complicate the study. In future research, simulations based on the model in EFA rather than the correct models are worth exploring, which can reveal the performance of the split-data strategy in practical applications.

Fourth, it is worth noting that this study only addressed the performance of the split-data strategy in EFA and CFA, and adopted model structures with no noise structure (e.g., residual covariance or trivial factors). However, some researchers proposed that when applied in real datasets, the use of the whole-sample strategy with only a single sample can potentially lead to more errors in determining the number of factors than the split-data strategy due to excessive statistical noise (e.g., Zhang, 2007). Another possible direction is to evaluate the split-data strategy in real datasets with statistical noise to see if the application of the split-data strategy for cross-validation can benefit researchers to avoid the influence of spurious random noise (Zhang & Stout, 1999).

Lastly, this thesis applied the traditional split-data strategy, but estimation for this strategy is not quite efficient because subsamples for both calibration and validation are much smaller than the complete

data set (de Rooij & Weeda, 2020). Further research can apply a more advanced split-data strategy and reuse the dataset to get more accurate results, such as two-sample cross-validation (Mosier, 1951), leave-one-out cross-validation (Stone, 1974), and K-fold cross-validation (Mosteller & Tukey, 1968). Specifically, researchers can split the dataset into K independent datasets. Each new dataset can be a validation set for the other $K-1$ sets. Currently, the K-fold cross-validation serves as a beneficial procedure in model selection (Refaeilzadeh et al., 2009). These methods can potentially provide a more stable evaluation of the model structures.

Chapter 6

Conclusion

The application of the split-data strategy is for cross-validation to avoid capitalization on chance. However, based on the two simulation studies in this thesis, the split-data strategy can provide less accurate conclusions regarding the number of factors and cross-loadings than the whole-sample strategy because of the sample size reduction after splitting the data. Applying the whole-sample strategy is preferred in small to medium samples based on the simulation conditions. With sufficient samples (e.g., greater than 1,000 for the models evaluated) and good model quality (e.g., large loadings, no cross-loading, and many items measuring a factor), using the split-data strategy is acceptable.

References

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Brown, A., Barker, E. D., & Rahman, Q. (2022). Development and psychometric validation of the Sexual Fantasies and Behaviors Inventory. *Psychological Assessment*, *34*(3), 217–232. https://doi.org/10.1037/pas0001082

Browne, M. W. (2001). An Overview of Analytic Rotation in Exploratory Factor Analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/s15327906mbr3601\{_}05

Buja, A., & Eyuboglu, N. (1992). Remarks on Parallel Analysis. *Multivariate Behavioral Research*, *27*(4), 509–540. https://doi.org/10.1207/s15327906mbr2704\{_}2

Burke, K., Dittman, C. K., Haslam, D., & Ralph, A. (2021). Assessing critical dimensions of the parent–adolescent relationship from multiple perspectives: Development and validation of the Parent-Adolescent Relationship Scale (PARS). *Psychological Assessment*, *33*(5), 395–410. https://doi.org/10.1037/pas0000992

Cabrera-Nguyen, P. (2010). Author Guidelines for Reporting Scale Development and Validation Results in the Journal of the Society for Social Work and Research. *Journal of the Society for Social Work and Research*, *1*(2), 99–103. https://doi.org/10.5243/jsswr.2010.8

Caleon, I. S., & King, R. B. (2021). Examining the Phenomenon of Resilience in Schools. *European Journal of Psychological Assessment*, *37*(1), 52–64. https://doi.org/10.1027/1015-5759/a000572

Cangur, S., & Ercan, I. (2015). Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality. *Journal of Modern Applied Statistical Methods*, *14*(1), 152–167. https://doi.org/10.22237/jmasm/1430453580

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102\{_}10

Cavalloro, P., Gendarme, C., Kronlöf, K., Mermet, J., Sas, V., Tiensyrjä, K., Voros, N., & van Sas, J. (2007). *System Level Design Model with Reuse of System IP*. Springer Publishing.

Chan, W. T., Bull, R., Ng, E. L., Waschl, N., & Poon, K. K. (2021). Validation of the Child Behavior Rating Scale (CBRS) using multilevel factor analysis. *Psychological Assessment*, *33*(11), 1138–1151. https://doi.org/10.1037/pas0001075

Child, D. (2006). *The Essentials of Factor Analysis* (3rd ed.). Bloomsbury Academic.

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 347–357. https://doi.org/10.1111/j.2044-8317.1991.tb00966.x

Çokluk Bökeoğlu, Ö., & Koçak, D. (2016). Using horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educational Sciences: Theory & Practice*, *16*, 537–551. https://doi.org/10.12738/estp.2016.2.0328

Colledani, D., Meneghini, A. M., Mikulincer, M., & Shaver, P. R. (2021). The Caregiving System Scale. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000673

Cooper, A. J., Smillie, L. D., & Corr, P. J. (2010). A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality and Individual Differences*, *48*(5), 688–691. https://doi.org/10.1016/j.paid.2010.01.004

Courtney, M. (2013). Determining the number of factors to retain in efa: Using the spss r-menu v2.0 to make more judicious estimations. *Practical Assessment, Research and Evaluation*, *18*, 1–14. https://doi.org/10.2147/JHL.S35483

Crasta, D., Rogge, R. D., Maniaci, M. R., & Reis, H. T. (2021). Toward an optimized measure of perceived partner responsiveness: Development and validation of the perceived responsiveness and insensitivity scale. *Psychological Assessment*, *33*(4), 338–355. https://doi.org/10.1037/pas0000986

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. https://doi.org/10.1177/0013164410379332

de Rooij, M., & Weeda, W. (2020). Cross-Validation: A Method Every Psychologist Should Know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Doherty, A. S., Mallett, J., Leiter, M. P., & McFadden, P. (2021). Measuring Burnout in Social Work: Factorial validity of the Maslach Burnout Inventory—Human Services Survey. *European Journal of Psychological Assessment*, *37*(1), 6–14. https://doi.org/10.1027/1015-5759/a000568

Eichenbaum, A. E., Marcus, D. K., & French, B. F. (2021). Item response theory analysis of the Triarchic Psychopathy Measure. *Psychological Assessment*, *33*(8), 766–776. https://doi.org/10.1037/pas0001022

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989x.4.3.272

Finch, W. H. (2019). Using Fit Statistic Differences to Determine the Optimal Number of Factors to Retain in an Exploratory Factor Analysis. *Educational and Psychological Measurement*, *80*(2), 217–241. https://doi.org/10.1177/0013164419865769

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Fokkema, M., & Greiff, S. (2017). How Performing PCA and CFA on the Same Data Equals Trouble. *European Journal of Psychological Assessment*, *33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Foldnes, N., & Olsson, U. H. (2016). A Simple Simulation Technique for Nonnormal Data with Prespecified Skewness, Kurtosis, and Covariance Matrix. *Multivariate Behavioral Research*, *51*(2-3), 207–219. https://doi.org/10.1080/00273171.2015.1133274

Franklin, S. B., Gibson, D. J., Robertson, P. A., Pohlmann, J. T., & Fralish, J. S. (1995). Parallel Analysis: a method for determining significant principal components. *Journal of Vegetation Science*, *6*(1), 99–106. https://doi.org/10.2307/3236261

Fu, Y., Wen, Z., & Wang, Y. (2021). A Comparison of Reliability Estimation Based on Confirmatory Factor Analysis and Exploratory Structural Equation Models. *Educational and Psychological Measurement*, *82*(2), 205–224. https://doi.org/10.1177/00131644211008953

Garofalo, C., Weller, J. A., Kirisci, L., & Reynolds, M. D. (2021). Elaborating on the longitudinal measurement invariance and construct validity of the triarchic psychopathy scales from the Multidimensional Personality Questionnaire. *Psychological Assessment*, *33*(9), 890–903. https://doi.org/10.1037/pas0001023

Gatignon, H. (2009). Confirmatory Factor Analysis. *Statistical Analysis of Management Data*, 59–122. https://doi.org/10.1007/978-1-4419-1270-1\{_}4

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 62–72. https://doi.org/10.1080/10705519609540030

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. https://doi.org/10.1177/0013164495055003002

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Gorsuch, R. (1983). *Factor Analysis, 2nd Edition* (2nd ed.). Lawrence Erlbaum Associates.

Green, S. B., Redell, N., Thompson, M. S., & Levy, R. (2015). Accuracy of Revised and Traditional Parallel Analyses for Assessing Dimensionality with Binary Data. *Educational and Psychological Measurement*, *76*(1), 5–21. https://doi.org/10.1177/0013164415581898

Hall, J. A., Steele, R. G., Christofferson, J. L., & Mihailova, T. (2021a). Development and initial evaluation of a multidimensional digital stress scale. *Psychological Assessment*, *33*(3), 230–242. https://doi.org/10.1037/pas0000979

Hall, J. A., Steele, R. G., Christofferson, J. L., & Mihailova, T. (2021b). Development and initial evaluation of a multidimensional digital stress scale. *Psychological Assessment*, *33*(3), 230–242. https://doi.org/10.1037/pas0000979

Henson, R. K., & Roberts, J. K. (2006). Use of Exploratory Factor Analysis in Published Research. *Educational and Psychological Measurement*, *66*(3), 393–416. https://doi.org/10.1177/0013164405282485

Hinkin, T. R. (1995). A Review of Scale Development Practices in the Study of Organizations. *Journal of Management*, *21*(5), 967–988. https://doi.org/10.1177/014920639502100509

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/bf02289447

Hornsby, B. W. Y., Camarata, S., Cho, S.-J., Davis, H., McGarrigle, R., & Bess, F. H. (2021). Development and validation of the Vanderbilt Fatigue Scale for Adults (VFS-A). *Psychological Assessment*, *33*(8), 777–788. https://doi.org/10.1037/pas0001021

Howard, M. C. (2015). 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. https://doi.org/10.1080/10447318.2015.1087664

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

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. https://doi.org/10.1207/s15327906mbr1002\{_}5

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *Journal*

of Organizational Behavior, *18*(6), 667–683. https://doi.org/10.1002/(SICI)1099-1379(199711)18:63.0.CO;2-T

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/bf02289343

Joyner, K. J., Daurio, A. M., Perkins, E. R., Patrick, C. J., & Latzman, R. D. (2021). The difference between trait disinhibition and impulsivity—and why it matters for clinical psychological science. *Psychological Assessment*, *33*(1), 29–44. https://doi.org/10.1037/pas0000964

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kupper, K., Krampen, D., Rammstedt, B., & Rohrmann, S. (2021). The German-Language Short Form of the Big Five Inventory for Children and Adolescents – Other-Rating Version (BFI-K KJ-F). *European Journal of Psychological Assessment*, *37*(2), 109–117. https://doi.org/10.1027/1015-5759/a000592

Lauriola, M., Donati, M. A., Trentini, C., Tomai, M., Pontone, S., & Baker, R. (2021). The Structure of the Emotional Processing Scale (EPS-25). *European Journal of Psychological Assessment*, *37*(6), 423–432. https://doi.org/10.1027/1015-5759/a000632

Li, C.-H. (2015). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, L. Y., Cicero, D. C., Dodell-Feder, D., Germine, L., & Martin, E. A. (2021). Comparability of social anhedonia across epidemiological dimensions: A multinational study of measurement invariance of the Revised Social Anhedonia Scale. *Psychological Assessment*, *33*(2), 171–179. https://doi.org/10.1037/pas0000972

Li, Y., Wen, Z., Hau, K.-T., Yuan, K.-H., & Peng, Y. (2020). Effects of Cross-loadings on Determining the Number of Factors to Retain. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 841–863. https://doi.org/10.1080/10705511.2020.1745075

Lim, S., & Jahng, S. (2019). Determining the number of factors using parallel analysis and its recent variants. *Psychological Methods*, *24*(4), 452–467. https://doi.org/10.1037/met0000230

Liu, D., Kahathuduwa, C., & Vazsonyi, A. T. (2021). The Pittsburgh Sleep Quality Index (PSQI): Psychometric and clinical risk score applications among college students. *Psychological Assessment*, *33*(9), 816–826. https://doi.org/10.1037/pas0001027

Lucas, N. (2004). *Effects of cross-loading items on convergence, variability of parameter estimates, and goodness-of-fit indices for confirmatory factor analysis* (Doctoral dissertation). University of New Mexico.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mai, Y., Zhang, Z., & Wen, Z. (2018). Comparing Exploratory Structural Equation Modeling and Existing Approaches for Multiple Regression with Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 737–749. https://doi.org/10.1080/10705511.2018.1444993

Maietta, J. E., Ahmed, A. O., Barchard, K. A., Kuwabara, H. C., Donohue, B., Ross, S. R., Kinsora, T. F., & Allen, D. N. (2021). Confirmatory factor analysis of imPACT cognitive tests in high school athletes. *Psychological Assessment*, *33*(8), 746–755. https://doi.org/10.1037/pas0001014

Mansolf, M., Blackwell, C. K., Cummings, P., Choi, S., & Cella, D. (2022). Linking the Child Behavior Checklist to the Strengths and Difficulties Questionnaire. *Psychological Assessment*, *34*(3), 233–246. https://doi.org/10.1037/pas0001083

Marsh, H., Balla, J., & Hau, K.-T. (1996). An evaluation of incremental fit indexes: A clarification of mathematical and empirical properties. *Advanced Structural Modeling Techniques*, 315–353.

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

McDonald, R. P. (1985). *Factor Analysis and Related Methods* (1st ed.). Psychology Press.

McHugh, M. L. (2013). The Chi-square test of independence. *Biochemia Medica*, 143–149. https://doi.org/10.11613/bm.2013.018

McLeod, B. D., Cecilione, J., Jensen-Doss, A., Southam-Gerow, M. A., & Kendall, P. C. (2021). Reliability, factor structure, and validity of an observer-rated alliance scale with youth. *Psychological Assessment*, *33*(10), 1013–1023. https://doi.org/10.1037/pas0001036

Moreira, P. A. S., Ramalho, S., & Inman, R. A. (2021). The Engagement/Disengagement in Sustainable Development Inventory (EDiSDI). *European Journal of Psychological Assessment*, *37*(5), 344–356. https://doi.org/10.1027/1015-5759/a000619

Mosier, C. I. (1951). Problems and Designs of Cross-Validation. *Educational and Psychological Measurement*, *11*(1), 5–11. https://doi.org/10.1177/001316445101100101

Muthén, L. K., & Muthén, B. O. (2012). *User's Guide - Mplus* (7th ed.). Los Angeles, CA.

Nevitt, J., & Hancock, G. R. (2000). Improving the Root Mean Square Error of Approximation for Nonnormal Conditions in Structural Equation Modeling. *The Journal of Experimental Education*, *68*(3), 251–268. https://doi.org/10.1080/00220970009600095

Nordgren, L., Ghaderi, A., Ljótsson, B., & Hesser, H. (2021). Identifying subgroups of patients with eating disorders based on emotion dysregulation profiles: A factor mixture modeling approach to classification. *Psychological Assessment*. https://doi.org/10.1037/pas0001103

Orcan, F. (2018). Exploratory and Confirmatory Factor Analysis: Which One to Use First? *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 414–421. https://doi.org/10.21031/epod.394323

Ortet-Walker, J., Mezquita, L., Vidal-Arenas, V., Ortet, G., & Ibáñez, M. I. (2022). Development of a 50-Item Abridged Form of the Junior Spanish Version of the NEO Questionnaire (JS NEO-A50). *European Journal of Psychological Assessment*, *38*(2), 101–112. https://doi.org/10.1027/1015-5759/a000648

Osborne, J. (2014). *Best Practices in Exploratory Factor Analysis (Best Practices in Quantitative Methods)*. CreateSpace Independent Publishing Platform.

Osborne, J., & Fitzpatrick, D. (2012). Replication analysis in exploratory factor analysis: What it is and: Why it makes your analysis better. *Practical Assessment, Research and Evaluation*, *17*, 1–8. https://doi.org/10.7275/h0bd-4d11

Ozkok, O., Zyphur, M. J., Barsky, A. P., Theilacker, M., Donnellan, M. B., & Oswald, F. L. (2019). Modeling Measurement as a Sequential Process: Autoregressive Confirmatory Factor Analysis (AR-CFA). *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02108

Partsch, M. V., & Danner, D. (2021). Measuring Self-Control in International Large-Scale Surveys. *European Journal of Psychological Assessment*, *37*(5), 409–418. https://doi.org/10.1027/1015-5759/a000618

Paulhus, D. L., Buckels, E. E., Trapnell, P. D., & Jones, D. N. (2021). Screening for Dark Personalities: The Short Dark Tetrad (SD4). *European Journal of Psychological Assessment*, 1–15. https://doi.org/10.1027/1015-5759/a000602

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. *Encyclopedia of Database Systems*, 532–538. https://doi.org/10.1007/978-0-387-39940-9\{_}565

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.2.5]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Revelle, W., & Rocklin, T. (1979). Very Simple Structure: An Alternative Procedure For Estimating The Optimal Number Of Interpretable Factors. *Multivariate Behavioral Research*, *14*(4), 403–414. https://doi.org/10.1207/s15327906mbr1404\{_}2

Rogoza, R., Cieciuch, J., Strus, W., & Kłosowski, M. (2021). Investigating the structure of the Polish Five Factor Narcissism Inventory: Support for the three-factor model of narcissism. *Psychological Assessment*, *33*(3), 267–272. https://doi.org/10.1037/pas0000901

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Savalei, V. (2018). On the Computation of the RMSEA and CFI from the Mean-And-Variance Corrected Test Statistic with Nonnormal Data in SEM. *Multivariate Behavioral Research*, *53*(3), 419–429. https://doi.org/10.1080/00273171.2018.1455142

Schmitt, T. A., & Sass, D. A. (2011). Rotation Criteria and Hypothesis Testing for Exploratory Factor Analysis: Implications for Factor Pattern Loadings and Interfactor Correlations. *Educational and Psychological Measurement*, *71*(1), 95–113. https://doi.org/10.1177/0013164410387348

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, *17*(5), 1004–1011. https://doi.org/10.1016/j.sapharm.2020.07.027

Sellbom, M., Liggins, C., Laurinaitytė, I., & Cooke, D. J. (2021). Factor structure of the Comprehensive Assessment of Psychopathic Personality-Self-Report (CAPP-SR) in community and offender samples. *Psychological Assessment*, *33*(10), 927–939. https://doi.org/10.1037/pas0001029

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Sorrel, M. A., Aluja, A., García, L. F., & Gutiérrez, F. (2022). Psychometric properties of the Five-Factor Personality Inventory for ICD-11 (FFiCD) in Spanish community samples. *Psychological Assessment*, *34*(3), 281–293. https://doi.org/10.1037/pas0001084

Ssebuggwawo, D., Hoppenbrouwers, S., & Proper, E. (2010). Assessing Collaborative Modeling Quality Based on Modeling Artifacts. *Lecture Notes in Business Information Processing*, 76–90. https://doi.org/10.1007/978-3-642-16782-9\{_}6

Stanton, K., Brown, M. F. D., McDanal, R., Carlton, C. N., & Emery, N. N. (2021). Informing the classification and assessment of positive emotional experiences: A multisample examination of hierarchical positive emotionality models. *Psychological Assessment*, *33*(11), 1038–1049. https://doi.org/10.1037/pas0001052

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502\{_}4

Stevens, J. (1995). *Applied Multivariate Statistics for the Social Sciences, Fifth Edition* (3rd ed.). Psychology Press.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Sun, J. (2005). Assessing Goodness of Fit in Confirmatory Factor Analysis. *Measurement and Evaluation in Counseling and Development*, *37*(4), 240–256. https://doi.org/10.1080/07481756.2005.11909764

Tabachnick, B., & Fidell, L. (2013). *Using Multivariate Statistics*. Pearson Education.

Thöne, A.-K., Junghänel, M., Görtz-Dorten, A., Dose, C., Hautmann, C., Jendreizik, L. T., Treier, A.-K., Vetter, P., von Wirth, E., Banaschewski, T., Becker, K., Brandeis, D., Dürrwächter, U., Geissler, J., Hebebrand, J., Hohmann, S., Holtmann, M., Huss, M., Jans, T., . . . Döpfner, M. (2021). Disentangling symptoms of externalizing disorders in children using multiple measures and informants. *Psychological Assessment*, *33*(11), 1065–1079. https://doi.org/10.1037/pas0001053

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. https://doi.org/10.1007/bf02291170

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327. https://doi.org/10.1007/bf02293557

Waddell, J. T., Corbin, W. R., Meier, M. H., Morean, M. E., & Metrik, J. (2021). The Anticipated Effects of Cannabis Scale (AECS): Initial development and validation of an affect- and valence-based expectancy measure. *Psychological Assessment*, *33*(2), 180–194. https://doi.org/10.1037/pas0000881

Wang, J., Shi, X., Zou, H., Pons, F., Xu, Q., Wang, Y., Tang, Y., & Jiang, S. (2021). Development and validation of the Emotional Intelligence Test for Adolescents in a Chinese sample. *Psychological Assessment*, *33*(12), 1200–1214. https://doi.org/10.1037/pas0001078

Watson, R., McCabe, C., Harvey, K., & Reynolds, S. (2021). Development and validation of a new adolescent self-report scale to measure loss of interest and pleasure: The Anhedonia Scale for Adolescents. *Psychological Assessment*, *33*(3), 201–217. https://doi.org/10.1037/pas0000977

Williams, L. J. (1995). Covariance structure modeling in organizational research: Problems with the method versus applications of the method. *Journal of Organizational Behavior*, *16*(3), 225–233. https://doi.org/10.1002/job.4030160305

Wilson, K. E., Almeida, F. A., Brito, F. A., Sweet, C. C., Katula, J. A., Michaud, T. L., Schwab, R., & Estabrooks, P. A. (2021). Psychometric assessment of the Brief Weight-Loss-Related Behavior Self-Efficacy Survey in adults with prediabetes. *Psychological Assessment*, *33*(11), 1089–1099. https://doi.org/10.1037/pas0001058

Worthington, R. L., & Whittaker, T. A. (2006). Scale Development Research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Xia, Y. (2021). Determining the Number of Factors When Population Models Can Be Closely Approximated by Parsimonious Models. *Educational and Psychological Measurement*, *81*(6), 1143–1171. https://doi.org/10.1177/0013164421992836

Yang, Y., & Green, S. B. (2010). A Note on Structural Equation Modeling Estimates of Reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(1), 66–81. https://doi.org/10.1080/10705510903438963

Yang, Y., & Xia, Y. (2014). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods*, *47*(3), 756–772. https://doi.org/10.3758/s13428-014-0499-2

Yuan, K.-H., Bentler, P. M., & Zhang, W. (2005). The Effect of Skewness and Kurtosis on Mean and Covariance Structure Analysis. *Sociological Methods & Research*, *34*(2), 240–258. https://doi.org/10.1177/0049124105280200

Zeelenberg, M., Seuntjens, T. G., van de Ven, N., & Breugelmans, S. M. (2021). Dispositional Greed Scales. *European Journal of Psychological Assessment*, 1–10. https://doi.org/10.1027/1015-5759/a000647

Zhang, J. (2007). Conditional Covariance Theory and Detect for Polytomous Items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213–249. https://doi.org/10.1007/bf02294536

Appendix A

Figures and Tables

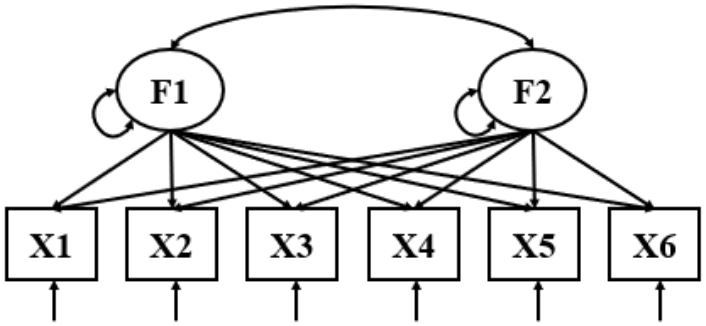

Figure 1: A 2-factor EFA model

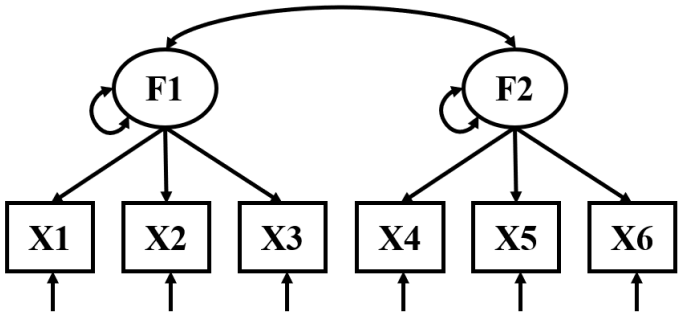

Figure 2: A 2-factor 6-item CFA model

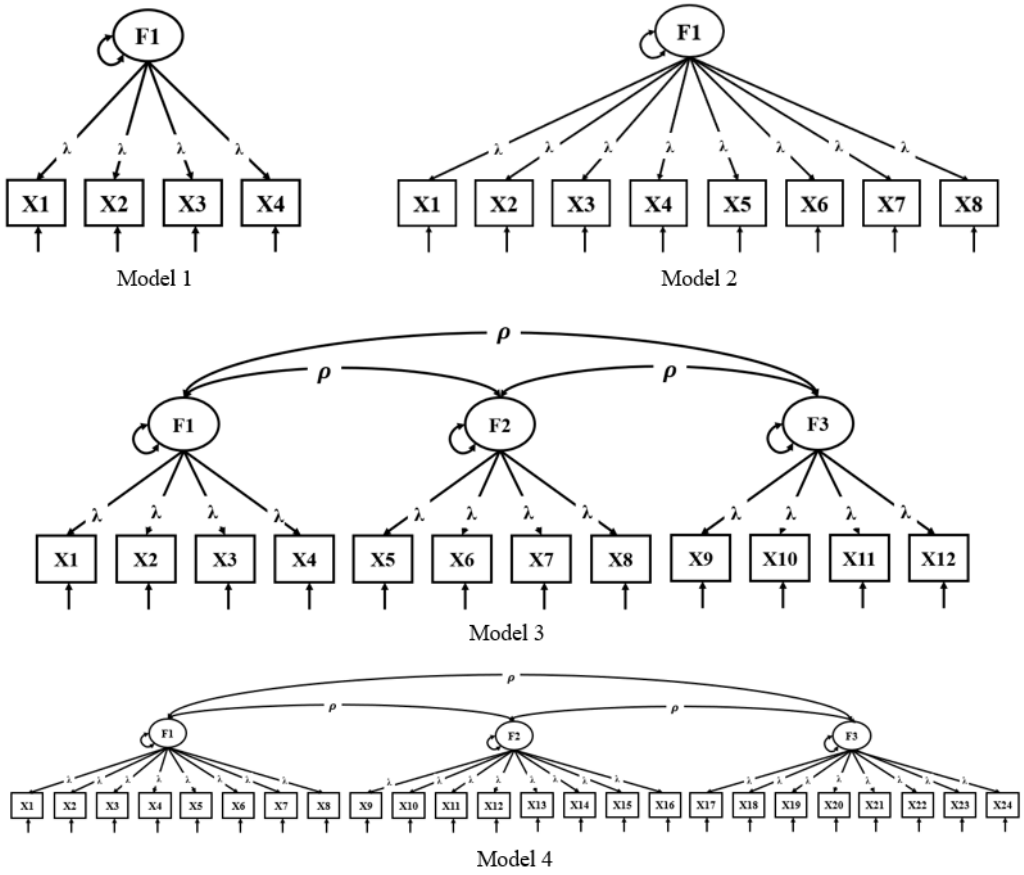

Figure 3: M1–M4 in Study 1

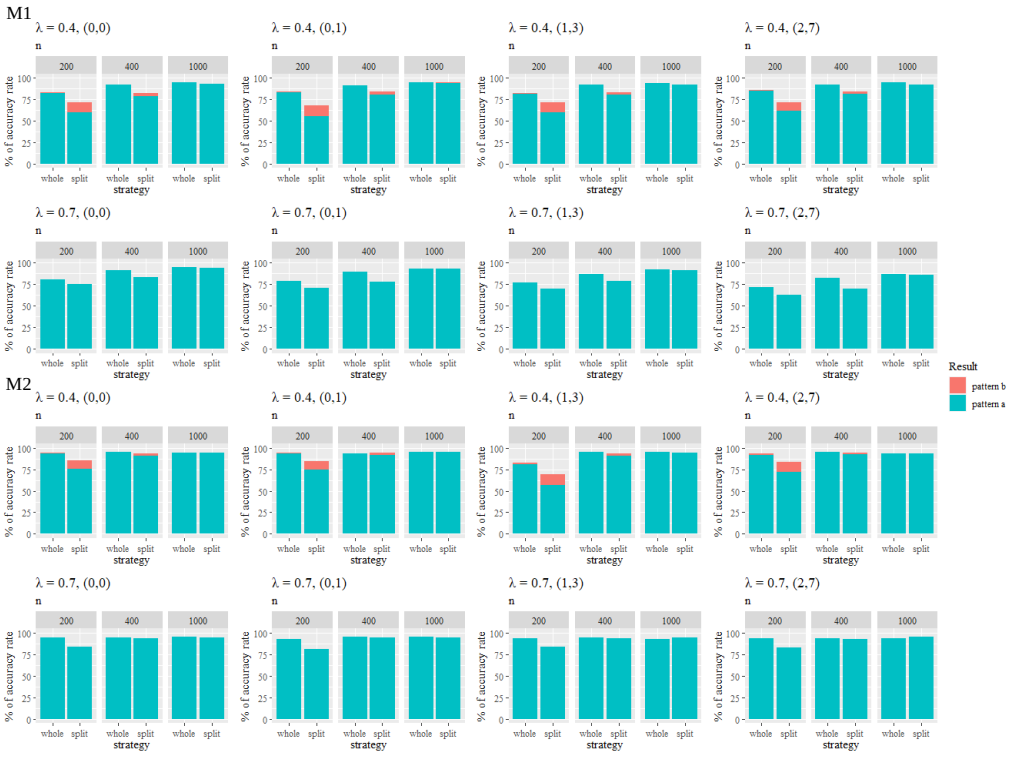

Figure 4: Proportions of Pattern a and b in M1–M2 in Study 1

Figure 5: Proportions of Pattern a and b in M3–M4 in Study 1

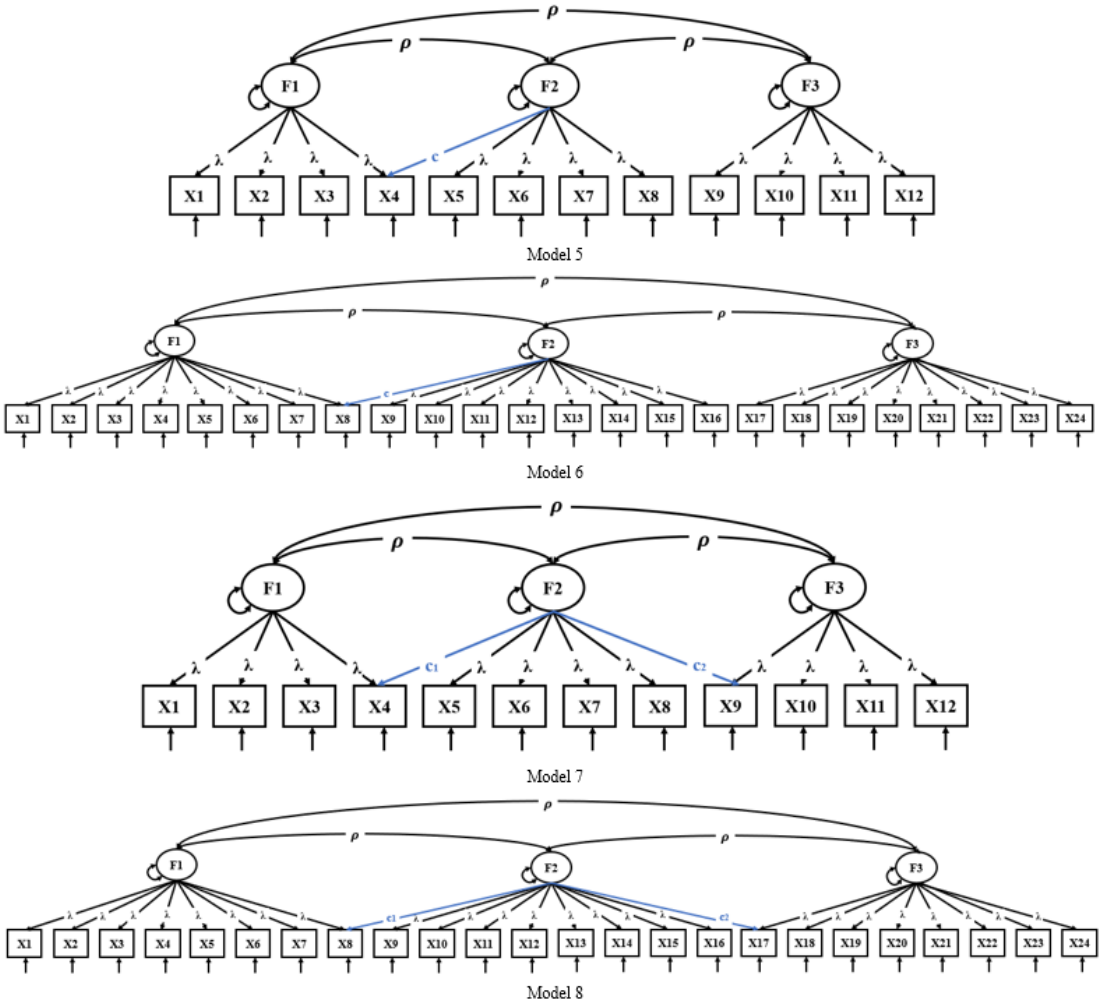

Figure 6: M5–M8 in Study 2

Figure 7: Proportions of Pattern a and b in M5 and M6 in Study 2

Figure 8: Proportions of Pattern a and b in M7 and M8 in Study 2

Table 1: Summary of Analytical Methods in Reviewed Articles

Title	Author	Analysis method	Sample size
Split data			
Identifying subgroups of patients with eating disorders based on emotion dysregulation profiles: A factor mixture modeling approach to classification	Nordgren, L., Ghaderi, A., Ljótsson, B., & Hesser, H.	split data	N = 857
Development and psychometric validation of the Sexual Fantasies and Behaviors Inventory	Brown, A., Barker, E. D., & Rahman, Q.	split data	N = 4280
The Pittsburgh Sleep Quality Index (PSQI): Psychometric and clinical risk score applications among college students	Liu, D., Kahathuduwa, C., & Vazsonyi, A. T.	split data	N = 976
Confirmatory factor analysis of imPACT cognitive tests in high school athletes	Maietta, J. E., Ahmed, A. O., Barchard, K. A., Kuwabara, H. C., Donohue, B., Ross, S. R., Kinsora, T. F., & Allen, D. N.	split data	N = 36091
Toward an optimized measure of perceived partner responsiveness: Development and validation of the perceived responsiveness and insensitivity scale	Crasta, D., Rogge, R. D., Maniaci, M. R., & Reis, H. T.	split data	N = 2334
Development and validation of a new adolescent self-report scale to measure loss of interest and pleasure: The Anhedonia Scale for Adolescents	Watson, R., McCabe, C., Harvey, K., & Reynolds, S.	split data	N = 2098
Examining the Phenomenon of Resilience in Schools	Caleon, I. S., & King, R. B.	split data	N = 1159
One factor analysis method			
Psychometric properties of the Five-Factor Personality Inventory for ICD-11 (FFiCD) in Spanish community samples	Sorrel, M. A., Aluja, A., García, L. F., & Gutiérrez, F.	EFA	N = 1409

Table 1(cont.)

Title	Author	Analysis method	Sample size
Informing the classification and assessment of positive emotional experiences: A multi-sample examination of hierarchical positive emotionality models	Stanton, K., Brown, M. F. D., McDanal, R., Carlton, C. N., & Emery, N. N.	Multi-EFA	N1 = 447, N2 = 375
Reliability, factor structure, and validity of an observer-rated alliance scale with youth	McLeod, B. D., Cecilione, J., Jensen-Doss, A., Southam-Gerow, M. A., & Kendall, P. C.	EFA	N = 51
Factor structure of the Comprehensive Assessment of Psychopathic Personality-Self-Report (CAPP-SR) in community and offender samples	Sellbom, M., Liggins, C., Laurinaitytė, I., & Cooke, D. J.	EFA	N1 = 960, N2 = 1047, N3 = 268
Item response theory analysis of the Triarchic Psychopathy Measure	Eichenbaum, A. E., Marcus, D. K., & French, B. F.	EFA	N = 937
Development and validation of the Vanderbilt Fatigue Scale for Adults (VFS-A)	Hornsby, B. W. Y., Camarata, S., Cho, S. J., Davis, H., McGarrigle, R., & Bess, F. H.	EFA	N = 580
Development of a 50-Item Abridged Form of the Junior Spanish Version of the NEO Questionnaire (JS NEO-A50)	Ortet-Walker, J., Mezquita, L., Vidal-Arenas, V., Ortet, G., & Ibáñez, M. I.	EFA	N1 = 400, N2 = 385
Dispositional Greed Scales	Zeelenberg, M., Seuntjens, T. G., van de Ven, N., & Breugelmans, S. M.	EFA	N1 = 300, N2 = 1000
Development and initial evaluation of a multidimensional digital stress scale	Hall, J. A., Steele, R. G., Christofferson, J. L., & Mihailova, T.	EFA	N1= 23, N2 = 247
Investigating the structure of the Polish Five Factor Narcissism Inventory: Support for the three-factor model of narcissism	Rogoza, R., Cieciuch, J., Strus, W., & Kłosowski, M.	EFA	N = 793

Table 1(cont.)

Title	Author	Analysis method	Sample size
The difference between trait disinhibition and impulsivity–and why it matters for clinical psychological science	Joyner, K. J., Daurio, A. M., Perkins, E. R., Patrick, C. J., & Latzman, R. D.	CFA	N1 = 400, N2 = 308
The German-Language Short Form of the Big Five Inventory for Children and Adolescents–Other-Rating Version (BFI-K KJ-F)	Kupper, K., Krampen, D., Rammstedt, B., & Rohrmann, S.	EFA	N = 258
Measuring Self-Control in International Large-Scale Surveys: Development and validation of a four-item scale in English, French, German, Japanese, Polish, and Spanish	Partsch, M. V., & Danner, D.	SEM	N1 = 973, N2 = 5557
EFA and CFA on different samples			
Development and validation of the Emotional Intelligence Test for Adolescents in a Chinese sample	Wang, J., Shi, X., Zou, H., Pons, F., Xu, Q., Wang, Y., Tang, Y., & Jiang, S.	EFA, CFA	EFA = 1536, CFA = 2568
The Caregiving System Scale	Colledani, D., Meneghini, A. M., Mikulincer, M., & Shaver, P. R.	EFA, CFA	EFA = 679, CFA = 742
Assessing critical dimensions of the parent–adolescent relationship from multiple perspectives: Development and validation of the Parent-Adolescent Relationship Scale (PARS)	Burke, K., Dittman, C. K., Haslam, D., & Ralph, A.	EFA, CFA	EFA = 256, CFA = 608
The Anticipated Effects of Cannabis Scale (AECS): Initial development and validation of an affect- and valence-based expectancy measure	Waddell, J. T., Corbin, W. R., Meier, M. H., Morean, M. E., & Metrik, J.	EFA, CFA	EFA = 303, CFA = 469

Table 1(cont.)

Title	Author	Analysis method	Sample size
Engagement/Disengagement in Sustainable Development Inventory (EDiSDI)	Moreira, P. A. S., Ramalho, S., & Inman, R. A.	EFA, CFA	EFA = 266, CFA = 510
Screening for dark personalities: The Short Dark Tetrad (SD4)	Paulhus, D. L., Buckels, E. E., Trapnell, P. D., & Jones, D. N.	EFA, CFA	EFA1 = 868, EFA2 = 999, CFA = 660
CFA and EFA (or ESEM) on the same datasets (whole-sample strategy)			
Linking the Child Behavior Checklist to the Strengths and Difficulties Questionnaire	Mansolf, M., Blackwell, C. K., Cummings, P., Choi, S., & Cella, D.	CFA EFA	N = 500
Validation of the Child Behavior Rating Scale (CBRS) using multilevel factor analysis	Chan, W. T., Bull, R., Ng, E. L., Waschl, N., & Poon, K. K.	Multi-CFA & EFA	N = 1375
Disentangling symptoms of externalizing disorders in children using multiple measures and informants	Thöne, A. K., Junghänel, M., Görtz-Dorten, A., Dose, C., Hautmann, C., Jendreizik, L. T., Treier, A. K., Vetter, P., von Wirth, E., Banaschewski, T., Becker, K., Brandeis, D., Dürrwächter, U., Geissler, J., Hebebrand, J., Hohmann, S., Holtmann, M., Huss, M., Jans, T., . . . Döpfner, M.	CFA, ESEM	N = 474
Psychometric assessment of the Brief Weight-Loss-Related Behavior Self-Efficacy Survey in adults with prediabetes	Wilson, K. E., Almeida, F. A., Brito, F. A., Sweet, C. C., Katula, J. A., Michaud, T. L., Schwab, R., & Estabrooks, P. A.	CFA, EFA	N = 599

Table 1(cont.)

Title	Author	Analysis method	Sample size
Elaborating on the longitudinal measurement invariance and construct validity of the triarchic psychopathy scales from the Multidimensional Personality Questionnaire	Garofalo, C., Weller, J. A., Kirisci, L., & Reynolds, M. D.	CFA, ESEM	N = 716
Comparability of social anhedonia across epidemiological dimensions: A multinational study of measurement invariance of the Revised Social Anhedonia Scale	Li, L. Y., Cicero, D. C., Dodell-Feder, D., Germine, L., & Martin, E. A.	CFA, ESEM	N = 14064
Measuring Burnout in Social Work: Factorial validity of the Maslach Burnout Inventory–Human Services Survey	Doherty, A. S., Mallett, J., Leiter, M. P., & McFadden, P.	CFA, ESEM	N = 1257
The Structure of the Emotional Processing Scale (EPS-25)	Lauriola, M., Donati, M. A., Trentini, C., Tomai, M., Pontone, S., & Baker, R.	CFA, ESEM	N = 350

Table 2: Possible Results in Split-data/Whole-sample.

	EFA/TPA	CFA
a: This strategy is supported	Correct number of factors	Correct
b: EFA/TPA fails to suggest the correct results, but the CFA will lead to the correct number of factors. The final result is still correct. However, the inconsistency results in difficulty in interpretation.	Incorrect	Correct
c: CFA fails to cross-validate the correct results obtained from EFA, or the CFA result is not converged; Additionally, the final conclusion is incorrect.	Correct	Incorrect
d: EFA/TPA and CFA fail to suggest the correct number of factors.	Incorrect	Incorrect

Table 3: Possible Results in Split-data/Whole-sample

	EFA	CFA
a: This strategy is supported	The model with the correct number of cross-loadings is acceptable	Acceptable
b: EFA fails to suggest the correct results, but CFA will lead to the correct evaluation of the existence of cross-loading(s). The final result is still correct. However, the inconsistency results in difficulty in interpretation.	Unacceptable	Acceptable
c: CFA fails to cross-validate the correct results obtained from EFA, or the CFA result is not converged; Additionally, the final conclusion is incorrect.	Acceptable	Unacceptable
d: Both EFA and CFA suggest that the model with the correct number of cross-loading(s) is not acceptable.	Unacceptable	Unacceptable