Journal of Psychoeducational Assessment
http://jpa.sagepub.com/

Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis

Thomas A. Schmitt

Journal of Psychoeducational Assessment 2011 29: 304
DOI: 10.1177/0734282911406653

Journal of Psychoeducational Assessment
29(4) 304–321
© 2011 SAGE Publications
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282911406653
http://jpa.sagepub.com

Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis

Thomas A. Schmitt[1]

Abstract

Researchers must make numerous choices when conducting factor analyses, each of which can have significant ramifications on the model results. They must decide on an appropriate sample size to achieve accurate parameter estimates and adequate power, a factor model and estimation method, a method for determining the number of factors and evaluating model fit, and a rotation criterion. Unfortunately, researchers continue to use outdated methods in each of these areas. The present article provides a current overview of these areas in an effort to provide researchers with up-to-date methods and considerations in both exploratory and confirmatory factor analysis. A demonstration was provided to illustrate current approaches. Choosing between confirmatory and exploratory methods is also discussed, as researchers often make incorrect assumptions about the application of each.

Keywords

factor analysis, exploratory factor analysis, confirmatory factor analysis, structural equation modeling

Using factor analysis (FA) procedures such as exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to investigate latent variables has become common for such areas as instrument development, longitudinal data analysis, comparing group means, and so on (see Cudeck & MacCallum, 2007). Despite more than 100 years existence of FA, both EFA and CFA remain popular and continue to be expanded and updated. Researchers are faced with numerous decisions when conducting FA, and the information for making these decisions is often scattered throughout the literature, difficult to understand, and/or inconsistent and inconclusive. Fortunately, there exist many general reviews of FA (e.g., Conway & Huffcut, 2003; Costello & Osborne, 2005; Fabrigar, Wegener, MacCallum, & Strahan, 1999; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009; Preacher & Maccallum, 2003; Worthington & Whittaker, 2006). Therefore, the goal of the this article is to build on these reviews of FA and provide an overview and illustration of current FA methods to help researchers decide on: (a) an appropriate sample size to achieve accurate parameter estimates and adequate power, (b) a factor

[1] Eastern Michigan University, Ypsilanti

Corresponding Author:
Thomas A. Schmitt, Psychology Department, Eastern Michigan University, 537K Mark Jefferson, Ypsilanti, MI 48197
Email: tschmit1@emich.edu

model and estimation method, (c) a method to determine the number of factors and evaluating model fit, and (d) a rotation criterion.

Sample Size

Although the FA literature has numerous recommendations concerning sample size rules of thumb, it is varied, ambiguous, and often lacks validity (MacCallum, Widaman, Zhang, & Hong, 1999; Marsh, Hau, Balla, & Grayson, 1998). As Marsh (2009) stated on SEMNET: "Golden rules or even guidelines about appropriate sample size are very tricky." This is often because such guidelines rely on studies that are limited in generalizability by the investigated conditions, such as the models considered, the estimator employed, and so on. Thus, researchers must remain aware of the limitations of such rules and ensure that a reasonable level of statistical precision and power for model parameter estimates can be obtained from the sample data. This can seem like an arduous task because precision and power are dependent on not just sample size but also on other factors such as the size of the hypothesized model (e.g., indicators per factor), the distribution of the variables (e.g., degree of multivariate normality), the estimation method (e.g., maximum likelihood estimation), the strength of association between variables (e.g., items and factors), the reliability of variables, and the amount and pattern of missing data and how it is dealt with or what missing data method is used (e.g., multiple imputation). Because of the numerous recommendations in FA literature, researchers are often confused. Fortunately, viable methods do exist to help determine appropriate sample size(s).

It is important to emphasize that appropriate sample size relies on the *precision* and *power* of the models parameter estimates. Precision is a test of how consistent or well the parameters and their standard errors are estimated. Power is defined as 1 minus the probability of committing a Type 2 error, with .80 most commonly defined as adequate power (Cohen, 1988). When considering the appropriateness of a model, it is important to have narrow confidence intervals around parameter estimates to ensure that the model's parameters are accurately estimated (see Kelley & Maxwell, 2003). Approaches to determine adequate sample size, commonly focus on statistical power because, generally, when adequate power is achieved, precision of parameter estimates will also be realized. But researchers often convolute the two, so it is important to realize that both precision and power must be considered.

One of the most well-known approaches for evaluating power of the likelihood ratio test in FA was developed by Satorra and Saris (1985; see also Brown, 2006; Kim, 2005). With the Satorra–Saris method, researchers compare a null model to an alternative model consisting of population or true values. The null model is the same as the alternative model except for the single parameter being tested. Note that the null model is nested in the alternative model. Unfortunately, this method is limited because (a) of the difficulty of defining an alternative model or an alternative parameter value to be tested, (b) of the difficulty in testing every parameter, (c) not all alternative models are testable, (d) researchers have to make exact estimates of the population values, and (e) it does not evaluate the precision of parameter estimates. Other authors have applied the Satorra–Saris method using bootstrapping (e.g., Yuan & Hayashi, 2003). The bootstrap method can be used with nonnormal and missing data, but it requires a large raw data set to determine power (see Brown, 2006).

Muthén and Muthén (2002) circumvented the raw data requirement and the lack of parameter precision estimates by using a Monte Carlo approach to simulate raw data from known parameters at various sample sizes. Like the Satorra–Saris method, the Muthén–Muthén approach requires parameter population values. The Muthén–Muthén approach can be used with different types of models, data, and estimation methods enabling researchers to specify a wide range of models that will reflect the particular types of variables (e.g., continuous, categorical, etc.) and distributions (normal, nonnormal, etc.) encountered in their work. Another important advantage of the Muthén–Muthén approach is that it randomly generates multiple samples from the population values, thus enabling researchers to evaluate the precision of the parameter estimates and their standard errors and, consequently, the confidence intervals. It is important to check precision because if parameter and standard error estimates are inaccurate at a sample size, the power estimates will be irrelevant.

Another method introduced by MacCallum, Browne, and Sugawara (1996) calculates power based on the root mean square error of approximation (RMSEA). The difficulty in applying the MacCallum–Browne–Sugawara method stems from the current debate about the strict use of cutoffs with approximate fit indexes (AFIs) and whether or not fit indexes are even appropriate for evaluating models (e.g., Marsh, Hau, & Wen, 2004; Vernon & Eysenck, 2007). In general, there exists little empirical support for the use of universal cutoff values for RMSEA to determine adequate model fit because to achieve a specific level of power the cutoff value of RMSEA depends on the specification of the model, the degrees of freedom, and the sample size (Chen, Curran, Bollen, Kirby, & Paxton, 2008). Unfortunately, the MacCallum–Browne–Sugawara method does not take all of these into account when calculating power based on RMSEA.

How then should researchers proceed when deciding on sample size? It is clear that determining an appropriate sample size based on rules of thumb is insufficient. There have been numerous rules-of-thumb recommendations of appropriate sample sizes based on sample size relative to the number of parameters being estimated (e.g., Jackson, 2007), the number of variables per factor (e.g., Marsh et al., 1998), and so on, but the limitation of these recommendations is the model(s) evaluated and the conditions studied. In other words, sample size is very much dependent on many factors that are inconsistent across models. Even though they are limited by requisite empirical data or prior knowledge, the Satorra–Saris method using bootstrapping and/or the Muthén–Muthén method using Monte Carlo simulation are reasonable approaches for determining sample size. Barrett (2007) and McIntosh (2007) provide a logical view that the Muthén–Muthén Monte Carlo approach is the best method for evaluating power because it enables researchers to integrate a wide variety of commonly encountered conditions into their model and evaluate precision of the parameter and standard error estimates.

It is important to note that regardless of what method is chosen for power determination and precision evaluation, researchers must be aware of the occurrence of *isopower*. Isopower is the phenomenon that different models, along with changes in other factors, can result in the same amount of power when testing a given null hypothesis (MacCallum, Lee, & Browne, 2010). As MacCallum et al. state, it is important for researchers to recognize this and consider altering the conditions and examine how this affects power or hold power constant and examine alternative sets of conditions that yield the same power. At minimum, researchers should state that their results are not isomorphic and that in all likelihood an infinite number of conditions exist that will yield the same power results.

Factor Models and Estimation Methods

A second consideration for researchers conducting FA is deciding on a model and the procedure to estimate the model parameters. The two main factor models associated with EFA or the unrestricted factor model include the *component model* and the *common factor model*, and numerous estimation or factor extraction methods exist for these models (see Gorsuch, 1983; Kaplan, 2009; Widaman, 2007). The main difference between these two models is that the component model assumes no measurement error and the common factor model attempts to account for measurement error. Principal component analysis (PCA) is one of the more frequently used component model–based factor extraction methods for EFA. Despite evidence that PCA can produce similar results to true factor analysis when measurement reliability is high and/or the number of factored variables/items increases (Gorsuch, 1983; Guadagnoli & Velicer, 1988, Thompson, 2004), PCA assumes measurement without error and is, therefore, less likely to generalize to CFA than EFA estimation methods of the common factor model.[1] In addition, PCA and EFA have different goals resulting in different outcomes, and PCA can produce inflated values of variance accounted for by the components (Gorsuch, 1997; McArdle, 1990; Widaman, 2007). All of this casts doubt on the use of PCA for depicting psychological and educational data. A widespread method used to estimate the common factor model is iterative principal axis factoring (PAF). PAF does afford the advantage of operating under the common factor model, thus taking into account measurement error. But along with PCA, PAF does not require data distributional assumptions and is, therefore, a *nonstatistical* estimation method (Kaplan, 2009). Thus, neither PCA nor PAF provide standard errors that would enable researchers to statistically test model fit and model parameters, such as factor loadings.

One of the most commonly used *statistical* estimation methods for estimating parameters of the common factor model is the maximum likelihood (ML) procedure. Because ML estimation acknowledges that sample data are being analyzed (i.e., makes distributional assumptions), researchers can statistically evaluate the hypothesis that there are a certain number of factors that predict the relationships among interfactor correlations, indicators/items, and factor loadings. Though ML is a common estimation method for CFA models, it is less commonly used for estimating EFA models. With the advent of ML-based EFA methods in well-known structural equation (SEM) modeling packages (e.g., M*plus*; Muthén & Muthén, 1998-2010) and papers promoting the use of EFA as an appropriate alternative to post hoc model modification within CFA (Asparouhov & Muthén, 2009; Browne, 2001; Sass & Schmitt, 2010; Schmitt & Sass, in press), ML-based EFA is likely to become more common in the psychological and educational sciences. With a sufficient sample size, proper model specification, and multivariate normality, ML will provide accurate standard errors, which can be used to test overall model fit, along with hypothesis tests of the interfactor correlations, factor loadings, and other model parameters.

The assumption of multivariate normality is especially important because response scales for measurement instruments are often not normally distributed when the ML parameter estimates are based on correlation or covariance matrices with ordinal variables. In other words, ordinal data, though purported to come from continuous normally distributed constructs, can result in categories (e.g., 1 = *strongly agree*, 2 = *agree*, 3 = *disagree*, 4 = *strongly disagree*) that are not continuous and are not normally distributed. Thus, the empirical attributes of the data do not match the assumptions of the estimation method. This can result in biased parameters and standard error ML estimates (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Lei, 2009; Rhemtulla, Brosseau-Liard & Savalei, 2010).

Three viable alternatives for estimating FA models for ordinal data are robust continuous ML estimation, robust least squares (LS) estimation, and robust weighted least squares (WLS) estimation (Beauducel & Herzberg, 2006; Bentler & Yuan, 1999; Flora & Curran, 2004; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Holgado -Tello, Chacón -Moscoso, Barbero-García, & Vila -Abad, 2010; Lei, 2009; Rhemtulla, Brosseau-Liard & Savalei, 2010).[2] Robust LS and WLS use polychoric correlations and robust ML uses standard Pearson correlations and all three rely on adjustments to the chi-square test statistic (χ^2), all of which can result in accurate parameter estimates and test statistics depending data and model conditions (see Rhemtulla, Brosseau-Liard & Savalei, 2010).

Like ML estimation, robust LS and WLS are a statistical estimation procedure; thus, standard errors are available for testing the overall factor structure and individual parameters (e.g., factor loadings). With robust LS and WLS, the differences in chi-square (χ^2) values do not follow a χ^2 distribution; therefore, a two-step procedure using the DIFFTEST option in M*plus* is necessary to obtain the correct χ^2 difference test when comparing nested models (Muthén & Muthén, 1998-2010).[3] Robust WLS is available in Mplus as weighted least square mean-and-variance adjusted (WLSMV) χ^2 test statistic estimation, robust LS as ULSMV, and robust ML as MLR.[4]

Two additional FA approaches that are important to mention because they are becoming more common in FA are formative measurement (e.g., Bollen & Davis, 2009) and Bayesian estimation (e.g., Muthén, 2010; Muthén & Asparouhov, 2010). In traditional FA items are *reflective*, which means the items depend on the latent construct or are reflections of the construct. Thus, variation in the construct results in variation in the indicators or items. From a *formative* perspective, the latent construct is dependent on the items, which means the latent construct is formed by the items. When the construct is formative, these models are often called causal indicator models because the items are "causing" the latent construct. Although there may be practical and statistical advantages to specifying indicators as formative, whether to use a formative of reflective specification should rest on the hypothesized theoretical relationship between the indicators and the latent construct(s). Roberts and Thatcher (2009), and Brown (2006) provide overviews for fitting models with formative indicators.

Bayesian estimation for FA models incorporates prior information to provide more accurate parameter estimates, and unlike ML estimation, Bayesian estimation does not depend on normally distributed large samples. Thus, distributions can be nonnormal, and performance is better for small samples when using Bayesian estimation methods. Another benefit of Bayesian estimation is that many models that are computationally difficult or impossible with ML estimation (e.g., models with categorical outcomes) can be estimated with Bayesian estimation. Lastly, Bayesian estimation allows for a wider range of models that can be analyzed. Though beyond the scope of this article, it is important that researchers begin to become familiar with Bayesian estimation because it is widely accepted and used in statistics, and is available in popular SEM software (e.g., M*plus*).

In summary, when data are continuous, PAF and ML estimation are viable options for estimating the common factor model. MLE has the advantage of producing test statistics for hypothesis testing, whereas PAF is feasible when sample participants are fewer than 50 (de Winter, Dodou, & Wieringa, 2009). Although PCA is popular and has been frequently employed by researchers, researchers should use it cautiously and realize its limitations for conducting FA with psychological and educational data. When data are ordinal with two to five categories researchers should consider robust LS, robust WLS or Bayesian estimation as each may produce more accurate parameter estimates when compared to robust ML continuous estimation. However, robust ML should also be considered for fewer categories when structural parameters (e.g., interfactor correlations) are of primary interest, when evaluating model fit, and when the underlying distribution is non-normal and/or when the thresholds are asymmetrical (e.g., when most of the response fall into one category). Researchers are also strongly encouraged to consider Bayesian analysis as it has been shown to outperform robust methods such as WLSMV (Asparouhov & Muthén, 2010a). Although other models and estimators are available for general latent variable analysis (see Asparouhov & Muthén, 2009; Bollen, Kirby, Curran, Paxton, & Chen, 2007; Esposito Vinzi, Chin, Hensler, & Wang, 2010), a discussion of these is beyond the scope of this article. The estimators discussed provide researchers with a wide range of options for analyzing different types of data in the context of FA.

Selecting the Number of Factors and Model Fit Criteria

Selecting the number of factors. Selecting the number of factors is an important part of construct validation in FA, which is commonly done in the context of EFA. Because over- or under-factoring (e.g., selecting too few/many factors) can result in significant modeling error, with underfactoring generally considered to be more detrimental of the two, appropriate methods must be used when determining the number of factors. As is true of most any statistical method, there are a multitude of methods for selecting the appropriate number of factors (see Fabrigar et al., 1999; Hayton, Allen, & Scarpello, 2004; Zwick & Velicer, 1986 for an overview). Some of the more well-known methods include the eigenvalue-greater-than-1 rule or Kaiser criterion (K1), the screen test, which is a visual plot of the eigenvalues, the minimum average partial (MAP) method, the χ^2 −based tests or the likelihood ratio test (LRT), and parallel analysis (PA). Other more recent methods that have been proposed include root mean square error adjustment (Browne & Cudeck, 1992), bootstrap methods (Lambert, Wildt, & Durand, 1990), and TETRAD (Glymour, 1982; Scheines, Spirtes, Glymour, Meek, & Richardson, 1998; Yu, Popp, DiGangi, & Jannasch-Pennell, 2007). Of these methods, PA and MAP has proven to be the most accurate and K1 the most inaccurate. Despite this, PA remains underutilized; thus, it will be the focus of this article (Hayton et al., 2004).

Parallel analysis uses a series of randomly generated data sets that "parallel" factors of the original data set in terms of sample size and number of variables (Horn, 1965). The rationale being that if real nonrandom factors exist then eigenvalues generated from the real data will be larger than the randomly generated eigenvalues. In general, simulation research has indicated that PA is the best empirical method for determining the number of factors in FA and PCA (Dinno, 2009) and has been recommended as the method of choice by journal editors (Thompson & Daniel, 1996) and others (e.g., Hayton et al., 2004; Henson & Roberts, 2006).

When modeling ordered categorical (e.g., Likert-type scales) item responses with FA, the distributional form may be nonnormal. It has been stated that PA is inappropriate for nonnormal distributions and modified approaches have been suggested (Hayton et al., 2004; Horn, 1965; Liu & Rijmen, 2008; O'Connor, 2000). But because PA randomly generates eigenvalues over multiple iterations, the central limit theorem should make the distribution of the data negligible. In fact, simulation work has shown that PA does not vary by the distributional assumptions made about the data (Dinno, 2009; Glorfeld, 1995). Another important point to note is FA based on the common factor model (e.g., EFA) must be modified. Fortunately, Dinno (2010) has developed a program called *paran* that is available in R (R Development Core Team, 2010; Dinno, 2001-2009) for doing PA for either EFA or PCA. If there is a large amount of missing data and/or data are not missing completely at random then researchers should handling missing data prior to conducting PA or consider using a PA method that allows for missing data, such as the approach outline by Liu and Rijmen (2008).[5]

Model-fit criteria. Despite evidence that indicates the LRT applied to EFA can results in too many factors, LRT is the only method grounded in distributional statistical theory (Hayashi, Bentler, & Yuan, 2007). Because of this and the fact that multiple criteria, along with sound theoretical reasoning, should be used when making factor retention decisions (Fabrigar et al., 1999; Henson & Roberts, 2006; Thompson & Daniel, 1996), LRT is a reasonably viable method that can help researchers select the appropriate number of factors (see Hayashi et al., 2007). Because of the LRT's tendency of overfactoring, researchers should use the LRT in conjunction with the standardized root mean square residual (SRMR) to evaluate improvement of fit due to each additional factor (Asparouhov & Muthén, 2009). Thus, it is recommended that the number of factors be determined with PA, and then evaluated using the LRT and the SRMR for improvement of model fit. This should work to prevent incorrect decisions being made with the LRT and still allow researchers to evaluate the hypothesized number of factors and factor structure within sound statistical theory. Although the LRT is used to compare models for factor selection, it should not be used to compare models when the base model is misspecified because it can result in inflated Type 1 and Type 2 errors (Yuan & Bentler, 2004).

Along with the χ^2 tests of model fit, AFIs that are sample independent are often used and/or required in applied FA research to evaluate and compare models with different numbers of factors (see Marsh, Hau, & Grayson, 2005, for an overview of AFIs). It is important to note that there is currently a great deal of debate concerning the validity of approximate fit statistics (e.g., Marsh et al., 2004; Vernon & Eysenck, 2007), so they should only be used as supplementary indicators to the χ^2 test of model fit. As Marsh et al. (2004) accurately summarized, rules-of-thumb cutoffs for AFIs should not be viewed as "golden rules," but as "preliminary interpretations that must be pursued in relation to the specific details of their research" (p. 321).

Bentler (2007) recommends limiting the reporting of fit indices to the SRMR or the average absolute standardized residual and at most two additional fit indices. Thus, the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the SRMR were used for the illustration below. It is recommended that the RMSEA be no greater than 0.06 (Hu & Bentler, 1999). Hu and Bentler stated that CFI values exceeding 0.95 indicate adequate model fit for continuous outcomes, and Yu (2002) found that a value of 0.96 was acceptable for ordered categorical outcomes. For the SRMR, values less than 0.08 indicate acceptable fit (Hu & Bentler, 1999). Nevertheless, it is important to remember these model fit statistics are simply guidelines and should not be interpreted as golden rules (Marsh et al., 2004).[6]

Another problem with assessing model fit is the well-known criticism of the χ^2 test for having severe dependence on sample size. This means that any model misspecifications will be detected by the χ^2 test with a large enough sample size, which has led to numerous AFIs (Hu & Bentler, 1998; Miles & Shevlin, 2007; Saris, Satorra, & van der Veld, 2009). Furthermore, as Saris et al. (2009) have shown, the χ^2 test is not only affected by sample size and the size of the misspecification but also by other model characteristics. Thus, global fit indexes, including χ^2 test and AFIs, should only be used for preliminary and exploratory interpretations as a researcher moves from a close fitting model to a near exact fitting model. It is then important to carefully evaluate a hypothesized model for specific sources of misfit (e.g., factor loadings) when using global fit indexes, such as the χ^2, because models with irrelevant misspecifications might be rejected if the sample is large, and models with substantial misspecifications might not be rejected if the sample is small (McDonald & Marsh, 1990; McIntosh, 2007; Saris et al., 2009). Because various factors influence both the χ^2 and AFIs sensitivity to detect model misspecification, it is important to move beyond global model evaluation methods as the only way to evaluate models.

Saris et al. (2009) have done just that in proposing a method for evaluating fit by examining individual parameters as sources of model misspecification. Because the χ^2 test and the AFIs cannot easily be used to test individual model parameters, the Saris–Satorra–van der Veld method should be used as complement to overall model testing. The Saris–Satorra–van der Veld approach examines specific misspecifications in the model by determining the power of the modification index (MI) test, which takes into account factors other than sample size. Saris et al. showed that the power of the MI test to detect a misspecification in a particular parameter (e.g., factor loadings) can be used in conjunction with the MI and the expected parameter change (EPC) for that parameter. The MI and EPC are both produced by standard SEM programs, such as M*plus*, for most relevant model parameters. The Saris–Satorra–van der Veld approach combines the significance or nonsignificance of the MI test and the high or low power of the MI test to distinguish four possible situations.

In the first situation, the MI is significant and the power of the MI test is low. Thus, because the test is not overly sensitive (i.e., low power) and the MI significant, it is clear that there is a misspecification in the model. In the second situation the reverse is true: The MI is not significant and the power of the MI is high. Thus, there is no misspecification of the parameter being evaluated. The third situation is a little more complex because the MI is significant, but the power of the test is high. This might be a misspecification, but it also might be that the MI is significant because it has high sensitivity. When this occurs, it is recommended that the EPC be evaluated. If the EPC is small it can be concluded there is no serious misspecification, but if the

EPC is large then the parameter is misspecified. The last situation occurs when the MI is not significant and the power of the MI test is low. There is then not enough information to make a decision. This highlights the fact that nonsignificant MIs do not necessarily indicate that misspecifications are not present; thus, it is important that researchers make model specification and modification decisions that have both substantive and statistical underpinnings. The Saris–Satorra–van der Veld method can be implemented using a program called Jrule or Judgment Rule Aid, which reads the MI and EPC values from either LISREL (Van der Veld, Saris, & Satorra, 2011) or M*plus* (Oberski, 2009).

Lastly, another underutilized component of EFA is statistical hypothesis testing of the factor pattern loadings (see Cudeck & O'Dell, 1994). Even in reviews of appropriate methods for conducting EFA, hypothesis testing of factor pattern loadings are rarely mentioned (e.g., Fabrigar et al., 1999; Henson & Roberts, 2006). Historically, hypothesis testing of factor loadings within EFA has been given little consideration due to computational complexity of the estimated standard errors (Jennrich, 2007). Fortunately, the standard errors have recently become more readily available for EFA in programs such as M*plus*. Hypothesis testing of standard errors are discussed and demonstrated in the Illustration, as well as in Schmitt and Sass (2011).

Rotation Criteria

Despite concerns that have been raised with EFA, it remains an accepted approach for researchers, partly because there has been a realization that CFA can be rather restrictive and even inappropriate when used in an exploratory fashion (Asparouhov & Muthén, 2009; Browne, 2001; Gorsuch, 1983). For example, CFA is often implemented assuming that each indicator perfectly depicts (i.e., all cross-loadings are zero) each factor. Because this unrealistic assumption often results in ill-fitting CFA models, researchers will turn to MIs for guidance in modifying their a priori hypothesized model. Unfortunately, this can result in post hoc model modifications that are not based on theory, which can lead to a model fitting by chance (MacCallum, Roznowski, & Necowitz, 1992). With the advent of advanced EFA methods, fitting CFA models by chance can be avoided.

One of the more important rationales for the continued use of EFA has been the introduction and application of exploratory structural equation modeling (ESEM; Asparouhov, & Muthén, 2009; Marsh et al., 2009; Marsh, Liem, Martin, Morin, & Nagengast, 2011). As discussed, a long-standing problem in applying FA is that CFAs are unable to replicate the factor structures produced by EFAs. This is commonly a result of the limiting nature of CFA because indicators only load on single factors and all cross-loadings are constrained to zero. This can distort the true factor structure and result in spuriously large interfactor correlations (Marsh et al., 2009; Schmitt & Sass, 2011). Because ESEM provides a viable method to circumvent some of the shortcomings of CFA and uses an EFA approach, being aware of and appropriately using different rotation criteria is as critical as ever.

In EFA, rotating factors is essential because even though clusters of variables may be obvious in the correlation matrix without factor rotation they are unlikely to be identified by the initial factor extraction methods (Gorsuch, 1983). Because researchers often choose rotation criteria based on the presence or absence of interfactor correlations (e.g., oblique or orthogonal), it is important that they become more acquainted with the different rotation methods. Unfortunately, it is common for researchers to give little rational for choice of rotation method (Henson & Roberts, 2006; for example, Promax, Quartimin, Equamax, etc.) or to consider how the selected rotation criterion may influence factor structure interpretation (Sass & Schmitt, 2010; Schmitt & Sass, in press). Instead, a rotation method is often arbitrarily based on how frequently it appears in the literature, which is generally the orthogonal Varimax criterion (Fabrigar et al., 1999; Ford, MacCallum, & Tait, 1986; Henson & Roberts, 2006; Russell, 2002). But most psychological and educational factors are correlated, so assuming factors are uncorrelated and using the Varimax criterion produces unrealistic factor structures. When factors are not allowed to correlate, item loadings will become inflated if the factors are truly correlated. Because oblique rotation methods generally produce accurate and comparable factor structures to orthogonal methods even when interfactor correlations are negligible, it is strongly recommend that researchers only use oblique rotation methods because they generally result in more realistic and more statistically sound factor structures.

Thus, researchers must be aware of the potential factor structure that may result from their data (see Schmitt & Sass, 2011). For example, more complex factor structures may result when researchers are developing and testing a new measure. With such measures, items may relate strongly to multiple factors (e.g., more and larger cross-loadings), so researchers should consider employing a rotation method that will allow for larger cross-loadings, such as CF-Equmax or CF-Facparsim. But if a measurement instrument has been well developed (e.g., fewer and smaller cross-loadings), researchers should consider Geomin or CF-Quartimax, because they are likely to produce a cleaner factor structures that are similar to CFA. Researchers may also want to consider multiple rotation criteria in an effort to better delineate the factor structure.

It is important to realize that the FA review in this article is not exhaustive; thus, researchers are encouraged to consult the references for further details. Researchers should also realize that each decision they make concerning how to conduct FA will have important implications for the validity of factor structures or lack thereof. Remaining cognizant of this, a demonstration will now be used to illustrate some of the previously discussed FA methods.

Illustration

A real data set was analyzed to illustrate some of the FA topics. The data consists of 26 psychological tests administered by Holzinger and Swineford (1939) to 145 students and has been used by numerous authors to demonstrate the effectiveness of FA. Of the 26 tests, 8 are used here and hypothesized to be formed by 2 constructs: a visual construct consisting of visual perception, cubes, paper form board, and flags, and verbal construct consisting of general information, paragraph comprehension, sentence completion, and word classification.

The first step when conducting FA is to determine the appropriate number of factors. To do this, a PA was performed using the R *paran* package (R Development Core Team, 2006; Dinno, 2009). Paran produces adjusted and unadjusted eigenvalues. The adjusted eigenvalues are corrected for sampling error that may result from finite or small samples. Because the sample for the Holzinger–Swineford data are not overly large and the bias statistics from paran (i.e., the difference between the adjusted and unadjusted eigenvalues for each factor) where rather large, adjusted eigenvalues were used to determine the appropriate number of factors. The PA results indicated a two-factor solution with adjusted eigenvalues of 2.76 and 0.36 using a cutoff of eigenvalues greater than zero for factor retention, which supports the hypothesized verbal and visual constructs. Note that most authors propose comparing the eigenvalues from the real and random data sets (e.g., Hayton et al., 2004), but an adjusted eigenvalue cutoff of greater than 1 for PCA and adjusted eigenvalue cutoff of greater than 0 for FA can be used because they are mathematically equivalent (Dinno, 2010; Horn, 1965).[7] Note that PA is used as a preliminary step of determining the number of factors, whereas model fit and evaluation statistics, such as the LRT, are used below to help "confirm" the factor structure.

The next step is to evaluate the sample size for both the precision and power of parameter estimates.[8,9] The Muthén–Muthén Monte Carlo approach using M*plus* and outlined by Muthén and Muthén (2002) was used (Brown, 2006, p. 424; Muthén & Muthén, 1998-2010, p. 375).

Population values for the factor loadings were set to 0.60, factor cross-loadings to 0.10, residual variance to 0.56, and interfactor correlation to 0.60.[10] Because the factor structure was hypothesized to have small cross-loadings, the Geomin rotation criterion was used. Geomin rotation is modified automatically in Mplus because the Epsilon parameter changes as a function of the number of parameters, which can be overridden by the user. For the current example epsilon was set to 0.05 because it resulted in better parameter and standard error estimates. Muthén and Muthén (2002) recommend three criteria for determining a sample size to achieving an adequate level of precision for a particular model: (1) bias for parameters and their standard errors should not exceed 10% for any parameter. (2) parameters that are being assessed in terms of power (e.g., factor loadings) should not have standard error bias greater than 5%, and (3) coverage should be between 0.91 and 0.98. If these minimum precision criteria are met, then an adequate sample size is achieved when power is 0.80 or greater. Note that Muthén and Muthén (2002) recommend 10,000 replications to achieve adequate values for evaluating precision.

In general, parameter bias for factor loadings, interfactor correlation, and residual variances were less than 10%, which met the first criteria. But the second criteria was not met because bias for the standard errors where larger than 5% for most parameters. Coverage was between 0.91 and 0.98 for all parameters except the factor cross-loadings. The power results for all parameters except the cross-loadings where above 0.80, but these results are not stable because of the significant bias exhibited in the standard error result of the Monte Carlo simulation. One obvious way to increase the precision of the standard errors would be to increase the sample size. But as de Winter et al. (2009) demonstrated, large factor loadings, fewer factors, and greater numbers of items can result in reliable solutions for small samples. Since this data set has a fixed sample size, but more items are available, it would seem prudent to reestimate power with more items. Although beyond the scope of the current demonstration, researchers are encouraged to run several Monte Carlo simulations varying different factors, such as the number of items, to optimize precision and power.

The Montel Carlo results can also be used to evaluate fit indices, such as the χ^2. Based on the Monte Carlo simulation, the χ^2 observed value of 0.05 for which the critical value was exceeded was the same as the expected value of 0.05, and the observed χ^2 value of 22.36 is close to the theoretical or expected value of 22.38. The bias for the χ^2 value is less than 0.1%, which indicates that the χ^2 distribution will be accurately approximated with a sample of 145 participants.

The third step is to fit the two-factor EFA model and evaluate model fit and parameter estimates. In order to demonstrate differences between EFA and CFA, a CFA model was also fit to the data. There are several important decisions that need to be made at this point, with the first being the choice of estimator. Since the Holzinger–Swineford data are comprised of continuous variables, ML estimation will be used. But if the data are ordinal or categorical, then it is recommended that researchers consider a robust WLS estimator, such as WLSMV or Bayesian estimation.[11,12] A nice characteristic of ML estimation in FA packages, such as M*plus*, Lisrel, AMOS, EQS, Mx, and so on, is that when the data contain missing responses all the available information is used to estimate the model. This results in consistent and efficient parameter estimates and test statistics, assuming data are missing completely at random (MCAR) or missing at random (MAR).[13,14] Conventional missing data methods (e.g., listwise deletion, simple or single imputation) will remove participants even if one response is missing or impute data using outdated methods, which can result in biased parameter estimates and standard errors (see Enders, 2010).

The next decision is the choice of rotation method.[15] Recall that there exist numerous rotation methods that can be divided up into those that reduce cross-loadings (e.g., simple structure) and those that allow for larger cross-loadings (e.g., complex structure; see Sass & Schmitt, 2010; Schmitt & Sass, in press). It should be noted that previous recommendations have focused on the obliqueness or orthogonality of the factor structures (e.g., Henson & Roberts, 2006), but since

Table 1. Rotated Factor Loading Pattern Results for Holzinger–Swineford Data

	Geomin		CF-Quartimax		CF-Equamax		CF-Facparsim		CFA	
Variable	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
Visual perception	**0.55**	0.11	**0.56**	0.10	**0.57**	0.09	**0.59**	0.06	**0.66**	0.00
Cubes	**0.56**	−0.05	**0.55**	−0.06	**0.55**	−0.06	**0.55**	−0.08	**0.50**	0.00
Paper form board	**0.43**	0.12	**0.44**	0.11	**0.45**	0.10	**0.47**	0.08	**0.53**	0.00
Flags	**0.76**	0.00	**0.76**	0.00	**0.76**	−0.01	**0.77**	−0.04	**0.72**	0.00
General information	0.14	**0.68**	0.18	**0.66**	**0.24**	**0.63**	**0.32**	**0.59**	0.00	**0.78**
Paragraph comprehension	0.01	**0.79**	0.06	**0.76**	0.13	**0.73**	**0.22**	**0.69**	0.00	**0.80**
Sentence completion	−0.17	**0.99**	−0.11	**0.96**	−0.02	**0.92**	0.09	**0.87**	0.00	**0.86**
Word classification	0.12	**0.64**	0.16	**0.62**	0.22	**0.59**	**0.29**	**0.55**	0.00	**0.72**
Interfactor correlation	0.55		0.51		0.44		0.37		0.59	

Note. Boldface numbers indicate statistically significant factor loadings.

most factors within the behavioral sciences are correlated (i.e., oblique) it seems fruitless to debate which is more theoretically appropriate. Moreover, oblique rotation methods will generally produce near-identical results to their orthogonal counterparts for uncorrelated factors; thus, this decision is moot and researchers should focus on hypothesized factor complexity when choosing an oblique rotation method. Because the Holzinger–Swineford data are thought to produce factor structures with few or mostly negligible cross-loadings, Geomin or CF-Quartimax would be appropriate choices for rotation methods. For illustration purposes Geomin, CF-Quartimax, CF-Equamax, and CF-Facparsim were used (see Table 1). Now that the estimator and rotation method have been decided upon, the model can be evaluated.

In order to evaluate the two-factor model the χ^2 test of model fit, the AFIs and the LRT were used to compare models. The one-factor EFA model provided evidence of ill fit, $\chi^2(20) = 69.07$, $p < .01$; RMSEA = 0.13, CFI = 0.88, SRMR = 0.08. For the two-factor solution the $\chi^2(13) = 10.70$, $p = .64$; RMSEA = 0.00, CFI = 1.00, SRMR = 0.02, and thus, was not statistically significant and provided evidence for a two-factor model. The three-factor solution also indicated adequate fit, $\chi^2(7) = 3.74$, $p = .81$; RMSEA = 0.00, CFI = 1.00, SRMR = 0.02; thus, it was used to compare the models using the LRT and the SRMR. The $\chi^2_{\text{diff}}(6) = 6.96$, $p = .32$, was not statistically significant and the ΔSRMR = 0.00 indicating the three-factor model did not fit significantly better than the two-factor model.

To further assess the appropriateness of the two-factor model the factor loadings were evaluated using their standard errors to determine whether or not they were statistically significant (Table 1). Using a correction procedure from Cudeck and O’Dell (1994) for correlated factors, the new α level can be computed as $\alpha^* = \alpha/d_u$, where α is the initial significance level (α =.05) and $d_u = im - m(m - 1)$, with i equal to the number of variables/items and m the number of factors. Bolded values are statistically significant at a two-tailed α^* =.004 ($z_{\alpha*} = 2.88$).[16] Expectedly, Geomin and CF-Quartimax produced statistically nonsignificant cross-loadings, whereas CF-Equamax and CF-Facparsim produced larger cross-loadings that in some cases where statistically significant. It can also be seen that as the factor loadings increased the interfactor correlations decreased. This is again a byproduct of the chosen rotation method where allowing for larger cross-loadings will generally result in a decrease in the interfactor correlations.

Because the cross-loadings were small indicating relatively simple factor structure, a CFA models was estimated and the Saris–Satorra–van der Veld procedure was also employed to test individual parameter misspecifications in the model. For the CFA model, its fit of the two-factor model, $\chi^2(19) = 22.40$, $p = .26$; RMSE = 0.04, CFI = 0.99, SRMR = 0.04, was similar to the EFA two-factor model. Notice that the CFA model had the highest interfactor correlations because the cross-loadings were fixed to zero. Of most interest in the current example were the individual factor loadings. To implement the Saris–Satorra–van der Veld procedure Jrule was used to examine the power and significance of potential cross-loadings by setting the misspecification cutoff, δ, to 0.40, Type 1 error rate to 0.05 and power to 0.80 (see Saris et al., 2009). The misspecification cutoff for examining the cross-loadings must be set by multiplying the standard deviation of the observed variables by 0.40 for all eight items. For example, the cutoff for the visual perception item would be $6.89 \times 0.40 = 2.76$, which is then entered in Jrule.[17] Results indicated that the CFA model was specified reasonably well as all eight of the cross-loadings were not misspecified by being set to zero. If Jrule had indicated misspecification(s) researchers could take several approaches including, but not limited to: evaluating statistical significance of possible cross-loadings, removing items, allowing the items to remain in the CFA, or modeling the cross-loadings in an ESEM model (see Marsh et al., 2011 and the CFA or EFA discussion below).

Like most hypothesis testing, researchers need to be aware that statistical significance does not automatically mean that a model does not fit and/or a factor loading is practically meaningful. In terms of the χ^2 test of model fit, it is good to have a large sample in order to find points of model misspecification, but because models are often complex and have hundreds, if not thousands, of degrees of freedom, there are many ways an FA model can be incorrect. Researchers also need to be aware that the detection of model misspecification increases with larger samples, and the same is true for the statistical tests of the factor loadings. A large sample can result in small cross-loadings being statistically significant, so researchers need also evaluate magnitude. This is not a criticism of statistical hypothesis testing; it is just meant to emphasize the need to integrate sound statistical tests with theory and practical significance.

CFA or EFA?

Within the context of this illustration, it is worth discussing some important differences between EFA and CFA. Researchers often erroneously assume that CFA is only used to verify or confirm hypothesized models, but researchers often apply CFA in an exploratory manner. Researchers often use MIs to modify CFA models, but when this occurs the perceived CFA becomes exploratory in nature (Bollen, 1989; Brown, 2001) and may be inappropriate (e.g., Asparouhov & Muthén, 2009; Byrne, 2001; Gerbing & Hamilton, 1996; Gorsuch, 1983; MacCallum et al., 1992; Marsh et al., 2011; Mulaik, 1972). Table 1 also illustrates when items relate only to a single factor, as in CFA, then both the factor loadings and interfactor correlations can become unrealistically inflated. In this sense, CFA very closely mirrors the results of the Geomin criterion within EFA.

On a final note, researchers should be aware that it is reasonable to follow up a poor-fitting CFA model with an EFA. Thus, researchers should consider a follow-up EFA when an adequate-fitting CFA model can only be obtained by model respecification based on the modification indices that are unsupported by theory, or when poor fit results and a large number of modification indices make CFA model respecification impossible. This does not mean that modification indices should never be used and/or CFA models should never be modified post hoc. It simply means that researchers should carefully consider all possibilities when a hypothesized model does not fit and realize that EFA is often more suitable for further "exploration" of poor fitting CFA models. In general, EFA can be used to (a) explore poorly fitting CFA models, (b) explore factor structures without strong hypotheses, and (c) confirm a factor structure based on strong hypotheses when the independent cluster assumption of CFA is unrealistic, such as ESEM (see

Marsh et al., 2011). Thus practically speaking, EFA and CFA are mostly differentiated by including or not including cross-loadings, respectively, and are not only "exploratory" or only "confirmatory" as CFA can be used to explore with MIs and EFA can be used to confirm when a priori cross-loadings are hypothesized.

Conclusion

The goal of this article was to outline and illustrate important and current methods in factor analysis. It is not meant to be exhaustive review of the literature or a complete step-by-step tutorial but is simply meant to provide a blueprint as to what researchers should consider and how they should proceed when conducting factor analysis. By no means is this the only way to proceed, but no matter how a researcher decides to conduct a factor analytic study and evaluate a model, it is important that they at least consider sample size, factor models and estimation methods, procedures for determining the number of factors and evaluating model fit, and rotation criteria. Researchers are encouraged to explore further the references provided when more depth is required and/or desired on a particular topic. And though relatively current, researchers conducting factor analysis need to stay abreast of the methods and realize that this article will soon become fodder for past-tense citations and more up-to-date methods. Hopefully, by carefully considering each of these important areas/methods, the science of factor analysis, and really the science of the interrelatedness of variables, will continue to move forward in an accurate and replicable fashion. Because, as is true of most sciences, factor analysis has come a long way, but it still has a long way to go.

Acknowledgment

The author is grateful to the following professionals for providing invaluable feedback on earlier versions of this article: Cameron McIntosh, Herb Marsh and his research team, Alexander Morin, Arief Liem, Andrew Martin, and Samuel Green.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Notes

1. PCA is often incorrectly assumed to be an EFA method of factor extraction, but it is technically not a member of the FA family of estimation methods that fall under the common factor model.
2. Robust WLS is a variant of full WLS which was developed for non-normal continuous data and is often called the asymptotically distribution free (ADF) estimator when all outcome variables are continuous (Brown, 1984; Flora & Curran, 2004).
3. The WLSMV estimator should not be used when comparing EFA nested models with Mplus, but WLS (full WLS) or WLSM (robust WLS with mean-adjusted χ^2 test statistic) can be used.
4. Similar procedures are available in Lisrel (Jöreskog & Sörbom, 2001) and EQS (Bentler, 2010).
5. The PA procedure developed by Liu and Rijmen (2008) was developed for use with ordinal data, but Dinno (2010) demonstrated that the proliferation of the data should not affect PA results.
6. The fit statistics and their respective cut-offs are appropriate for ML estimation methods. Researchers should check for appropriate fit statistics based on the estimator they use.
7. Random eigenvalues and a plot of eigenvalues is available from paran.

8. Because the sample size n = 145 is set for the current example, it was evaluated for power and accuracy of parameter estimates post hoc. In general, it would be more likely for researchers to evaluate different sample sizes a priori to determine an adequate sample size. This can be done in Mplus by specifying different sample sizes for the hypothesized model.
9. As discussed below, ML estimation and Geomin rotation were used for the Monte Carlo simulation.
10. To standardize the item variance scale to one, the residuals were calculated based on the factor loadings, cross-loadings, and interfactor correlation (see Gorsuch, 1983, pp. 29-30; Sass & Schmitt, p. 83).
11. WLSMV uses all available data with limited pair-wise information, which is a more restrictive variation of the full information ML estimation method, thus it has been show that Bayesian estimation outperforms WLSMV with categorical or ordinal data in the context (Asparouhov & Muthén; 2010a; Asparouhov & Muthén; 2010b).
12. At this time Bayesian estimation for EFA is not available in Mplus.
13. ML used when missing data is present is commonly called full information maximum likelihood (FIML), but be aware that ML is always a full information estimator.
14. ML estimation with missing data requires raw data input as opposed to a correlation or variance-covariance matrix that can be used when there is no missing data.
15. Note that the χ^2 test statistic along with approximate fit indexes will not change across rotation methods.
16. Note that other correction procedures are available.
17. The variance of the latent variables must be set to one or if the first item is used to set the scale then standard deviation of the observed variable must be divided by the standard deviation of the latent trait and then multiplied by 0.40.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438.

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus*. Technical Report. Version 4. Retrieved from http://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. (2010b). *Weighted least squares estimation with missing data* (Technical Report). Retrieved from http://www.statmodel.com/download/GstrucMissingRevision.pdf

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

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13*, 186-203.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*, 825-829.

Bentler, P. M. (2010). *EQS structural equation modeling software*. Encino, CA: Multivariate Software.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, 34, 181-197.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Bollen, K. A., & Davis, W. R. (2009). Causal indicator models: Identification, estimation, and testing. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 498-522.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: A comparison of Two Stage Least Squares (2SLS) to Full Information Maximum Likelihood estimators (FIML). *Sociological Methods and Research, 36*, 48-86.

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York, NY: Guilford.

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

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

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

Byrne, B. M. (2001). *Structural equation modeling with Lisrel, Prelis, and Simplis: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods and Research, 36*, 462-494.

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

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

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

Cudeck, R., & MacCallum, R. C. (Eds.). (2007). Factor analysis at 100: Historical developments and future directions. Mahwah, NJ: Lawrence Erlbaum.

Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin, 115*, 475-487.

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research, 44*, 147-181.

Dinno, A. (2001-2009). *Paran. Performs Horn's parallel analysis for principal component (or factor) retention. Packages written for use with STATA, and for R*. Retrieved from http://www.doyenne.com/stata/paran.html

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

Dinno, A. (2010). *Gently clarifying the application of Horn's parallel analysis to principal component analysis versus factor analysis*. Retrieved from http://doyenne.com/Software/files/PA_for_PCA_vs_FA.pdf

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Esposito Vinzi, V., Chin, W. W., Henseler, J., & Wang, H. (Eds.). (2010). *Handbook of partial least squares*. New York, NY: Springer.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272-299.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

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

Forero, C., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*, 625-641.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*, 62-72.

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*, 377-393.

Glymour, C. (1982). Casual inference and causal explanation. In R. McLaughlin (Ed.), *What? Where? When? Why? Essays on induction, space, and time,explanation* (pp. 179-191). Boston, MA: D. Reidel Publishing Company.

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

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

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

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

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

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

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality and Quantity, 44*, 153-166.

Holzinger, K., & Swineford, F. (1939). A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monograph, no. 48*. Chicago: University of Chicago Press.

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

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

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

Jackson, D. L. (2007). The effect of the number of observations per parameter in misspecified confirmatory factor analytic models. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 48-76.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: A review and some recommendations. *Psychological Methods, 14*, 6-23.

Jennrich, R. I. (2007). Rotation methods, algorithms, and standard errors. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 315-335). Mahwah, NJ: Lawrence Erlbaum.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8 user's reference guide*. Lincolnwood, IL: Scientific Software International.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: SAGE.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*, 305-321.

Kim, K. H. (2005). *The relation among fit indexes, power, and sample size in structural equation modeling, 12*, 368-390.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1990). Assessing sampling variation relative to number-of-factors criteria. *Educational and Psychological Measurement, 50*, 33.

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality & Quantity, 43*, 495-507.

Liu, O. L., & Rijmen, F. (2008). A modified procedure for parallel analysis for ordered categorical data. *Behavior Research Methods, 40*, 556-562.

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

MacCallum, R. C., Lee, T., & Browne, M. W. (2010). The issue of isopower in power analysis for tests of structural equation models. *Structural Equation Modeling, 17*, 23-41.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*, 490-504.

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

Marsh, H. W. (2009, May 27). Re: There are no golden rules about sample size. Retrieved from http://bama.ua.edu/archives/semnet.html

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Marsh, H. W., Muthén, B., Asparouhov, A., Lüdtke, O., Robitzsch, A., Morin, A. J. S., . . . Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439-476.

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275-340). Mahwah, NJ: Lawrence Erlbaum.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological-measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. Special issue in the Journal of Psychoeducational Assessment.

McArdle, J. J. (1990). Principles versus principals of structural factor-analyses. *Multivariate Behavioral Research, 25*, 81-87.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness-of-fit. *Psychological Bulletin, 107*, 247-255.

McIntosh, C. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*, 859-857.

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences, 42*, 869-874.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York, NY: McGraw-Hill.

Muthén, B. (2010). *Bayesian analysis in Mplus: A brief introduction*. (Technical Report). Version 3. Retrieved from http://www.statmodel.com/download/IntroBayesVersion%203.pdf

Muthén, B., & Asparouhov, T. (2010, September). *Bayesian SEM: A more flexible representation of substantive theory*. Submitted for publication. Retrieved from http://www.statmodel.com/download/BSEMv4.pdf

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599-620.

Oberski, D. L. (2009). Jrule for Mplus version 0.91 [computer software]. Retrieved from http://wiki.github.com/daob/JruleMplus/

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. B*ehavior Research Methods, Instruments, & Computers, 32*, 396-402.

Preacher, K. J., & Maccallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*, 13-43.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rhemtulla, M., Brosseau-Liard, P., & Savalei, V. (2010). How many categories is enough to treat data as continuous? A comparison of robust continuous and categorical SEM estimation methods under a range of non-ideal situations. Retrieved from http://www2.psych.ubc.ca/~mijke/files/HowManyCategories.pdf

Roberts, N., & Thatcher, J. (2009). Conceptualizing and testing formative constructs: Tutorial and annotated example. *ACM SIGMIS Database archive, 40*, 9-39.

Russell, D. (2002). In search of underlying dimensions: the use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and social psychology bulletin, 28*, 1629-1646

Saris, W. E., Satorra, A., & van der Veld, W. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561-582.

Sass, D. A., & Schmitt, T. A. (2010). A Comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 1-33.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83-89.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The TETRAD project: Constraint-based aids to causal model specification. *Multivariate Behavioral Research, 33*, 65-117.

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*, 95-113.

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

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement, 56*, 197-208.

Van der Veld, W., Saris, W. E., & Satorra, A. (2008). *Jrule 2.0: User manual* (Unpublished Manuscript, Internal Report). Radboud University Nijmegen, the Netherlands.

Vernon, T., & Eysenck, S. (2007). Introduction to special issue on Structural Equation Modeling. *Personality and Individual Differences, 42*, 813.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177-203). Mahwah, NJ: Lawrence Erlbaum.

Worthington, R. W., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*, 806-838.

Yu, C. H., Popp, S. O., DiGangi, S., & Jannasch-Pennell, A. (2007). Assessing unidimensionality: A comparison of Rasch Modeling, Parallel Analysis, and TETRAD. *Practical Assessment, Research, and Evaluation, 12*. Retrieved from http://pareonline.net/pdf/v12n14.pdf

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Doctoral dissertation, University of California, Los Angeles.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and z-tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757.

Yuan, K.-H., & Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *British Journal of Mathematical and Statistical Psychology, 56*, 93-110.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

Bio

Thomas A. Schmitt, is an Assistant Professor of Psychology at the Eastern Michigan University. His specialization is in applied statistics and measurement as related to education and psychology. His research interests include methodological issues related to latent variable modeling such as factor analysis, structural equation modeling, mixture modeling, latent class analysis, multilevel modeling, item response theory, adaptive testing, and test and instrument construction.