Journal of MultiDisciplinary Evaluation
Volume 11, Issue 25, 2015

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

How Good Are Our Measures? Investigating the Appropriate Use of Factor Analysis for Survey Instruments

Megan Sanders
The Ohio State University

P. Cristian Gugiu
The Ohio State University

Patricia Enciso
The Ohio State University

Background: Evaluation work frequently utilizes factor analysis to establish the dimensionality, reliability, and stability of surveys. However, survey data is typically ordinal, violating the assumptions of most statistical methods, and thus is often factor-analyzed inappropriately.

Purpose: This study illustrates the salient analytical decisions for factor-analyzing ordinal survey data appropriately and demonstrates the repercussions of inappropriate analyses.

Setting: The data used for this study are drawn from an evaluation of the efficacy of a drama-based approach to teaching Shakespeare in elementary and middle school.

Intervention: Not applicable.

Research Design: Survey research.

Data Collection and Analysis: Four factor analytic methods were compared: a traditional exploratory factor analysis (EFA), a full-information EFA, and two EFAs within the confirmatory factor analysis framework (E/CFA) conducted according to the Jöreskog method and the Gugiu method.

Findings: Methods appropriate for ordinal data produce better models, the E/CFAs outperform the EFAs, and the Gugiu method demonstrates greater model interpretability and stability than the Jöreskog method. These results suggest that the Gugiu E/CFA may be the preferable factor analytic method for use with ordinal data. Practical applications of these findings are discussed.

Keywords*: factor analysis; ordinal data; E/CFA; survey research.*

Within the field of evaluation, survey research is one of the most predominant methodologies. In a review of empirical work published in 2014 in several evaluation journals (*New Directions for Evaluation, Educational Evaluation and Policy Analysis, Evaluation and the Health Professions, Evaluation Review, Evaluation and Program Planning*, and *American Journal of Evaluation*), 45% (74 of 165) of the articles reviewed utilized survey data in some form. Of these 74 studies, 24% (18) reported using factor analysis to examine the dimensionality, reliability, or stability of the survey used in the study. However, even though surveys are a popular methodology for use in evaluation, survey data is frequently analyzed incorrectly (Kampen & Swyngedouw, 2000). Most survey data are ordinal, but the most commonly-used methods of factor analysis assume that data are continuous. The reviewed literature underscores this mismatch: of the 74 reviewed studies that collected survey data, only 4% (3) appear to have explored the quality of their measures using factor analysis in a way appropriate for the nature of the data that was collected. Because factor analysis is such a commonly-used method, representing 24% of the reviewed articles that utilized survey data, it is important to discuss methods of factor analysis that are appropriate for use with ordinal data. The purpose of this article is to illustrate the repercussions of utilizing inappropriate factor analytic methods and to demonstrate how to conduct appropriate factor analytic methods, using an illustrative, evaluation-based example. To make this discussion maximally beneficial for the reader, this study illustrates points that are well known in the psychometric literature using real data that was collected as part of an evaluation, rather than emphasizing the mathematics of how one method is superior to the other or using Monte Carlo simulations, which utilize simulated data that frequently do not match the complexities of real-world data (e.g., employ designs with few variables, simple correlation matrixes, and often no method effects).

The evaluation used as an example in this study aimed to assess the efficacy of a drama-based approach to teaching Shakespeare. The Royal Shakespeare Company created the Stand Up for Shakespeare (SUFS) program to change the way students encounter Shakespeare in school (Strand, 2009). The program prepares teachers to help students engage with Shakespeare the way actors would—interacting with the plays as scripts to be acted rather than texts to be read. Drama-based pedagogy programs like SUFS have been show to positively impact students' academic and social outcomes (Lee, Patall, Cawthon, & Steingut, 2015); through this kind of pedagogy, SUFS in particular aims to increase students' positive attitude towards Shakespeare and both their interest and ability in reading. As part of the evaluation of the SUFS program, surveys were developed to measure student attitudes toward Shakespeare, to determine whether SUFS increased students' positive attitudes toward Shakespeare, and to in turn examine whether these positive attitudes facilitated students' knowledge acquisition. This paper will focus on the validation of the instrumentation using factor analysis to illustrate the impact of using more and less appropriate methods, to highlight the salient analytical decisions that distinguish each method, and to provide guidance for the reader aiming to appropriately conduct factor analysis with ordinal data.

The evaluation used as an example in this study aimed to assess the efficacy of a drama-based approach to teaching Shakespeare. The Royal Shakespeare Company created the Stand Up for Shakespeare (SUFS) program to change the way students encounter Shakespeare in school (Strand, 2009). The program prepares teachers to help students engage with Shakespeare the way actors would—interacting with the plays as scripts to be acted rather than texts to be read. Drama-based pedagogy programs like SUFS have been show to positively impact students' academic and social outcomes (Lee, Patall, Cawthon, & Steingut, 2015); through this kind of pedagogy, SUFS in particular aims to increase students' positive attitude towards Shakespeare and both their interest and ability in reading. As part of the evaluation of the SUFS program, surveys were developed to measure student attitudes toward Shakespeare, to determine whether SUFS increased students' positive attitudes toward Shakespeare, and to in turn examine whether these positive attitudes facilitated students' knowledge acquisition. This paper will focus on the validation of the instrumentation using factor analysis to illustrate the impact of using more and less appropriate methods, to highlight the salient analytical decisions that distinguish each method, and to provide guidance for the reader aiming to appropriately conduct factor analysis with ordinal data.

More specifically, we compared four approaches to factor analysis: a traditional exploratory factor analysis (EFA), which is characteristic of how factor analysis is used in evaluation work, a full-information or ordinal EFA (Jöreskog & Moustaki, 2006), and two exploratory factor analyses within the confirmatory factor analysis framework (E/CFA): one according to the

Jöreskog model specification search method (1969; Jöreskog & Sörbom, 1979) and the other according to the Gugiu method (Gugiu, 2011; Gugiu, Coryn, Clark, & Kuehn, 2009). These methods differed in the observed input correlation matrix, the method of estimation used to extract factors, the method of factor selection, and the method of model modification used in refining the models. The appropriateness and strength of the four methods were assessed by determining how well the extracted models replicated in an independent data set. To best illustrate the salient analytical differences between these methods, and how these differences are more and less appropriate for the nature of the data, each will first be described in detail before comparing the models produced by each method.

Traditional EFA

Before discussing the traditional EFA method, it is worth drawing a distinction between factor analysis and components analysis (CA). The two methods are similar and often confused, but only factor analysis is appropriate for exploring latent factor structure such as the SUFS survey of attitudes toward Shakespeare. The difference lies in how the two methods model variance. CA models variance with components, which are a linear combination of *all* of the variance in the set of indicators used in the CA. Principal components analysis (PCA) is a special case of CA wherein the research does not retain all the components. Although PCA does not explain all the variability, its method for extracting components does not rely on the assumption of a latent factor structure. Hence, mathematically the model is identical to that of CA. Factor analysis, on the other hand, models variance using latent factors, a linear combination of only the *common* variance in the set of indicators used in the EFA (Tabachnick & Fidell, 2001). Thus, the first step is to determine whether one is interested in modeling all of the variance (CA and PCA) or just the common variance (EFA). In general, if a common trait or construct is thought to predict a set of behaviors, indicators, or responses to a set of items, then the appropriate method is EFA, not PCA.

The second step is to define the input correlation matrix that will be modeled with latent factors in the EFA. Input matrixes, unless otherwise specified, are always generated using the Pearson product-moment correlation coefficient, which assumes that the data are continuous. However, using the Pearson coefficient to calculate correlations for ordinal variables decreases variability, as all scores within a given range on the latent variable are assigned to the same category of the observed variable. This reduced variability leads to underestimated associations between variables (Gilley & Uhlig, 1993), as well as decreased parameter estimates in factor analyses using Pearson correlation matrixes as input (DiStefano, 2002; Olsson, 1979b). To illustrate the impact of inappropriately inputting a Pearson correlation matrix with non-normal and non-continuous data, the matrix was used in the traditional EFA with the SUFS data.

Factor selection—arguably the most important step of factor analysis—uses the correlation matrix produced in factor extraction to calculate the number of factors that should be retained in subsequent analyses. Scree plots and the Kaiser criterion are frequently used to select factors although the accuracy of these methods is questionable (Hayton, Allen, & Scarpello, 2004). This study employed parallel analysis to select the number of factors (Horn, 1965). This method plots the eigenvalues calculated from the actual data (identical to the scree plot) against the eigenvalues extracted from random data that matches key characteristics of the actual data, including the sample size and number of variables. Principal components analysis is typically used to extract eigenvalues from the random dataset, and the process of generating random data and extracting eigenvalues is repeated between 50 and 1,000 times, with greater repetitions leading to more accurate results (Hayton et al., 2004). The current study used 100 repetitions since in our experience the decision beyond this number is never altered. The 95th percentile of eigenvalues extracted from the random data is used to counter the tendency of parallel analysis to overfactor (Glorfeld, 1995). Finally, the actual eigenvalues and the 95th percentile eigenvalues from the random data are plotted in descending magnitude. The point at which the two plotted lines cross indicates the number of factors to retain. A factor is worth retaining when its associated eigenvalue is greater than the eigenvalue expected by chance alone (Gugiu, Coryn, Clark, & Kuehn, 2009; Hayton et al., 2004).

After the input correlation matrix is specified, the next step is to select a method for estimating the factor model. The most frequently used methods of estimation are principal axis factors (PF) and maximum likelihood (ML), which are appropriate for use when data are continuous and, in the case of ML, normal (Brown, 2006). Although PF does not carry a distributional assumption, it only produces a limited range of goodness-of-fit statistics and does not allow for

statistical significance testing, limitations not faced by ML (Fabrigar, Wegener, MacCallum, & Strahan, 1999). To allow for model-fit comparison across the four factor analytic methods, ML was used in the traditional EFA despite being inappropriate given the ordinal nature of the data.

Once estimated, the original model may be modified by eliminating items with factor loadings that fall below a threshold of 0.3 (Tabachnick & Fidell, 2001). Below this threshold, the latent factors account for less than 10% of the variance in the item. Therefore, these items are not strong indicators of the latent variable and can be eliminated. Finally, reliability of the final model is typically, though problematically, estimated by Cronbach's alpha. Similar to other standard procedures, Cronbach's alpha misestimates reliability unless specific conditions hold, such as tau equivalence and the absence of correlated measurement errors, and underestimates the true reliability when data are measured on an ordinal scale (Brown, 2006; Zumbo, Gadermann, & Zeisser, 2007).

Full-information EFA

Full-information or ordinal EFA (Jöreskog & Moustaki, 2006) differs from the traditional EFA in the coefficient used for the input correlation matrix and the method of model estimation. First, full-information EFA uses the polychoric correlation coefficient rather than the Pearson correlation coefficient. This coefficient is estimated from the bivariate frequency distribution (crosstab) of the observed ordinal scores, under the assumption of bivariate normality. The estimated relationships are closer to the Pearson correlations that would be found if the variables were measured on an interval rather than ordinal scale (Brown, 2006; Olsson, 1979a), Consequently, the coefficients are more accurate, yielding less attenuated parameter estimates in factor analysis.

As with the traditional EFA, parallel analysis was used in the full-information EFA to select factors. However, the full-information EFA uses a different factor extraction method; namely, diagonal weighted least squares (DWLS) in conjunction with the asymptotic covariance matrix. Unlike ML, DWLS and the asymptotic covariance matrix adjust parameter estimates for violations of normality and so are appropriate for use with non-normal and categorical data (Brown, 2006). The asymptotic covariance matrix is used to compute a weight matrix used to adjust the fit statistics and standard errors for nonnormality (Brown, 2006). Essentially, items with less asymptotic variance (i.e., greater precision) are given more weight than variables with more variance (i.e., more sampling error) (Schumacker & Lomax, 2010). After factor selection and model estimation, items with loadings less than 0.3 are eliminated (Tabachnick & Fidell, 2001), paralleling the modification process in the traditional EFA method.

The inter-item reliability of the full-information EFA model is reported in terms of Raykov's (2001, 2004) coefficient of scale reliability ρ, which avoids many of the problems associated with Cronbach's alpha. The coefficient represents the proportion of true score variance to total score variance and is calculated by the expression:

$$\rho = \frac{\left(\Sigma\lambda_i\right)^2}{\left(\Sigma\lambda_i\right)^2 + \Sigma\theta_{ii} + 2\Sigma\theta_{ij}}$$

where λ_i represents the unstandardized factor loading for the ith item, θ_{ii} the unstandardized measurement error variances for the ith item, and θ_{ij} any correlated measurement errors between items i and j.

E/CFA

As a method, exploratory factor analysis within the CFA framework falls between EFA and CFA, utilizing a CFA imposed with the same limitations as an EFA (Jöreskog, 1969). Specifically, factor variances are set to unity; the item with the highest loading on a factor is chosen as an anchor and its cross-loadings are fixed to zero; and factor covariances are freely estimated. Although the model fit of the E/CFA is the same as that of a parallel EFA, the E/CFA provides additional information, such as modification indices and the statistical significance of factor loadings, which can be used to refine the model before validating it in a CFA. This specification search tends to produce better fitting initial models that are more likely to replicate in an independent CFA (Brown, 2006; for applied examples, see Gugiu, Coryn, Clark, & Kuehn, 2009; Brown, White, Forsyth, & Barlow, 2005).

Similar to the full-information EFA, the Jöreskog and Gugiu E/CFA methods utilize the polychoric correlation matrix as input, DWLS and the asymptotic covariance matrix to extract factors, and parallel analysis to select factors. The two methods of E/CFA are similar in specification search and extraction method but differ in model

modification. The Jöreskog approach (1969; Jöreskog & Sörbom, 1979) relies on identifying large modification indices (MI) in an initial model. MIs represent the amount the model chi-square will decrease if the corresponding correlated error is freed. Freeing errors with MIs greater than 3.84, the critical chi-square value at $\alpha = 0.05$, will result in significantly better model fit, as indicated by a significant χ^2 difference test between the simpler (nested) and more complex (null) model. Because freeing a single significant MI can have substantial and unpredictable effects on model fit indices, significant MIs should be freed one at a time in an iterative process. Furthermore, the correlated errors corresponding to the MI should only be freed if substantially justified by theory. After the highest correlated error is freed, the modified model is compared to the previous model using a χ^2 difference test, and this process is repeated until both the χ^2 for the model and the χ^2 difference test are nonsignificant, indicating that the last freed error covariance did not significantly improve model fit.

An alternative method of model specification within the E/CFA framework is the Gugiu approach (Gugiu, 2011; Gugiu, Coryn, Clark, & Kuehn, 2009). Rather than freeing correlated errors, this approach deletes items from the model that contribute to model misfit. Candidates for deletion are selected by examining the residual table for the largest misfitting standardized residual. Standardized residuals represent the difference between the estimated and observed covariances divided by the asymptotic standard errors—the square root of the asymptotic variance. Thus, when a standardized residual is greater than 1.96 in absolute value (critical value at $\alpha = 0.05$), the two items that correspond to this covariance contribute a great deal to the model misfit (Schumacker & Lomax, 2010). To reduce model misfit using the Gugiu E/CFA method, the item with the greatest number of misfitting standardized residuals is removed from the model. Similar to the Jöreskog method, each reduced model is compared to the previous models, and the process is repeated until the χ^2 difference test is nonsignificant (Gugiu, 2011; Gugiu, Coryn, Clark, & Kuehn, 2009). Although the Jöreskog method relies primarily on MIs whereas the Gugiu method relies on standardized residuals, both approaches use additional information provide by the CFA framework to refine the initial models.

Model Comparison

In the current study, models of the SUFS pretest data were specified according to the four methods of factor analysis and then a confirmatory factor analysis (CFAs) was performed on the posttest data. SAS 9.3 was used to specify the traditional EFA model and the full-information EFA model; all other model specification and validation was conducted in LISREL 8.8. The EFAs specified in SAS were also run in LISREL to obtain model fit statistics, which are not normally produced in an EFA framework.

Models were compared based on four goodness-of-fit statistics, the number of misfitting standardized residuals and significant MIs, and reliability. The first goodness-of-fit statistic used was the χ^2 statistic for the model, which indicates whether the difference between the observed and estimated model is significant. A nonsignificant χ^2 statistic suggests that the model fits the data well. The normal theory weighted least squares χ^2 (NTWLS χ^2) is appropriate when data are normal, whereas the Satorra-Bentler scaled χ^2 (SB χ^2; Satorra & Bentler, 1994) is appropriate when data do not meet the normality assumption. The models were also compared in terms of the root mean square error of approximation (RMSEA), which measures absolute fit with a penalty for non-parsimonious models (Brown, 2006). A RMSEA value of 0 indicates perfect fit, while values less than or equal to 0.05 indicate that the model fits the data well (Browne & Cudeck, 1993; Steiger & Lind, 1980). The standardized root mean square residual (SRMR), a measure of absolute fit, was also used to compare the models. SRMR values below 0.05 indicate good model fit, with smaller values indicating better fit (Schumacker & Lomax, 2010). The last goodness-of-fit statistic used was the Tucker-Lewis index (TLI; Tucker & Lewis, 1973). Like RMSEA, TLI also includes a penalty for model complexity but measures model fit relative to the null model. The closer the TLI value is to 1.0 the better the model; TLI values above 0.95 are desirable (Hu & Bentler, 1999).

Models were also compared in terms of the number of observed and expected misfitting standardized residuals and significant MIs. The number of misfitting residuals or significant MIs expected by chance may be calculated by $p * k(k-1)/2$, where k denotes the number of variables, $k(k-1)/2$ is the total number of residuals or MIs, and p is the Type I error rate (Gugiu, Coryn, Clark, & Kuehn, 2009). A larger-than-expected number of observed misfitting

standardized residuals or significant MIs suggests that the model does not capture important relationships in the data. Models were compared in terms of whether and by how much they exceeded the number of misfits expected by chance. Finally, the reliabilities of the models were also compared to assess the stability of the latent scores produced by each model. Comparing the models produced through each of the four methods will highlight the impact of these different analytical decisions.

Method

Sample

Schools and teachers in a large, midwestern city were recruited to participate in the SUFS drama-based pedagogy program: 503 students participated in the study (54% female; 5% Asian, 15% Black or African American, 4% Hispanic, 18% multi-racial, 49% white, 10% not reported). These students ranged from grades 3 to 8 and were drawn from the classrooms of 14 teachers across 5 schools in 2 public school districts.

Instrument

The primary purpose of the evaluation was to determine whether the drama-based pedagogy intervention improved students' positive attitudes towards Shakespeare. Thus, students were administered a survey about their exposure to Shakespeare, attitudes toward Shakespeare, and attitude toward school in September 2011 and May 2012. This was modeled on the survey used during the original implementation of the SUFS program in England (Strand, 2009). Questions pertaining to Shakespeare were measured on a 3-point scale: "No" (1), "Don't Know" (2), and "Yes" (3). Although it is unclear where "Don't Know" should logically fall on the dimension of "No" to "Yes," it appeared as the middle anchor point on the survey (see Appendix), and a Rasch analysis confirmed that students indeed treated "Don't Know" as a middle point (Yeomans-Maldonado, Gugiu, & Enciso, 2013). Thus, instead of dichotomizing the scale by collapsing the "Don't Know" responses into the "No" category, the 3-point ordinal scale was retained. This study focused on the 12 questions about student attitudes toward Shakespeare.

Procedure

Teachers collected permission from parents for students to participate in the study. A survey was administered before students had been exposed to the SUFS pedagogy and again after teachers had implemented SUFS pedagogy. Research assistants read the questions out loud to students at the end of a regular class period, and students who did not have permission to participate were asked to sit quietly.

Results

Factor Selection

First, missing values were imputed so as to prevent the sampling bias that would have resulted from listwise deletion. Then, parallel analysis was used to select factors from the Pearson correlation matrix, in the case of the traditional EFA, and from the polychoric correlation matrix, in the case of the full-information EFA, the Jöreskog E/CFA, and the Gugiu E/CFA. The resulting plot indicated that in both the case of the Pearson and the polychoric correlation matrix, one factor should be retained.

Model Specification Search

Traditional EFA. All 12 items in the traditional EFA loaded on the latent factor above the 0.3 threshold and so were retained in the model. The goodness of fit statistics indicated that the final model did not fit the observed data particularly well (see Table 1). Furthermore, the observed number of misfitting residuals (20) and significant MIs (20) in the EFA model exceeded the number expected (3.3), suggesting considerable model misfit. The model reliability was acceptable as measured by Cronbach's alpha, $\alpha = 0.845$ and acceptable according to Raykov's $\rho = 0.847$ (calculated so the reliability of all four models could be compared).

Table 1

Goodness-of-Fit Indices for Models of Attitude Toward Shakespeare (n=400)

Model Building: Pretest Sample

Final Model	c^2	*df*	RMSEA (90% CI)	SRMR	NNFI (TLI)	Expected Misfits[c]	Observed Misfits (Residuals, MI)	Raykov's Reliability
EFA (traditional)	207.599***[a]	54	0.084 (0.072, 0.097)	0.056	0.948	3.3	20, 20	0.847
EFA (full-info.)	146.675***[b]	54	0.066 (0.053, 0.078)	0.063	0.978	3.3	8, 13	0.903
E/CFA (Jöreskog)	46.516[b]	45	0.009 (0.000, 0.035)	0.038	1.000	3.3	0, 1	0.951
E/CFA (Gugiu)	4.569[b]	14	0.000 (0.000, 0.000)	0.020	1.007	1.1	0, 0	0.862

Model Validation: Posttest Sample

Final Model	c^2	*df*	RMSEA (90% CI)	SRMR	NNFI (TLI)	Expected Misfits[c]	Observed Misfits (Residuals, MI)	Raykov's Reliability	Test-Retest
EFA (traditional)	291.886***[a]	54	0.105 (0.093, 0.117)	0.065	0.914	3.3	23, 23	0.833	0.495
EFA (full-info.)	151.127***[b]	54	0.067 (0.055, 0.080)	0.069	0.979	3.3	8, 14	0.910	0.544
E/CFA (Jöreskog)	74.014**[b]	45	0.040 (0.023, 0.056)	0.048	0.993	3.3	3, 8	0.938	0.559
E/CFA (Gugiu)	28.896*[b]	14	0.052 (0.024, 0.078)	0.050	0.989	1.1	1, 6	0.864	0.525

[a] Normal Theory Weighted Least Squares c^2. [b] Satorra-Bentler c^2. [c] Refers to either the number of expected misfitting standardized residuals or the number of expected significant modification indices, not the total number of misfits.

Full-information EFA. As was the case in the traditional EFA, all 12 items in the full-information EFA loaded on the latent factor above the 0.3 threshold and thus were retained in the model (see Figure 1). Although the conceptual model was the same for both the traditional and the full-information EFAs, the goodness-of-fit indices and model reliability improved as a result of employing the polychoric correlation matrix and DWLS estimation method. Furthermore, both the 8 observed misfitting standardized residuals and the 13 significant MIs were closer to the expected number of misfits (3.3) (see Table 1). Similarly, the reliability of the full-information EFA was greater than that of the traditional EFA, $\rho = 0.903$. Moreover, the parameter estimates differed between the traditional EFA and the full-information EFA. Systematically, the full-information EFA produced factor loadings that were, on average, greater by 0.1 than the traditional EFA, while the full-information EFA error variance estimates were smaller.

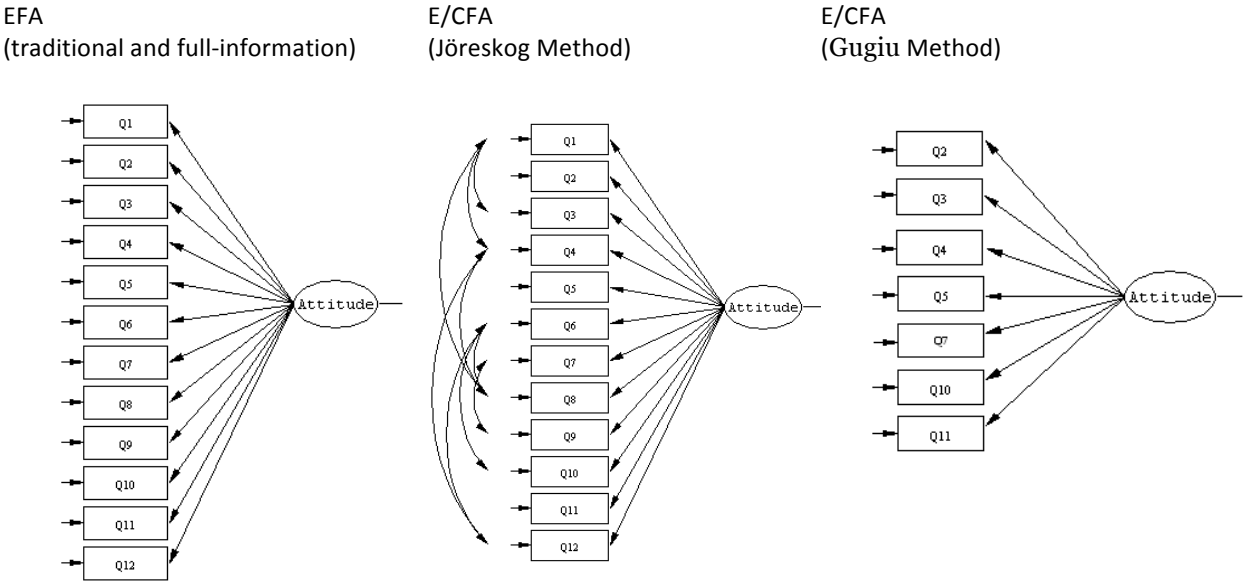

Figure 1. Final conceptual models of attitude toward Shakespeare resulting from the four model-building methods

Jöreskog E/CFA. In the SUFS dataset, all items could reasonably be related to one another, as all focused on some aspect of students' attitude toward Shakespeare. Therefore, correlated errors were freed beginning with the error associated with the largest MI until both the SB χ^2 for the model and the SB χ^2 difference test were nonsignificant. Following this approach resulted in a final modified model with 9 freed error covariances (see Figure 1). The goodness-of-fit indices for the model built with this approach showed significant improvement over the previous EFAs (see Table 1). Examination of the standardized residuals and MIs also indicated very good model fit with no misfitting standardized residuals and fewer than expected significant MIs. Furthermore, the reliability of the E/CFA model was higher, $\rho = 0.951$, and the parameter estimates greater than in the traditional EFA.

Gugiu E/CFA. Through the process of identifying large residuals, deleting corresponding items, and computing the SB χ^2 difference test, 5 items were removed, resulting in a final model that included 7 items (see Figure 1). To avoid overfitting the model, a reasonable case could have been made for retaining the last item (question 9), given the very small and nonsignificant SB χ^2 value, the well-fitted model indicated by the goodness-of-fit indices, and the fewer-than-expected observed misfitting standardized residuals and significant MI. However, the 7-item model was used as the final model because model fit improved and the construct validity was not adversely affected by the removal of item 9; the domain of "attitudes toward Shakespeare" was still well represented by the other questions (see Appendix).

The final model created by the Gugiu E/CFA approach also fit the data much better than the models created through EFA (see Table 1), underscored by the lack of any misfitting

standardized residuals or significant MIs, despite 1.1 misfits expected by chance. Although the goodness-of-fit indices trump the other models, the reliability of this model was somewhat lower, ρ = 0.862, owing to the fewer number of items. Hence, to make the basis of comparison more even, the Spearman-Brown prophecy formula was used to calculate the reliability under the assumption that a revised survey would contain 12 items of equal psychometric quality as the 7 items retained in this analysis. Under this assumption, the model reliability jumped to ρ = 0.914; the implication of this statistic, however, is that 5 new items would need to be written.

Model Validation

Posttest scores were used to test the models specified on the pretest data (see Table 1). Not unexpectedly, the model created through traditional EFA was not validated in the CFA. Several goodness-of-fit indices suggested poor model fit, echoed by the large number of misfitting standardized residuals (23) and significant MIs (23). The reliability of the validated model was acceptable according to Cronbach's alpha (0.840) and Raykov's ρ (0.833), with a low test-retest reliability of 0.495.

The CFA of the model created through the full-information EFA suggested that this model was better able to capture the relationships in the posttest data than the traditional EFA model (see Table 1). Although the traditional and the full-information EFA models were the same, the model created through full-information EFA utilized the polychoric correlation rather than the Pearson correlation in the input matrix, which resulted in less attenuated parameters and better model fit. However, although the goodness-of-fit indices suggest moderate fit of the full-information EFA model, the number of misfitting standardized residuals and significant MIs (8 and 14, respectively) was greater than the expected number (3.3). The reliability was greater than that of the traditional EFA, ρ = 0.910, as was the test-retest reliability, ρ = 0.544.

The models created by the two E/CFA methods both fared better in validation than the models created through EFA. The model specified by the Jöreskog method was able to capture the relationships in the posttest data very well, as demonstrated by the goodness-of-fit indices (see Table 1), but examination of the standardized residuals and MIs suggested there were associations in the posttest data that were not well represented by the model (8 significant MIs observed and 3.3 expected). In particular, the pattern of significant MI and corresponding error covariances to free differed between the pretest and posttest data. Of the nine freed correlated errors in the model of the pretest data, only four were significantly different from zero in the posttest data (see Table 2). Similarly, the posttest CFA highlighted eight large MI that were negligible in the pretest model. Of the 132 possible MI, the pre- and posttest models differ on 13, representing 10% disagreement. Despite these inconsistencies, the model showed a high degree of reliability, ρ = 0.938, and an acceptable level of test-retest reliability, ρ = 0.559.

Table 2

Correlated Errors for the Model of Attitude Toward Shakespeare (E/CFA, Jöreskog Method) in Pre- and Posttest Samples

Error Covariances Significant at Pre					Error Covariances Significant at Post but Not Pre (Modification Indices)[a]		
Freed Error Covariance	Pretest		Posttest		Freed Error Covariance	Pretest	Posttest
1 and 8	0.235	***	0.235	***	1 and 2	0.001	4.145
6 and 10	0.202	***	0.238	***	1 and 10	0.020	5.044
4 and 8	0.196	***	0.224	***	2 and 7	0.022	4.605
1 and 4	0.171	**	0.058		3 and 10	0.031	5.426
7 and 9	-0.196	***	0.050		4 and 10	0.231	4.899
6 and 8	0.152	**	0.064		4 and 11	0.000	4.267
6 and 12	0.138	**	0.088		7 and 12	3.069	6.059
4 and 12	0.130	*	0.212	***	9 and 11	2.072	6.711
1 and 3	-0.143	**	-0.082				

[a] Modification indices greater than 3.84, the critical value for a chi square distribution with df=1, indicate correlated errors that are likely to significantly improve model fit if allowed to freely covary.
* p < .05. ** p < .01. *** p < .001.

The Gugiu E/CFA model also fit the posttest data associations better than the EFAs, but slightly less well than the Jöreskog E/CFA model (see Table 1). As with the other models, the misfitting standardized residuals (1) and significant MIs (6) suggested some lack of fit (1.1 expected). The reliability was lower than that of the Jöreskog E/CFA model, $\rho = 0.864$, as was the test-retest reliability, $\rho = 0.525$. However, when estimated with 12 items instead of 7 using the Spearman-Brown prophecy formula, the internal ($\rho = 0.916$) and test-retest ($\rho = 0.655$) reliabilities were comparable to those of the Jöreskog E/CFA model.

Discussion

Repercussions of Inappropriate Analyses

Several lessons can be learned from the results of these models. First, it is clear that using analyses that have assumptions that are incompatible with the nature of the data can have severe implications for the results. In the case of the traditional EFA, an inappropriate correlation matrix resulted in underestimated factor loadings and contributed to poor model fit, as compared to the results of the full-information EFA. Unfortunately, ordinal data—the norm in evaluation survey studies—does not meet the assumptions of most of the methods of statistical analysis. Thus, traditional EFAs are generally not recommended unless data are continuous.

The goodness-of-fit indices suggest that the full-information EFA estimates did a slightly better job of capturing the observed relationships than those produced by the traditional EFA. The impact of using a polychoric correlation matrix rather than the Pearson correlation matrix is highlighted by the difference in the parameter estimates between the traditional EFA and full-information EFA. Because the Pearson estimates relationships from the coarse categories rather than continuous scores, it underestimates the associations between variables, which manifests in smaller parameter estimates and poorer model fit statistics. Although the full-information EFA is appropriate given the nature of the data and provides better model fit than the traditional EFA, it does not share the benefits of the E/CFA models. Without the ability to refine EFA models, not only does the model exhibit relatively poor fit, but it may also not be validated by a CFA, as was illustrated in this study. Thus, the EFA approach to model specification is not ideal. Although the full-information EFA should be chosen over the traditional EFA, the two methods of E/CFA are preferable for the purpose of establishing internal validity of a survey instrument, particularly when the stability of the latent construct over time is important.

Both E/CFA methods avoided the problems encountered by the two EFAs, thereby yielding robust initial models that also fit the data well in a CFA context. However, the Jöreskog method faced a challenge not faced by the other three methods;

namely, what constitutes a sufficiently substantive, theory-based rationale for freeing correlated errors. In the case of this study, it is unclear how the presence of so many correlated errors can be justified, particularly because the reason used to justify the presence of a correlated error must also explain why it does not apply in the case of absent correlated errors. On this basis, it is not intuitive why these nine pairs of questions share significant amounts of variance with each other but not the other items (see Appendix), and an *a priori* theoretical prediction of these relationships is highly unlikely. If only the errors that could be substantively justified were freed, as is often recommended, then the Jöreskog model in this study would have resembled that of the full information EFA as few, if any, of the correlated errors could be substantiated on theoretical grounds.

Furthermore, not only is the pattern of correlated errors difficult to justify and interpret, but the overall model is less stable. The size and significance of 10% of the MI differed between the pre- and posttest models, representing a small but non-ignorable amount of instability. This may result from the use of a 3.84 (critical value at α = 0.05) cut-off for large MIs, which may not be appropriate when used in conjunction with DWLS. The interpretation of MIs as chi-square difference is only appropriate for ML or robust maximum likelihood (RML) estimation methods with continuous variables; MI values are not directly analogous to chi-square differences under DWLS (K. Jöreskog, personal communication, June 23, 2013). In other words, the significance of MIs cannot be interpreted in the same way with the type of data found in the SUFS study. Therefore, even though the Jöreskog E/CFA method produced well-fitting models that were validated by a CFA on posttest data, the interpretability and instability of the models, as well as the uncertain appropriateness of the MI criteria, suggest that this method may also not yield stable results in the context of evaluation work.

The Gugiu E/CFA method is not limited by the issues of other methods but still produced well-fitting models that replicated in posttest CFAs. To its credit, the method produces models without correlated errors that are more easily interpretable than Jöreskog E/CFA models and that do not require *ad hoc* theoretical justifications. The method places a greater emphasis on standardized residuals than on MIs as a modification criterion and thus may be used with categorical and non-continuous data without concerns regarding the interpretation of the modification criteria. However, the limited number of items retained by this method does raise two concerns. First, with fewer items, models may be less reliable, a fact reflected in the models of the SUFS data. Second, iteratively reducing the number of items runs the risk of decreasing the model's construct validity. Fortunately, these issues can be easily addressed by adding more items and retaining items that are theoretically important for the construct. As suggested by the Spearman-Brown prophecy formula, adding more items with parallel psychometric properties will increase the reliability of these models to levels comparable to, if not better than, the reliability of models created through the other methods.

Recommendations for Practice

The model comparisons from the SUFS evaluation-based example illustrate a number of important take-aways for evaluation practice. First of all, these model comparisons demonstrate that using the appropriate analyses makes a meaningful difference, both in the resulting factor structure and in its stability over time. When this is considered in light of the fact that survey data is ubiquitous in evaluation work, utilized in 45% of the reviewed articles, it is clear that using the appropriate factor analytic method is fundamental in determining the quality of the data. From the current study, there are several steps that evaluators can take to help ensure well-fitting, stable factor structures.

First, for many of the constructs of interest within the field, including attitudes, beliefs, and knowledge, factor analysis is more appropriate than principal components analysis. When choosing a correlation coefficient for use in the input correlation matrix, the polychoric correlation coefficient is most appropriate for ordinal data (Brown, 2006), and the Pearson correlation coefficient is most appropriate for data that is truly continuous, with at minimum 15 answer choices (Jöreskog & Sörbom, 1996; Schumacker & Lomax, 2010). Parallel analysis should be utilized for factor selection, because it is more accurate and more clearly interpretable than Scree plots or the Kaiser criterion (Hayton, Allen, & Scarpello, 2004). For the method of estimation, diagonal weighted least squares (DWLS) in conjunction with the asymptotic covariance matrix is best used when the data are ordinal, whereas maximum likelihood can be used when data are continuous. For all types of data, the model can be modified using the 0.3 cut-off (Tabachnick & Fidell, 2001). The results of the current study also

suggest that the Gugiu method may be a particularly effective way of modifying the model, without some of the disadvantages of other methods. Taken together, these recommendations can ensure that the surveys used in evaluation work have well-fitting and stable factor structures, which in turn will ensure that clear and accurate conclusions are drawn from the data.

References

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Brown, T. A., White, K. S., Forsyth, J. P., & Barlow, D. H. (2005). The structure of perceived emotional control: Psychometric properties of a revised Anxiety Control Questionnaire. *Behavior Therapy, 35*(1), 75-99.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Beverly Hills, CA: Sage.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*(3), 327-346.

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

Gilley, W., & Uhlig, G. (1993). Factor analysis and ordinal data. *Education, 114*(2), 258-264.

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.

Gugiu, P. C. (2011). *Exploratory Factor Analysis within a Confirmatory Factor Analysis Framework (E/CFA)*. Expert lecture presented at the 2011 American Evaluation Association conference in Anaheim, California.

Gugiu, P. C., Coryn, C., Clark, R., & Kuehn, A. (2009). Development and evaluation of the short version of the Patient Assessment of Chronic Illness Care instrument. *Chronic Illness, 5*(4), 268-276.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 32*, 179-185.

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

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202.

Jöreskog, K. G. & Moustaki, I. (2006). *Factor analysis of ordinal variables with full information maximum likelihood*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jöreskog, K. G., & Sörbom, D. (1979). *Advances in factor analysis and structural equation models* (p. 105). J. Magidson (Ed.). Cambridge, MA: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Lincolnwood, IL: Scientific Software International.

Kampen, J., & Swyngedouw, M. (2000). The ordinal controversy revisited. *Quality and Quantity, 34*(1), 87-102.

Lee, B. K., Patall, E. A., Cawthon, S. W., & Steingut, R. R. (2015). The effect of drama-based pedagogy on preK–16 outcomes: A meta-analysis of research from 1985 to 2012. *Review of Educational Research, 85*, 3-49.

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

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*(4), 485-500.

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*(2), 315-323.

Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. *Behavior Therapy, 35*(2), 299-331.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. Clogg (Eds.), *Latent variable analysis: Applications for developmental*

research (pp. 399-419). Thousand Oaks, CA: Sage.

Schumacker, R. E. & Lomax, R. G. (2010). *A beginner's guide to SEM*. Manwah, NJ: Lawrence Erlbaum Associates.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Strand, S. (2009). *Attitude to Shakespeare among Y10 students: Final report to the Royal Shakespeare Company on the Learning and Performance Network student survey 2007-2009*. Warwick, England: Centre for Educational Development, Appraisal and Research.

Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics. Boston: Pearson.

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

Yeomans-Maldonado, G., Gugiu, C. P., & Enciso, P. (2013). *Using the Rasch model to assess the interest for Shakespeare and illustrate rating scale diagnostics*. Poster presented at the Modern Modeling Methods Conference, Storrs, CT.

Yuan, Y. C. (2000). *Multiple imputation for missing data: Concepts and new development* (SAS Tech. Rep. No. P267-25). Rockville, MD: SAS Institute.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.

Appendix

Attitudes Toward Shakespeare Survey, Based on Warwick Survey (Strand, 2009)

What I think about Shakespeare	No	Don't know	Yes
1. Everyone should read Shakespeare	❑	❑	❑
2. Shakespeare is fun	❑	❑	❑
3. Shakespeare's plays are difficult for me to understand	❑	❑	❑
4. Shakespeare's plays help us understand each other better	❑	❑	❑
5. I would like to do more Shakespeare	❑	❑	❑
6. Some of the people in Shakespeare's plays are like people you meet today	❑	❑	❑
7. I tell my friends in other classes about Shakespeare	❑	❑	❑
8. It is important to study Shakespeare's plays	❑	❑	❑
9. Shakespeare is only for old people	❑	❑	❑
10. Things that happen in Shakespeare's plays can happen in real life	❑	❑	❑
11. Shakespeare is boring	❑	❑	❑
12. I have learned something about myself by learning about Shakespeare	❑	❑	❑