A Comparison of Algorithms for Dimensionality Analysis

Sedat Sen, Allan S. Cohen, and Seock-Ho Kim

1 Introduction

Item response theory (IRT) models have been widely used for various educational and psychological testing purposes such as detecting differential item functioning (DIF), test construction, ability estimation, equating, and computer adaptive testing. The main assumption underlying these models is that local independence holds with respect to the latent ability being modeled (Lord and Novick 1968). It is important, therefore, to show that the unidimensionality assumption holds before any unidimensional IRT modeling is applied. Otherwise, violations of the unidimensionality assumption may have a considerable and negative effect on parameter estimation (Ackerman 1989; Reckase 1979). Ackerman (1992) also showed that the presence of multidimensionality may also cause DIF. Correct identification of the internal test structure also helps to examine how well the test measures the underlying structure. Tate (2003) noted that strict dimensionality and essential dimensionality are two types of dimensionality in the traditional IRT context. The former refers to the minimum number of examinee latent abilities required to estimate a monotone and locally independent model (McDonald 1981; Stout 1990) while the latter refers to a test with a single dominant factor and one or more minor factors (Stout 1987, 1990).

Because of the centrality of the unidimensionality assumption to many applications of IRT, the dimensionality assessment problem has been the focus of considerable study. Excellent reviews are provided by Hattie (1984) and Tate (2003). Dimensionality assessment is more problematic for categorical variables than continuous variables. When the variables are continuous, traditional factor analysis techniques can be used to identify factors that may be used to explain the observed data. Data in social science are often categorical in nature (e.g., dichotomous

S. Sen (✉) • A.S. Cohen • S.-H. Kim
University of Georgia, Athens, GA 30602, USA
e-mail: sedatsen@uga.edu; acohen@uga.edu; shkim@uga.edu

R.E. Millsap et al. (eds.), *New Developments in Quantitative Psychology*,
Springer Proceedings in Mathematics & Statistics 66, DOI 10.1007/978-1-4614-9348-8_14,

and polytomous item responses). These types of data normally fail to meet the distributional requirements of the traditional linear factor analysis. As a result, factor analysis may not be directly applicable to categorical variables because spurious factors (called difficulty factors) may emerge when using Pearson product-moment correlations (Ackerman et al. 2003; McLeod et al. 2001). As a result, the number of dimensions may be overestimated (Bock et al. 1988). In order to deal with this situation, tetrachoric correlations can be used instead of Pearson correlations to deal with dichotomous nature of item scores (Hulin et al. 1983; Knol and Berger 1991; Parry and McArdle 1991). However, it should be noted that tetrachoric matrices for item-level data may not always be positive definite, as required for modern factor analysis techniques. Another problem with this method is the estimation of tetrachoric correlations which can be difficult to implement when correlations are very close to unity (Thissen and Wainer 2001).

A number of different methods have been proposed to assess test dimensionality for item-level, beginning with work by Christoffersson (1975) and Muthén (1977). Some relatively new methods based on item factor analysis (IFA) have also been proposed. There are a wide range of IFA models within structural equation modeling (SEM) and IRT including full-information maximum-likelihood (FIML) estimation (Bock et al. 1988), the algorithm in the software package LISCOMP (Muthén 1978), nonlinear factor analysis (McDonald 1982), and factor analysis of the tetrachoric correlations between all item pairs (Knol and Berger 1991). The FIML estimation method is based on analyzing the entire item response pattern while the other three use bivariate information. These parametric approaches also differ in the estimation algorithms used. There are several methods available for IFA model parameter estimations. Among these are FIML, unweighted least squares (ULS), weighted least squares (WLS), and its modified extensions such as modified WLSM and WLSMV. In addition to these parametric approaches, there are also some nonparametric approaches for dimensionality assessment such as the algorithm in the computer software DIMTEST (Nandakumar and Stout 1993) and in the software DETECT (Kim 1994; Zhang and Stout 1999a,b). These techniques are designed to test essential dimensionality of a set of test items.

More recently, a number of studies of dimensionality have focused on comparison of different methods (e.g., Nandakumar 1994; Nandakumar and Yu 1996; Tate 2003), the effect of applying unidimensional IRT to multidimensional items (e.g., Ackerman 1989), and the effect of guessing parameter (Tate 2003; Stone and Yeh 2006). Although it has been more than three decades since Lord's (1980) call for a statistical significance test for assessing dimensionality of a test, there is still no general test for dichotomous items. Hattie (1984) noted that most indices were inappropriate for dimensionality assessment for the case of dichotomous variables.

Even though substantial work has been done on techniques used for dimensionality checking, there has been a lack of study on the effectiveness of different software packages implementing these techniques. The purposes of this study were to (1) compare two popular software packages, Mplus and TESTFACT, with respect to their effectiveness for checking dimensionality in multiple-choice tests and (2) compare different criteria used in these programs. We also included SAS in our empirical analyses to examine what would happen if Pearson correlations instead of tetrachoric correlations were used. In addition to use of Pearson correlations, we also analyzed the empirical data set with SAS to provide a tetrachoric correlation for completeness. Guessing parameters and the size of correlations between dimensions were manipulated to explore possible interaction between these effects. Three indices based on the proportion of variance, RMSR reduction, and a chi-square difference test were used to examine dimensionality. The research included two parts, a simulation study using a Monte Carlo approach and an application with data from a large midwestern university mathematics placement testing program.

1.1 Software

There are a number of computer programs used for both parametric and nonparametric approaches. Because the focus of this study is on parametric approaches, software packages designed for nonparametric approaches (e.g., DIMTEST) are not discussed in detail. IFA-based procedures for applications with dichotomously scored items can be implemented in software programs, including Mplus (Muthén and Muthén 2010), NOHARM (Fraser and McDonald 1988), and TESTFACT (Wilson et al. 2003). Although the goal of these three programs is the same, the methods employed by each are different. They differ in sample statistics, estimation methods, and how guessing is handled (Stone and Yeh 2006).

1.1.1 Mplus

Mplus can handle categorical, continuous, and ordinal types of data. The software permits users to perform both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to estimate unidimensional or multidimensional models. Estimation of dichotomous data is done using tetrachoric correlations via the following methods: ULS, WLS, WLSM, and WLSMV. Mplus also provides several fit indices including chi-square test statistics, root mean square residuals (RMSR), root mean square error of approximation (RMSEA), and comparative fit index (CFI). In addition, both orthogonal (varimax) and oblique (promax) rotations of the initial solution are available. There is no option for handling the guessing parameter in the three-parameter model. The Mplus manual also indicates that the relationship between the extracted factors and the observed indicators is provided using probit regression of items on factors.

1.1.2 TESTFACT

TESTFACT was designed to perform nonlinear, exploratory full-information IFA on dichotomous items. This software uses marginal maximum likelihood (MML) estimation in combination with an expectation-maximization (EM) algorithm. The estimates are obtained in TESTFACT using all of the information in the item responses rather than use of an item covariance or correlation matrix as is implemented in Mplus, and TESTFACT can handle the guessing parameter for factor analyses. TESTFACT prompts the user to specify the number of factors and the guessing parameters, if guessing is assumed in the model. The guessing parameter can be input by either assuming a single value or providing estimated guessing parameters for each item from another software package such as BILOG or MULTILOG. TESTFACT calculates chi-square statistics which can be used for model comparison. However, TESTFACT requires nonzero frequencies for each item pattern in order to calculate this value. Problematic correlations due to extreme proportions are replaced with admissible values using Thurstone's centroid method (Tate 2003). A smoothing option is also available if the correlation matrix is nonpositive definite. Although TESTFACT can produce the output of a residual matrix, there is no residual-based fit index. RMSR value can be calculated from residual matrix. As with Mplus, varimax and promax rotations of the initial solution can be obtained in TESTFACT.

1.1.3 SAS

SAS provides a way of doing common-factor and component analysis using the proc factor statement. It offers a range of methods in EFA to select the number of factors, extraction and rotation methods. These analyses can be done using either raw data or correlation/covariance matrix. SAS is often used for continuous variables with Pearson correlation coefficients. Although it is not very practical, one can conduct factor analysis for dichotomous type data by providing a tetrachoric correlation matrix. The extraction methods available in SAS include principal component analysis, principal factor analysis, iterated principal factor analysis, ULS factor analysis, maximum likelihood (canonical) factor analysis, alpha factor analysis, image component analysis, and Harris component analysis. Proc factor produces the residual correlation matrix and the partial correlation matrix. EQUAMAX, ORTHOMAX, QUARTIMAX, PARSIMAX, and VARIMAX; and two oblique rotation methods, PROCRUSTES and PROMAX, can be obtained with proc factor statement. In order to help in determining the number of components or factors, the scree plot, percentage of variance, and Kaiser's rule can be obtained from output.

2 Method

Dimensionality assessment results for the simulated data are given first, followed by results for the real data. Only the results of applying Mplus and TESTFACT are presented in the simulation study. Additional results from SAS are reported for the real data study. As mentioned earlier, number of dimensions, correlation, and guessing parameter were manipulated. Results from uncorrelated factors and those from correlated ($r = 0.3$) factors are presented for both Mplus and TESTFACT in results section. Values in the each cell represent the correct number of identifications out of ten replications. First rows of two tables are the same since only the uncorrelated condition is possible for unidimensional data. EFA was carried out using WLSM for all Mplus analyses. Similarly, exploratory analyses in TESTFACT were conducted using FIML for one to five factors. Hereafter, we refer Mplus as it is applied with WLSM and TESTFACT as it is applied with FIML in simulated data analyses. Maximum likelihood extraction method was used for SAS analyses in empirical data set.

2.1 Simulated Data

Examinees' responses to ten different 60-item tests were simulated based on the dichotomous, multidimensional logistic IRT model. Each of the ten tests was replicated ten times. One-, two-, and three-dimensional data sets were simulated for each replication. Two guessing conditions were simulated in which the guessing parameter was set at 0 and 0.25. There is a certain amount of correlation among factors in most educational tests. To simulate this, a correlation of 0.3 was used in addition to correlations of 0 between factors. Data were generated for 2,000 respondents for each test using WINGEN 3.0 (Han 2006) software. Following the conditions in (Yeh 2007), distribution of latent traits was normal with mean of zero and standard deviation of 0.1 for unidimensional data. While mean of latent traits remained the same for each dimension, different values for standard deviations were used to obtain the desired correlation ($r = 0.3$) between dimensions. Because this was the only way to obtain correlated dimensions in WINGEN. Item parameter distributions were $N(1,\ 0.36)$ and $N(0,\ 1.43)$ for discrimination and difficulty parameters, respectively. Ten data conditions were simulated by changing correlation, guessing, and the number of dimensions.

2.2 Real Data

The data used in this study were from a test designed to measure calculator proficiency in pre-calculus mathematics. A total of 765 students took a special, experimental form of this 28-item test. Each item had five choices. Students were allowed to use a calculator on the first 14 items, but were not allowed to do so on the second 14 items. Only the second 14 items, which allowed no calculator use, were analyzed for this study. The test was originally constructed as a unidimensional instrument.

The multidimensional item response theory (MIRT) model for dichotomously scored items with a guessing parameter (Bock et al. 1988) was used to analyze the data. The probability of a correct response to item j can be given as

$$P(U_j = 1|\theta) = g_j + (1-g_j)\Phi[z_j(\theta)] = g_j + (1-g_j)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_j(\theta)} \exp(-t^2/2)dt, \tag{1}$$

where
$$\Phi[z_j(\theta)] = \Phi\left(c_j + \sum_{k=1}^{K} a_{jk}\theta_k\right) = \Phi\left(\frac{\delta_j + \sum_{k=1}^{K} \alpha_{jk}\theta_k}{\sigma_j}\right), \tag{2}$$

g_j is the guessing parameter, c_j is the intercept or easiness parameter, a_{jk}s are the slopes, θ_ks are latent variables equivalent to the vector θ, δ_j is the standard difficulty or negative threshold (i.e., $-\gamma_j$), α_{jk}s are items regression coefficients or factor loadings to the respective dimensions from 1 to K (i.e., λ_{jk}), and $\sigma_j = \sqrt{1 - \sum_{k=1}^{K} \alpha_{jk}^2}$. If we let $d_j = \sqrt{1 + \sum_{k=1}^{K} a_{jk}^2}$, then $\alpha_{jk} = a_{jk}/d_j$ and $\delta_j = c_j/d_j$ (cf. McLeod et al. 2001, p. 199).

TESTFACT was used to obtain the a_{jk}s and c_j for each item under MMLE. The g_j parameters are not estimated with other parameters in TESTFACT and must be specified by the user. BILOG-MG (Zimowski et al. 2002; see also Mislevy and Bock 1990) was used to obtain the lower asymptote estimates using all default options with an exception of the option for items with five choices.

2.3 *Decision Criteria*

Several methods for determining the number of factors have been proposed. Eigenvalues, fit indices, and proportions of variance are typically used to examine the factor structure of a set of items. Scree plots involve plotting the eigenvalues for all possible numbers of factors and looking for the elbow in the plot (i.e., the point at which the eigenvalues tend to stop decreasing). The number of factors is taken as one fewer than the solution corresponding to the elbow. This approach is criticized as being very subjective because the location of the elbow is not always very clear. Kaiser (1960) proposed a heuristic rule called the eigenvalue-greater than-one (K1) rule in which each eigenvalue greater than one is taken to indicate a component, and his rule was applied by some to common-factor analysis (Mulaik 2009, p. 186). The proportion of variance is an index for the substantive importance of factors. This procedure is fairly straightforward and suggests keeping the number of factors needed to account for a specified percentage of the variance (e.g., 80% or 90%).

In addition to using eigenvalues, there are several residuals and fit indices that can be used for dimensionality assessment such as chi-square fit statistics, RMSEA, and RMSR. These statistics indicate the differences between observed values and estimated values. Smaller values are taken to indicate better fit. A cutoff value of 0.05 or less for the RMSR and RMSEA statistic has been suggested as a guide indicating an acceptable number of factors (Browne and Cudeck 1993). Hu and Bentler (1999) offer different cutoff values for these indices, specifically RMSR < 0.08 and RMSEA < 0.06. The chi-square test evaluates whether the observed data correspond to the expected data. The chi-square statistic is dependent on sample size, but RMSEA is not. Thus, for larger samples, it may be more appropriate to use RMSR and RMSEA to assess the model fit. In addition to using cutoff values, the model fit decision can be made based on the percentage of reduction of the

RMSR (Tate 2003). Tate (2003) suggests that factors be added to the model until the percent of RMSR reduction is less than 10%.

Dimensionality decisions in this study were based on the following three criteria: percentage of the RMSR reduction, chi-square difference test, and proportion of variance. As mentioned earlier, the assessment of test dimensionality in TESTFACT can be done using a test of the change of the chi-square fit statistic due to adding a factor to the model. In Mplus, RMSR reduction approach was used. However, proportion of variance criterion was used for all of the software packages.

3 Results

3.1 *Simulated Data Results*

3.1.1 One-Dimensional Tests

One-dimensional data with two guessing situations were analyzed in Mplus and TESTFACT programs. The fit statistics of the bifactor model were compared with those for the 1-factor model. As can be seen in the first rows of the two tables, TESTFACT and Mplus did not correctly identify the unidimensional structure when no guessing was simulated. When guessing was simulated, however, TESTFACT performed better than Mplus as expected (Table 1).

3.1.2 Two-Dimensional Tests

Within each test form, the correlations between factors were fixed at either 0 or 0.30. Mplus provided no correct identification when no guessing was simulated regardless of the simulated correlation. TESTFACT correctly identified 80% and 50% in the no-guessing simulation, however, for uncorrelated and correlated cases, respectively. As in the one-dimensional case, TESTFACT did better than Mplus, when guessing was simulated for two-dimensional data. Mplus correctly identified four cases when two-dimensional uncorrelated data were simulated with a guessing effect.

Table 1 Number of correct identification for TESTFACT and Mplus for 1- to 3-factor models ($r = 0$)

	$c = 0$		$c = 0.25$	
# of dimensions	TESTFACT	Mplus	TESTFACT	Mplus
1	0/10	0/10	10/10	6/10
2	8/10	4/10	5/10	4/10
3	4/10	2/10	10/10	10/10

Table 2 Number of correct identifications for TESTFACT and Mplus for 1- to 3-factor models ($r = 0.3$)

	$c = 0$		$c = 0.25$	
# of dimensions	TESTFACT	Mplus	TESTFACT	Mplus
1	0/10	0/10	10/10	6/10
2	5/10	0/10	8/10	0/10
3	9/10	1/10	7/10	0/10

3.1.3 Three-Dimensional Tests

The results from applying the 3-factor models indicated that TESTFACT performed better than Mplus in each of the four conditions. Correct identification rates range for TESTFACT ranged from 40% to 100%. Similar rates for Mplus were low in each of the three conditions except for the case for which guessing with zero correlation was simulated (Table 1).

3.2 *Real Data Results*

3.2.1 Full-Information Item Factor Analysis with TESTFACT

Summary indices for TESTFACT, Mplus, and SAS are presented in Table 3 for 1- to 4-factor solutions. The TESTFACT/BILOG rows show indices for the MIRT model with the g_j estimates from BILOG-MG since TESTFACT cannot estimate the lower asymptote. The TESTFACT/C rows show the results from the same MIRT model but the g_j were assumed to have a fixed value of 0.20 (because all items had five choices). The g_j parameters in this case are not separately estimated. The TESTFACT rows contain MIRT models without the g_j term.

The difference between the chi-squared goodness of fit values from the 1-factor solution to 2-factor solution was not significant for all the three cases with TESTFACT. (The critical value at the 0.05 level is $\chi^2(13) = 19.19$.) The respective critical values at the 0.05 nominal level are $\chi^2(12) = 21.20$ and $\chi^2(11) = 8.52$ for the 2-factor to 3-factor solution and for the 3-factor to the 4-factor solution. Although the 2-factor solution to the 3-factor solution shows a significant reduction in the goodness of fit values, the 1-factor solution seems to be a reasonable choice for the data.

The cumulative proportions of the variance accounted for appear to increase as the number of factors increases. The 1-factor solution for TESTFACT/C yielded a higher proportion of variance accounted for than was observed for a higher number of factors. Table 3 contains the summary of items with high Promax loadings (i.e., α_{jk} or $\lambda_{jk} > 0.30$). Although all TESTFACT methods yielded proper extraction results for the 4-factor solution for the TESTFACT/BILOG, TESTFACT/C, and TESTFACT cases, the Promax rotation failed to yielded reasonable loading results

Table 3 Numbers of items with high promax loadings and correlations between factors

	One factor	Two factors		Three factors			Four factors			
	I	I	II	I	II	III	I	II	III	IV
Mplus/WLSMV	14	6	6	6	3	4	Heywood case			
Mplus/WLS	14	6	7	Heywood case due to over-factoring						
Mplus/ULS	14	8	5	4	3	4	8	1	3	1
TESTFACT/BILOG	14	10	1	6	2	3	Not available			
TESTFACT/C	14	10	1	6	2	3	Not available			
TESTFACT	14	9	1	6	1	3	Not available			
SAS	13	8	7	7	5	6	4	6	4	4
SAS/Tetrachoric	14	9	7	7	3	6	7	4	4	3
Correlation between factors										
Mplus/WLSMV										
II		0.71		0.58						
III				0.70	0.53					
IV							Heywood case			
Mplus/WLS										
II		0.68								
III				Heywood case due to over-factoring						
IV										
Mplus/ULS										
II		0.64		0.57			0.48			
III				0.68	0.52		0.75	0.51		
IV							0.39	0.34	0.32	
TESTFACT/BILOG										
II		0.65		0.66						
III				0.73	0.59					
IV							Not available			
TESTFACT/C										
II		0.65		0.67						
III				0.73	0.59					
IV							Not available			
TESTFACT										
II		0.64		0.68						
III				0.75	0.61					
IV							Not available			
SAS										
II		0.31		0.25			0.25			
III				0.27	0.17		0.28	0.11		
IV							0.16	0.07	0.17	
SAS/Tetrachoric										
II		0.43		0.42			0.38			
III				0.36	0.33		0.20	0.31		
IV							0.22	0.33	0.30	

Table 4 TESTFACT/BILOG loadings for 1- to 4-factor solutions

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.50	0.55	−0.09	0.56	−0.09	−0.04	Not available			
2	0.44	0.21	0.29	0.27	0.31	−0.07				
3	0.45	0.48	−0.04	−0.03	−0.12	0.59				
4	0.34	−0.22	0.60	−0.18	0.62	−0.05				
5	0.49	0.35	0.19	−0.02	0.16	0.41				
6	0.49	0.49	0.01	0.29	−0.02	0.26				
7	0.47	0.27	0.25	0.19	0.25	0.09				
8	0.44	0.47	−0.04	0.54	−0.03	−0.07				
9	0.48	0.25	0.29	0.16	0.28	0.12				
10	0.47	0.54	−0.09	0.50	−0.10	0.06				
11	0.42	0.41	0.01	0.37	0.00	0.07				
12	0.49	0.43	0.09	0.03	0.04	0.39				
13	0.50	0.32	0.23	0.34	0.24	−0.05				
14	0.50	0.52	−0.03	0.49	−0.04	0.04				
Correlation between factors										
Factor										
II		0.65		0.66						
III				0.73	0.59					
IV							Not available			

and, therefore, are reported as "Not available." For the 2-factor solution, one item consistently loaded on the second factor while other items mainly loaded on the first factor. High correlations were obtained between pairs of the factors under the 2- and 3-factor solutions.

Tables 4–6 contain the loadings for the 1-factor, 2-factor, and 3-factor solutions for TESTFACT/BILOG, TESTFACT/C, and TESTFACT, respectively. In Table 4, the 2-factor solution yielded only one item, Item 4, on the second factor. This item asks for the complete factoring of $12ax^2 - 9ax - 3a$. The same item as well as Item 2 yielded relatively high loadings on the second factor. Items 3, 5, and 12 had high loadings on the third factor for the 3-factor solution. Also for the 3-factor solution, the number of items loading on the first factor decreased from ten on the 2-factor solution to six on the 3-factor solution. Similar patterns of loadings were observed for the TESTFACT/C and TESTFACT solutions.

3.2.2 Factor Analysis with Mplus

Summary results are presented in Table 3 for results from Mplus for each of the three different estimation methods. For the EFA, WLSMV (i.e., weighted least squares parameter estimates using a diagonal weight matrix with standard errors and mean- and variance-adjusted chi-square test statistics that use a full weight matrix) is the default estimation in Mplus (Muthén and Muthén 2010, pp. 531–532). Two other

Table 5 TESTFACT/C loadings for 1- to 4-factor solutions

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.50	0.56	−0.11	0.56	−0.19	−0.05	Not available			
2	0.44	0.21	0.29	0.27	0.30	−0.08				
3	0.44	0.46	−0.03	−0.05	−0.12	0.59				
4	0.34	−0.23	0.61	−0.18	0.63	−0.06				
5	0.49	0.34	0.20	0.00	0.16	0.38				
6	0.49	0.48	0.01	0.29	−0.02	0.25				
7	0.47	0.27	0.26	0.18	0.27	0.09				
8	0.44	0.47	−0.04	0.53	−0.02	−0.07				
9	0.49	0.25	0.30	0.16	0.29	0.11				
10	0.49	0.55	−0.09	0.52	−0.10	0.05				
11	0.42	0.42	0.01	0.37	0.00	0.07				
12	0.47	0.41	0.09	0.03	0.04	0.35				
13	0.50	0.32	0.23	0.33	0.25	−0.04				
14	0.49	0.50	−0.04	0.47	−0.04	0.03				
Correlation between factors										
Factor										
II		0.66		0.67						
III				0.73	0.59					
IV							Not available			

Table 6 TESTFACT loadings for 1- to 4-factor solutions

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.50	0.56	−0.10	0.57	−0.07	−0.03	Not available			
2	0.42	0.27	0.21	0.31	0.26	−0.09				
3	0.41	0.43	−0.03	−0.06	−0.12	0.58				
4	0.31	−0.19	0.60	−0.19	0.63	−0.04				
5	0.47	0.35	0.16	0.00	0.15	0.37				
6	0.46	0.47	−0.00	0.27	−0.03	0.26				
7	0.41	0.28	0.18	0.19	0.19	0.08				
8	0.41	0.48	−0.09	0.57	−0.09	−0.08				
9	0.46	0.30	0.22	0.17	0.25	0.12				
10	0.45	0.53	−0.11	0.51	−0.12	0.04				
11	0.40	0.42	−0.02	0.39	−0.01	0.03				
12	0.50	0.45	0.07	0.03	0.04	0.45				
13	0.40	0.28	0.16	0.28	0.19	−0.01				
14	0.43	0.45	−0.03	0.37	−0.03	0.10				
Correlation between factors										
Factor										
II		0.64		0.68						
III				0.75	0.61					
IV							Not available			

estimation methods used in this study were WLS and ULS. The proportions of variance accounted for by the respective factors were based on the varimax rotated loadings instead of the initial extraction. The total variance accounted for is reported as sum (see Table 7).

In terms of the model selection using the various indices from Mplus-type computer programs, Hu and Bentler (1999) recommend a model with a value of RMSR less than 0.08 for selection of a model. That recommendation, however, was not based on analysis of binary variables. The values of RMSR's from the Mplus runs using Hu and Bentler's RMSR < 0.08 suggested 1-factor solution provided reasonably good fit under all estimation methods. Stone and Yeh (2006) suggested a model with a value of RMSR less than 0.05 could be chosen in conjunction with factor analysis for a set of dichotomously scored items. Using the Stone and Yeh suggestion, then WLSMV and ULS estimation methods yielded a 2-factor solution rather than a 1-factor solution.

Hu and Bentler (1999) suggested a value of an RMSEA less than 0.06 as indicating good fit. Using this criterion, WLSMV and WLS both would suggest a 1-factor solution. Stone and Yeh (2006) recommended an RMSEA of less than 0.05. Using this criterion, a 1-factor solution would be recommended. In addition to RMSR and RMSEA, Stone and Yeh also suggested a chi-square divided by its degrees of freedom of less than 1.4 as an indicator of reasonable fit. Using this latter criterion, the 1-factor solution would be selected based on WLSMV and WLS estimates. Tate (2003) recommended a 10% reduction. Using this criterion, WLSMV would have yielded a 3-factor solution, WLS a 2-factor solution; and ULS a 4-factor solution.

As can be seen in Table 3, Heywood cases resulted for both WLSMV and WLS estimation, possibly due to over-factoring. The patterns of factor loadings were different from those with TESTFACT although high correlations were obtained between pairs of the available Promax factors. Tables 8–10 contain the factor loadings for 1- to 4-factor solutions for the three estimation methods using Mplus. The 2-factor solution presented in Table 8 shows six items as loading on the first factor (Items 1, 6, 8, 10, 11, and 14) and six items on the second factor (Items 2, 3, 4, 5, 9, and 12). For the 3-factor solution, six items (Items 1, 6, 8, 10, 11, and 14) loaded on the first factor, three items (Items 2, 4, and 9) on the second factor, and four items (Items 3, 5, 6, and 12) on the third factor. Similar patterns were observed for WLS and ULS.

3.2.3 Factor Analysis with SAS

Two different sets of SAS results are reported based on Phi coefficients (Table 11) and Tetrachoric correlations (Table 12). Adding factors increased the cumulative proportions of variance (see Table 7). Using the 20 % criterion suggested in Reckase (1979), results for both coefficients yielded a 1-factor solution.

The 2-factor solution shown in Table 11 indicated nine items loaded on the first factor (Items 1, 3, 4, 6, 8, 10, 11, 12, and 14), and seven items on the second factor

Table 7 Indices for item factor analyses

	One factor	Two factors		Three factors			Four factors			
	I	I	II	I	II	III	I	II	III	IV
Mplus/WLSMV										
$\chi^2(df)\ p$	73.7(65) 0.22	55.7(55) 0.45		40.4(46) 0.70			Heywood case			
RMSEA	0.013	0.004		0.000			Heywood case			
RMSR	0.0540	0.0462		0.0377			Heywood case			
Mplus/WLS										
$\chi^2(df)\ p$	97.7(77) 0.06	67.1(64) 0.37		Heywood case due to over-factoring						
RMSEA	0.019	0.008		Heywood case due to over-factoring						
RMSR	0.0649	0.0570		Heywood case due to over-factoring						
Mplus/ULS										
RMSR	0.0538	0.0457		0.0371			0.0318			
TESTFACT/BILOG										
$\chi^2(df)$	1533.46(736)	1512.80(723)		1488.25(711)			1478.97(700)			
$\Delta\chi^2(df)$		20.66(13)		24.55(12)			9.28(11)			
TESTFACT/C										
$\chi^2(df)$	1540.08(736)	1519.16(723)		1493.91(711)			1485.50(700)			
$\Delta\chi^2(df)$		20.92(13)		25.25(12)			8.41(11)			
TESTFACT										
$\chi^2(df)$	1538.78(736)	1519.59(723)		1498.39(711)			1489.87(700)			
$\Delta\chi^2(df)$		19.19(13)		21.20(12)			8.52(11)			
Proportion of variances accounted for by factors										
Mplus/WLSMV										
Proportion	0.27	0.16	0.14	0.14	0.09	0.11	Heywood case			
Sum	0.27	0.30		0.34			Heywood case			
Mplus/WLS										
Proportion	0.30	0.17	0.16	Heywood case due to over-factoring						
Sum	0.30	0.33		Heywood case due to over-factoring						
Mplus/ULS										
Proportion	0.27	0.18	0.12	0.14	0.09	0.11	0.15	0.06	0.11	0.07
Sum	0.27	0.30		0.34			0.39			
TESTFACT/BILOG										
Proportion	0.22	0.21	0.02	0.19	0.02	0.02	0.17	0.02	0.02	0.01
Cumulative	0.22	0.21	0.23	0.19	0.21	0.23	0.17	0.19	0.21	0.22
TESTFACT/C										
Proportion	0.30	0.20	0.02	0.18	0.02	0.02	0.17	0.02	0.02	0.01
Cumulative	0.30	0.20	0.22	0.18	0.20	0.22	0.17	0.19	0.21	0.22
TESTFACT										
Proportion	0.17	0.18	0.02	0.17	0.02	0.02	0.17	0.02	0.02	0.01
Cumulative	0.17	0.18	0.20	0.17	0.19	0.21	0.17	0.19	0.21	0.22
SAS										
Proportion	0.21	0.21	0.08	0.21	0.08	0.07	0.21	0.08	0.07	0.07
Cumulative	0.21	0.21	0.29	0.21	0.29	0.36	0.21	0.29	0.36	0.44
SAS/Tetrachoric										
Proportion	0.32	0.32	0.08	0.32	0.08	0.07	0.32	0.08	0.07	0.07
Cumulative	0.32	0.32	0.40	0.32	0.40	0.47	0.32	0.40	0.47	0.54

Table 8 Mplus/WLSMV factor loadings for models with 1- to 4-factors

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.67	0.73	0.01	0.81	0.03	−0.11	Heywood case			
2	0.48	0.21	0.31	0.27	0.39	−0.07				
3	0.45	0.08	0.41	0.01	−0.08	0.57				
4	0.33	−0.13	0.48	−0.17	0.65	0.01				
5	0.57	−0.01	0.63	−0.00	0.23	0.45				
6	0.57	0.34	0.27	0.33	−0.01	0.30				
7	0.46	0.20	0.29	0.20	0.25	0.10				
8	0.48	0.61	−0.08	0.58	−0.06	−0.01				
9	0.57	0.14	0.47	0.17	0.40	0.12				
10	0.53	0.55	0.03	0.55	−0.07	0.09				
11	0.47	0.41	0.11	0.42	0.07	0.04				
12	0.68	0.07	0.68	−0.05	0.12	0.76				
13	0.44	0.26	0.22	0.26	0.24	0.02				
14	0.48	0.37	0.15	0.34	0.02	0.17				
Correlation between factors										
Factor										
II		0.71		0.58						
III				0.70	0.53					
IV							Heywood case			

(Items 2, 4, 5, 7, 9, 12, and 13). For the 3-factor solution, seven items loaded on the first factor (Items 1, 2, 4, 8, 10, 11, and 14), five items on the second factor (Items 2, 3, 5, 6, and 12), and six items (Items 2, 4, 5, 7, 9, and 13) on the third factor. For the 4-factor solution, four items loaded on the first factor (Items 3, 5, 6, and 12), six items loaded on the second factor (Items 1, 2, 4, 8, 10, and 11), four items loaded on the third factor (Items 4, 7, 13, and 14), and four items (Items 2, 4, 5, and 9) on the fourth factor. Results in Table 12 yielded complex patterns similar to those in Table 11. The lower part of Tables 11 and 12 contains the correlations between promax factors for SAS and SAS/Tetrachoric, respectively.

4 Discussion

The primary purpose of this study was to compare two popular software packages, Mplus and TESTFACT, on their capabilities for checking dimensionality in multiple-choice tests. Consistent with previous research (Stone and Yeh 2006; Tate 2003), analyses of the guessing condition indicated that TESTFACT was more accurate at detecting the simulated number of dimensions than Mplus. Both TESTFACT and Mplus, however, failed to detect unidimensionality, when no guessing was simulated. TESTFACT detected unidimensionality, when guessing was simulated, but Mplus overestimated the number of factors, because it has no

Table 9 Mplus WLS factor loadings for models with 1- to 4-factors

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.68	0.72	0.05	Heywood case due to over-factoring						
2	0.53	0.27	0.28							
3	0.49	0.17	0.37							
4	0.36	−0.17	0.54							
5	0.61	0.07	0.57							
6	0.60	0.46	0.19							
7	0.50	0.09	0.44							
8	0.50	0.67	−0.10							
9	0.61	0.05	0.60							
10	0.54	0.55	0.04							
11	0.52	0.34	0.24							
12	0.72	0.10	0.68							
13	0.44	0.13	0.36							
14	0.50	0.35	0.21							
Correlation between factors										
Factor										
II		0.68								
III				Heywood case due to over-factoring						
IV										

option for handling guessing. With respect to the estimated number of dimensions, TESTFACT generally was more accurate than Mplus for both guessing and no guessing conditions. Mplus with WLSM using RMSR criteria tended to over estimate the number of dimensions when guessing was simulated. Similarly, Mplus performed less well when factors were correlated. TESTFACT performed similarly with correlated and uncorrelated factors.

In the real data analysis example, both TESTFACT and Mplus yielded similar results. Although the true underlying factor structure of the data was unknown, the mathematics test itself was designed to be unidimensional. According to results for both algorithms, a 1-factor solution appeared to be a reasonable choice for the data. In addition, results for SAS were consistent with those for TESTFACT and Mplus. The results for TESTFACT were consistent with previous research by Stone and Yeh (2006) and Tate (2003).

A second purpose of this study was to compare different indices used for detection of dimensionality for dichotomous items. The main finding was that the proportion of variance was not a good indication of dimensionality. The RMSR reduction in Mplus, recommended by Tate (2003), also did not appear to work well, whereas the chi-square test was successful in most conditions. The RMSR reduction criterion of 10% (Tate 2003) was more sensitive, overestimating the simulated dimensionality under most conditions. RMSR reduction yielded a 3-factor solution for the real data. The RMSR criterion of < 0.08 proposed by Hu and Bentler (1999)

Table 10 Mplus/ULS factor loadings for the models of one factor, two factors, three factors, and four factors

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.67	0.71	−0.01	0.83	0.00	−0.08	0.87	−0.11	−0.16	0.11
2	0.48	0.19	0.35	0.29	0.44	−0.14	0.12	0.03	0.03	0.69
3	0.44	0.32	0.15	−0.04	−0.07	0.61	0.10	−0.05	0.49	−0.11
4	0.33	−0.23	0.66	−0.20	0.61	0.06	−0.14	0.91	−0.08	0.03
5	0.57	0.22	0.42	−0.00	0.29	0.41	0.04	0.07	0.17	0.14
6	0.56	0.48	0.12	0.26	−0.02	0.39	0.46	0.05	0.19	−0.14
7	0.46	0.22	0.29	0.14	0.23	0.18	0.31	0.20	0.08	−0.07
8	0.48	0.57	−0.06	0.52	−0.03	0.04	0.57	−0.05	−0.05	0.03
9	0.57	0.19	0.47	0.19	0.44	0.08	0.22	0.13	0.19	0.22
10	0.53	0.60	−0.04	0.49	−0.05	0.15	0.57	−0.07	0.03	0.01
11	0.48	0.44	0.08	0.42	0.09	0.03	0.42	−0.05	0.14	0.14
12	0.67	0.37	0.37	−0.05	0.17	0.69	−0.10	−0.07	0.93	0.05
13	0.44	0.25	0.24	0.23	0.22	0.07	0.38	0.18	−0.01	−0.04
14	0.48	0.43	0.08	0.28	0.03	0.22	0.38	0.01	0.14	−0.04
Correlation between factors										
Factor										
II		0.64		0.57			0.48			
III				0.68	0.52		0.75	0.51		
IV							0.39	0.34	0.32	

seemed to work well, given the conditions simulated, but the criterion of < 0.05 recommended by Stone and Yeh (2006) suggested a 2-factor solution. Results for RMSEA using the criteria from both Yeh and Stone and Tate yielded a 1-factor solution. Overall results provided no clear-cut answer to the practical question of which method should be used in all circumstances. Results from the chi-square test in TESTFACT were similar to previous research by Stone and Yeh and by Tate whereas results for the RMSR reduction index with Mplus were not consistent with these studies.

Although Mplus is easy to use and provides more fit indices, one suggestion is that using the chi-square test in TESTFACT might be more useful based on the higher number of correct identifications. Additionally, it would seem wise at this point to use a combination of these indices rather than relying on a single one. Substantive theory also should be considered as a meaningful explanation is more important than simply fitting a statistical model (Cudeck 2000). Finally, factor loadings should also be examined when determining the number of factors.

Table 11 SAS/Phi factor loadings for 1- to 4-factor solutions

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.52	0.53	0.07	0.59	0.06	0.06	0.12	0.57	−0.07	0.22
2	0.36	0.03	0.46	0.34	−0.32	0.50	−0.16	0.34	−0.16	0.70
3	0.40	0.42	0.04	−0.09	0.73	−0.05	0.78	−0.09	−0.09	−0.11
4	0.30	−0.30	0.79	−0.30	0.03	0.78	−0.01	−0.35	0.43	0.53
5	0.51	0.24	0.41	0.01	0.39	0.35	0.51	0.00	−0.08	0.36
6	0.50	0.50	0.09	0.17	0.52	0.02	0.54	0.16	0.03	−0.01
7	0.44	0.21	0.35	0.10	0.23	0.31	0.11	0.04	0.47	0.11
8	0.44	0.56	−0.08	0.65	0.01	−0.08	−0.09	0.60	0.25	−0.05
9	0.52	0.15	0.53	0.12	0.14	0.51	0.21	0.10	0.08	0.50
10	0.49	0.58	−0.03	0.47	0.25	−0.06	0.21	0.44	0.14	−0.04
11	0.43	0.42	0.08	0.54	−0.04	0.08	−0.02	0.51	0.04	0.20
12	0.61	0.44	0.30	0.11	0.54	0.23	0.53	0.08	0.18	0.13
13	0.41	0.21	0.32	0.19	0.12	0.30	−0.11	0.11	0.71	0.02
14	0.44	0.46	0.04	0.33	0.27	0.01	0.07	0.27	0.52	−0.18
Correlation between factors										
Factor										
II		0.31		0.25			0.25			
III				0.27	0.17		0.28	0.11		
IV							0.16	0.07	0.17	

Table 12 SAS/Tetrachoric factor loadings for the 1- to 4-factor solutions

	One factor	Two factors		Three factors			Four factors			
Item	I	I	II	I	II	III	I	II	III	IV
1	0.70	0.68	0.11	0.66	0.12	0.08	0.61	0.13	0.26	−0.01
2	0.53	0.13	0.53	0.37	−0.21	0.55	0.30	−0.18	0.80	−0.10
3	0.49	0.50	0.04	−0.10	0.86	−0.11	−0.07	0.87	−0.11	−0.07
4	0.37	−0.33	0.89	−0.28	0.02	0.86	−0.36	−0.02	0.46	0.60
5	0.61	0.29	0.45	0.00	0.49	0.35	−0.02	0.51	0.44	−0.06
6	0.61	0.59	0.09	0.20	0.60	−0.02	0.19	0.59	−0.03	0.08
7	0.51	0.24	0.39	0.15	0.20	0.34	0.09	0.15	0.05	0.50
8	0.53	0.65	−0.08	0.76	0.07	−0.07	0.70	−0.09	0.00	0.16
9	0.62	0.18	0.59	0.14	−0.16	0.54	0.08	0.17	0.58	0.09
10	0.58	0.68	−0.05	0.57	0.23	−0.09	0.54	0.22	0.06	−0.02
11	0.53	0.50	0.10	0.60	−0.05	0.10	0.54	−0.04	0.28	−0.02
12	0.70	0.50	0.32	0.14	0.59	0.21	0.11	0.58	0.17	0.12
13	0.50	0.24	0.36	0.32	−0.02	0.36	0.23	−0.11	−0.08	0.76
14	0.53	0.54	0.04	0.47	0.18	0.01	0.41	0.13	−0.18	0.40
Correlation between factors										
Factor										
II		0.43		0.42			0.38			
III				0.36	0.33		0.20	0.31		
IV							0.22	0.33	0.30	

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113–127.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37–51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

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). Newbury Park, CA: Sage.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

Cudeck, R. (2000). Exploratory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 265–296). San Diego, CA: Academic.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267–269.

Han, K. T. (2006). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*, 457–459.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.

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

Hulin, C. L., Drasgow, F., & Parsons, L. K. (1983). *Item response theory*. Homewood, IL: Dow-Jones-Irwin.

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

Kim, H. R. (1994). *New techniques for the dimensionality assessment of standardized test data* (Unpublished doctoral dissertation). University of Illinois, Urbana–Champaign.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457–477.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mentaltest scores*. Reading, MA: Addison-Wesley.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*, 379–396.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG-MG 3: Item analysis and test scoring with binary logistic models* [Computer software]. Chicago, IL: Scientific Software International.

Mulaik, S. A. (2009). *The foundations of factor analysis*. New York: CRC.

Muthén, B. (1977). *Statistical methodology for structural equation models involving latent variables with dichotomous indicators* (Unpublished doctoral dissertation). Department of Statistics, University of Uppsala, Uppsala.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.

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

Nandakumar, R. (1994). Assessing the dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*, 17–35.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

Nandakumar, R., & Yu, F. (1996). Empirical validation of DIMTEST on nonnormal ability distributions. *Journal of Educational Measurement, 33*, 355–368.

Parry, C. D., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15*, 35–46.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207–230.

Stone, C. A., & Yeh, C.-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the multistate bar examination. *Educational and Psychological Measurement, 66*, 193–214.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159–203.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, D. T., Wood, R., & Gibbons, R. (2003). *TESTFACT: Test scoring, item statistics, and item factor analysis* [Computer program]. Chicago, IL: Scientific Software International.

Yeh, C.-C. (2007). The effect of guessing on assessing dimensionality in multiple-choice tests: A Monte Carlo study with application. Unpublished dissertation. University of Pittsburg.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*, 129–152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 213–249.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). *BILOG-MG for windows* [Computer software]. Lincolnwood, IL: Scientific Software International.