BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of International Education and Practice
http://ojs.bilpublishing.com/index.php/jiep/

ARTICLE

Guessing and Nature of Multidimensionality Matter: A Cautionary Note on the Use of Fit Indices to Assess Unidimensionality of Binary Data

Yong Luo[1]*   Dimiter M. Dimitrov[1,2]

1. National Center for Assessment in Higher Education, Riyadh, 12395, Saudi Arabia

2. George Mason University, Fairfax, Virginia, 22030, USA

ARTICLE INFO

Article history:
Received: 26 November 2018
Accepted: 13 December 2018
Published: 31 December 2018

Keywords:
Cut-off value
Model fit index
Guessing
Compensatory model
Noncompensatory model
Unidimensionality

ABSTRACT

Use of cutoff values for model fit indices to assess dimensionality of binary data representing scores on multiple-choice items is a popular approach among researchers and practitioners, and the commonly used cutoff values are based on simulation studies that used as the generating model factor analysis models, which are compensatory models without modeling guessing. Consequently, it remains unknown how those cutoff values for model fit indices would perform when (a) guessing exists in data, and (b) data follow a noncompensatory multidimensional structure. In this paper, we conducted a comprehensive simulation study to investigate how guessing affected the statistical power of commonly used cutoff values for RMSEA, CFA, and TLI (RMSEA > 0.05; CFA < 0.95; TLI < 0.95) to detect violation of unidimensionality of binary data with both compensatory and noncompensatory models. The results indicated that when data were generated with compensatory models, increase of guessing values resulted in the systematic decrease of the power of RMSEA, CFA, and TLI to detect multidimensionality and in some conditions, a small increase of guessing value can result in dramatic decrease of their statistical power. It was also found that when data were generated with noncompensatory models, use of cutoff values of RMSEA, CFA, and TLI for unidimensionality assessment had unacceptably low statistical power, and while change of guessing magnitude could considerably change their statistical power, such changes were not systematic as in the compensatory models.

1. Introduction

As one of the pivotal assumptions of item response theory (IRT), unidimensionality stipulates that item responses are driven by a single underlying latent variable. Numerous studies have shown that violations of the assumption of unidimensionality can lead to serious psychometric consequences such as biased item parameter estimates, equating errors, and misclassification

**Corresponding Author:*
Yong Luo,
National Center for Assessment in Higher Education, Riyadh, 12395, *Saudi Arabia*
Email: jackyluoyong@gmail.com

of examinees[1][2][3][4]. Aside from its critical importance in the valid application of IRT, the other two reasons that unidimensionality is of particular interest to researchers and practitioners, as summarized by Stout, are that the primary ability a test intends to measure should not be contaminated by other abilities and a unified latent variable is the precondition of meaningful comparison of individuals[5].

Probably due to its statistical importance and conceptual attractiveness, unidimensionality has received extensive attention in the psychometric literature[6][7][8][9] and a plethora of methods have been developed to assess unidimensionality[10][11][12][13][14][15][16][17][18]. Among them, factor analytic methods, "an important tool"[19] for dimensionality assessment of IRT models, are especially attractive to many structural equation modeling (SEM)[20] researchers since such methods are housed in the familiar SEM framework and require no additional IRT-based software programs other than common SEM ones. In this paper we focus on one such factor analytic method, namely the use of cutoff values of fit indices within SEM framework to assess unidimensionality with binary data. We assume that the binary data are item scores of multiple choice items and consequently, guessing is expected to exist within the data.

The use of fit indices for unidimensionality assessment makes intuitive sense in light of the mathematical equivalence between factor models with categorical variables and IRT[21][22][23][24][25]: if fit indices can be used to assess whether a one-factor model fits data satisfactorily, why cannot they be used to test whether a two-parameter normal ogive model, the IRT analog of the one-factor model with categorical variables, represents the data well? If these fit indices indicate good model fit based on some well-established cutoff values, it is concluded that the unidimensionality assumption is not violated. Despite its logical intuitiveness, this fit-index-based approach with binary data makes two implicit assumptions: first, those well-established cutoff values of model fit indices are applicable to cases of unidimensionality assessment with binary data; second, such cutoff values are robust to the existence of guessing.

The cutoff criteria for model fit indices proposed by Hu and Bentler[26] have been hugely popular among researchers interested in assessing the latent structure of their data. Although unidimensionality assessment is not included as a condition in their simulation study, these cutoff values have been nevertheless used by SEM researchers for unidimensionality assessment[27]. Despite their tremendous popularity, researchers have raised concerns about indiscriminate use of those indices[28][29][30][31][32], on the grounds that these fit indices are sensitive to different type of model misspecifications and consequently, establishment of cutoff values for model fit indices that are universally applicable is, if not impossible, very difficult. As pointed out by Huggins-Manley and Han[33], Hu and Bentler's simulation study[26], as well as other similar simulation studies[34][35] that address establishment of cutoff values, focus on misspecifications of factor loadings and/or latent variable correlations in multidimensional models. It remains unclear how model fit indices would perform when a unidimensional model is imposed upon data generated with multidimensional models.

While the impact of model misspecification type upon performances of model fit indices has been extensively studied in the literature, measurement quality, which can tremendously change the statistical behavior of model fit indices, fails to receive attention from researchers and practitioners with a few exceptions[36][37]. As demonstrated by McNeish, An, and Hancock, change of measurement quality (operationalized through the change of magnitude of standardized factor loadings) can result in drastically different distributions of model fit indices[36] and consequently, the model fit indices are meaningless without taking into consideration the standardized factor loadings. If measurement quality is conceptualized as the strength of the relation between indicators and the target latent variable, we argue that measurement quality can deteriorate with either the decrease of magnitude of standardized factor loadings, or with the introduction of guessing and the increase of guessing magnitude. As measurement quality can also be affected by the existence of guessing, we believe that the expected ubiquitous existence of guessing due to the common use of multiple-choice questions in educational setting, is of huge relevance when it comes to the dimensionality assessment of binary data. Since the common cutoff values of model fit indices were proposed in the factor analysis framework and based on data generating models that do not incorporate guessing, we believe their performances will be negatively affected by guessing based on previous studies that investigated the effect of guessing in factor analysis models[38][39][40]. To date, there have been no simulation studies that systematically investigate the guessing effect upon the performances of cutoff values of model fit indices.

Aside from guessing, another factor that has not received sufficient attention in the literature regarding the use of model fit indices for model assessment is the multidimensional nature of data (whether data exhibits compensatory or noncompensatory multidimensionality).

DOI: https://doi.org/10.30564/jiep.v1i1.356

While educational and psychological researchers usually focus on compensatory multidimensionality, noncompensatory multidimensionality occurs occasionally in some educational test items that require multiple skills and inadequacy in one skill cannot be compensated by other skills[41]. In the aforementioned simulation studies that dealt with categorical indicators[34][35], the researchers generated data based on factor analysis models that are equivalent to the compensatory IRT models, and it remains unknown how model fit indices, which are based on the factor analysis framework and therefore compensatory IRT models, will perform with data generated with noncompensatory IRT models. Previous studies have indicated that common dimensionality assessment methods that perform well with compensatory models may fall short with noncompensatory ones[42][43]. In addition, it is equally unclear whether and how guessing systematically affect the performances of cutoff values of model fit indices with data following noncompensatory structures.

The purpose of this study is to systematically investigate the impact of guessing upon the statistical power of cutoff criteria of model fit indices to refute unidimensionality when data are generated with both compensatory and noncompensatory IRT models. Specifically, since the standardized root mean square residual (SRMR) is not recommended for dichotomous items[35], we focus on the comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA), which are the also the fit indices reported in the popular latent variable modeling software Mplus[44] with the default weighted least squares mean- and variance-adjusted[45] estimator for categorical variables. Due to the existence of a large body of literature that provides excellent review of model fit indices, we do not review CFI, TLI, and RMSEA in this paper but refer interested readers to [46] for a comprehensive introduction[46].

The remainder of this paper is organized as follows. First, we review some influential simulation studies in which the commonly used cutoff values of fit indices were either established or validated. Next, we review previous studies that have investigated the effect of guessing in the factor analysis framework, followed by a review of studies dealing with dimensionality assessment of data generated with noncompensatory IRT models. In the method section we present two simulation studies conducted to investigate how commonly used cutoff values of CFI, TLI, and RMSEA perform with binary data generated with both compensatory and noncompensatory IRT models. We conclude this paper with conclusions and discussions, as well as some advice for applied researchers and practitioners who are interested in using model fit indices for unidimensionality assessment.

2. Literature Review

2.1 Simulation Studies on Model Fit Indices

In their highly influential study, Hu and Bentler[26] generated continuous data based on two model types (complex and simple), both of which assumed fifteen observed variables and three factors. They fixed the factor variances to 1.0 and the correlation between factors to 0.5, 0.4, and 0.3. For the simple model type, five variables load on each factor and there are no cross loadings; for the complex model type, one out of five variables that loads on one specific factor has a cross loading with another factor. They created seven data generation conditions by manipulating factors such as normality and correlation between factors and errors. To create scenarios of model misspecification, they either constrained the between factor correlation or some cross factor loadings to be zero. They created 200 replicated datasets within each condition and established the following cutoff values for model fit indices based on their simulation results: RMSEA < 0.06, CFI > 0.95, TLI > 0.95.

As Yu[35] decisively pointed out, Hu and Bentler's simulation study was based on maximum likelihood (ML) estimation method with continuous data, which is not suitable for categorical data usually estimated with robust diagonally weighted least square (DWLS) estimation methods. Since DWLS and ML use different fit functions and hence the chi-square values are different, the behavior of chi-square-based model fit indices might be different across different estimation methods and the cutoff values proposed out of simulation studies using one estimation method should not be generalized to other methods. She followed a similar simulation study design as Hu and Bentler's but focused on categorical outcomes and WLSMV estimator, the robust DWLS estimation method implemented in Mplus. She found that SRMR is not a good mode fit index for binary outcomes. TLI > 0.95 seems to perform satisfactorily when the sample size is greater than 250; for CFI, she found that CFI > 0.96 seems to perform better than CFI > 0.95; with RMSEA, she found that RMSEA < 0.05 outperforms RMSEA < 0.06.

Driven by the realization that Hu and Bentler's study was based on ML estimator with continuous outcomes and consequently, their proposed cutoff values of model fit indices might not generalize to cases of categorical outcomes estimated with DWLS estimator, Nye and Drasgow [34] conducted a simulation study to investigate how the cutoff values proposed by Hu and Bentler performed with binary data estimated with the DWLS estimator imple-

DOI: https://doi.org/10.30564/jiep.v1i1.356

mented in LISREL 8.71[47]. Specifically, they simulated data from a two-factor model (the between factor correlation was fixed at 0.3) with 15 variables that either load on one factor or both factors; they manipulated the sample size to have three levels (400, 800, and 1600); they further manipulated the underlying distribution to have three levels (multivariate normal, moderately skewed, and severely skewed); they created misspecification scenarios by either constraining some factor loadings to be zero or both some factor loadings and between factor correlation to zero. They found that the commonly used cutoff values did not have enough power with DWLS estimator and more stringent values need to be used, and they concluded that simple cutoff values for model fit indices would not work since model fit can only be evaluated effectively in combination with the specific data.

While both Yu[35] and Nye and Drasgow[34] investigated the performances of these model fit indices with DWLS estimator and binary outcomes, neither included guessing in their data generation process and as a result, the effect of guessing upon the performance of model fit indices remains unclear. In addition, both simulations studies simulated data base on some factor analysis models and consequently, it is unknown how these model fit indices would perform with data generated with noncompensatory IRT models. Another difference is the nature of model misspecification: in both studies model misspecification takes the form of erroneous between-factor correlations or factor loadings, which is different from imposing a unidimensional structure upon data with multidimensional nature-the misspecification scenario we focus on in the current study.

2.2 Guessing in the Factor Analysis Framework

The effect of guessing is rarely investigated in the factor analysis framework. Among the few studies available, Carrol[48] found that when guessing was modeled, the tetrachoric correlations were corrected and hence stronger relation among the indicators were expected. In other words, if guessing was not modeled, the tetrachoric correlations would be attenuated relative to the true values. Considering that DWLS estimator is based on the estimation of tetrachoric correlations, such attenuation effects are expected to exist with data generated with guessing in the factor analysis framework.

Subsequent studies corroborate Carrol's findings. Tate[39] conducted a simulation study to investigate how guess affects decisions regarding dimensionality and parameter recovery in both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). He found that with guessing parameter fixed to be 0.2 in various multidimensional models, both EFA and CFA based on tetrachoric correlations uncorrected for guessing resulted in lower power to identify the true dimensionality. In terms of parameter recovery, the item thresholds and factor loadings in both EFA and CFA showed downward biases, which range in magnitude from 0.1 to 0.8 for item thresholds, and from 0.1 to 0.5 for factor loadings. Such biases were exacerbated with extreme item difficulties and discriminations.

Stone and Yeh[38] also investigated the guessing effect in EFA implemented in TESTFACT[49] using the Multistate Bar Examination data. They found that when guessing was modeled, the first eigenvalue of exploratory factor analysis (EFA) become larger and more items loaded substantially on factors. In addition, the average tetrachoric correlation increased from 0.07 to 0.11. Yeh conducted a large-scale simulation study to investigate guessing effect in EFA implemented in Mplus and TESTFACT[40]. Using a fixed sample size of 2,000 and a test length of 60 items, she systematically manipulated the number of dimensions, item discrimination parameters, between dimension correlations, and guessing magnitude to create various simulation conditions. Within each condition, 100 datasets were generated and estimated with EFA procedures implemented in both Mplus and TESTFACT. She found that TESTFACT, which allows the users to provide guessing values, outperformed Mplus in most simulation conditions regarding the ability to confirm the correct dimensionality.

2.3 Noncompensatory MIRT Model

Multidimensional IRT (MIRT)[50] models consist of compensatory and noncompensatory cases. Whereas the compensatory MIRT model is mathematically equivalent to a multi-factor model with categorical indicators (which is also known as a nonlinear factor model), the noncompensatory MIRT model does not have an equivalent counterpart in the factor analysis framework. The mathematical equation for a three parameter logistic noncompensatory MIRT model[51] takes the following form:

$$P(U_{ij}=1\,|\,\mathbf{\grave{e}}_{\mathbf{i}},\mathbf{a}_{\mathbf{j}},\mathbf{b}_{\mathbf{j}},c_j)=c_j+(1-c_j)\prod_{d=1}^{D}\frac{1}{1+\exp(-a_{jd}(\theta_{id}-b_{jd}))} \tag{1}$$

where U_{ij} is the response of examinee i to item j, D is the number of dimensions, θ_{id} is the ability of examinee i on dimension d, a_{jd} and b_{jd} are the discrimination parameter and difficulty parameter of item j on dimension d, and c_j is the lower asymptote of item j . As indicated by the Pi notation, the noncompensatory MIRT model assumes that inadequacy in one dimension cannot be completely compensated by adequacy in another dimension.

Comparing to a large number of methodological studies investigating the performance of various dimensionality assessment techniques with data generated using the compensatory model, there are considerably fewer ones focusing on data with noncompensatory structures. Among those few, Hattie, Krakowski, Rogers, and Swaminathan[42] investigated the performance of Stout's index of essential unidimensionality implemented in the DIMSEST software[5] with data generated using both the compensatory and noncompensatory model. They found that the DIMTEST procedure performed poorly when the data was generated with the latter model, and attributed its poor performance to the problems in estimating tetrachoric correlations. A more recent study is a simulation study conducted by Svetina[43], in which she generated data using the 2PL noncompensatory MIRT model and compared the performance of two methods (exploratory vs. cross validated) based on DETECT (Dimensionality Evaluation To Enumerate Contributing Traits)[52][18][53] and three methods[54][55][39] based on NOHARM (Normal Ogive Harmonic Analysis Robust Method)[56]. Having found that the performances of those methods can only be considered acceptable in a small number of conditions, she suggested that further studies be conducted before consideration of applying those methods to data suspected of having noncompensatory structure. It should be noted that in the above two studies, the magnitude of guessing was not systematically investigated: in the first study Hattie and et al. manipulated the lower asymptote to be either 0 or 0.15; Svetina only considered a 2PL model in which the guessing is assumed not to exist.

3. Methodology

3.1 Outcome Variable

A one-factor model was fit to each generated data set using Mplus with WLSMV estimator and RMSEA, CFI, and TLI were computed. For each of these three model fit indices, we computed its empirical power rate, which is the number of times that the model fit is considered poor divided by the number of replications within a simulation condition, using the following cutoff values: RMSEA > 0.05, CFI < 0.95, TLI < 0.95. It should be noted that for the sake of simplicity, in the following sections we use terms such as the power of RMSEA to refer to the power of using the cutoff value RMSEA > 0.05, the power of CFI for the power of using the cutoff value CFI < 0.95, and the power of CFI for the power of using the cutoff value TLI < 0.95.

3.2 Study Design

In both the compensatory and noncompensatory cases, the number of dimensions were fixed to three and following factors were manipulated:

1) Sample size (500, 1,000, or 2,000)
2) Number of items (30 or 60)
3) Between dimension correlation (0, 0.3, 0.5, and 0.7)
4) Pseudo-guessing value (0, 0.1, 0.2, 0.3, and 0.4)

For both the compensatory and noncompensatory cases, we have a fully crossed design with 3*2*4*5=120 conditions. Within each condition, we generated 1,000 datasets based on the corresponding MIRT model.

3.3 Item Response Generation

For both compensatory and noncompensatory cases, latent abilities were generated from three-dimensional multivariate normal distributions with a mean vector of 0s and a variance vector of 1s, and different levels of between dimension correlation values as specified in the previous section. Another commonality between the compensatory and noncompensatory cases is the systematic manipulation of pseudo-guessing values. The two cases differ regarding the generated item discrimination and difficulty parameter values.

For the compensatory case, the following three-dimensional three-parameter logistic (3PL) item response theory (IRT) model was used to generate item responses:

$$P(U_{ij}=1\,|\,\boldsymbol{\theta}_{\mathbf{i}},\mathbf{a}_{\mathbf{j}},d_j,c_j)=c_j+(1-c_j)\frac{1}{1+\exp(-\mathbf{a}_{\mathbf{j}}\boldsymbol{\theta}_i+d_j))} \qquad (2)$$

where a_j is a vector of item *j*'s discrimination parameters on the three dimensions, θ_i is a vector of examinee i's scores on the three dimensions, d_j is item *j*'s multidimensional difficulty parameter, and c_j is the pseudo-guessing parameter. For item discrimination and difficulty parameters a_j and d_j , we used values provided by Reckase[50] as a realistic approximation of tests of simple structure. Since there are only 30 sets of item parameters, we generated another 30 similar items: for a_j , we added a random value drawn from N (0, 0.02) to each of the original 30 sets of discrimination parameters; for d_j , we added a random value drawn from *N* (0, 0.1) to each of the original 30 difficulty parameters. When the number of items is 30, only the first 30 sets of item discrimination and difficulty parameters were used for item response generation; when the number of items is 60, all 60 sets of item discrimination and difficulty parameters were used.

For the noncompensatory case, we followed the same item generating scheme adopted by Svetina[43]: item difficulty parameters were generated to fall in the range of -1.5 to 1.5 with an increment of 0.75, and item discrimination parameters on the dominant dimension were generated to range from 0.8 to 1.6 with an increment of 0.2, while on the remaining two dimensions they were generated to be 0.2 smaller than their counterparts on the dominant dimension. Items responses were generated based on the model specified in Equation 1.

DOI: https://doi.org/10.30564/jiep.v1i1.356

4. Results

4.1 Compensatory Model

Table 1 lists the power rates of RMSEA to correctly reject unidimensionality across different simulation conditions; those of CFI and TLI appear in Tables 2-3. Specifically, the value within a cell indicates the number of times to reject unidimensionality divided by 1,000, when using the cutoff value for a given model fit index for data generated under a certain simulation condition. For example, the value 0.022 on the second row of Table 1 means that when sample size was fixed to 500 students and test length to 30 items, by applying RMSEA > 0.05 only 22 datasets were correctly identified as multidimensional out of the 1,000 datasets generated based on the 3PL MIRT model in Equation 2 with the between-dimension correlation equal to 0.5 (denoted as C3 in the table) and the pseudo-guessing parameter equal to 0.1 (denoted as G2). Similarly, the value 0.033 on the third row of Table 2 means that when sample size was fixed to 2,000 students and test length to 30 items, by applying CFI < 0.95 only 33 datasets were correctly identified as multidimensional out of the 1,000 datasets generated based on a 3PL MIRT model with the between-dimension correlation equal to 0.7 (denoted as C4 in the table) and the pseudo-guessing parameter equal to 0.2 (denoted as G3). As can be seen, a common pattern for the three model fit indices is that their power decreases with the decrease of sample size and the increase of guessing magnitude, between-dimension correlation, and test length.

Table 1. Power of RMSEA to Reject Unidimensionality in Compensatory Models

	SS = 500				SS = 1,000				SS = 2,000			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
30 Items												
G1	1	1	0.945	0	1	1	1	0	1	1	1	0
G2	1	0.986	0.022	0	1	1	0.015	0	1	1	0.049	0
G3	1	0.051	0	0	1	0.099	0	0	1	0.306	0	0
G4	0.284	0	0	0	0.707	0	0	0	0.949	0	0	0
G5	0	0	0	0	0	0	0	0	0	0	0	0
60 Items												
G1	1	1	0.013	0	1	1	0.198	0	1	1	0.941	0
G2	1	0.022	0	0	1	0.491	0	0	1	1	0	0
G3	0.305	0	0	0	1	0	0	0	1	0	0	0
G4	0	0	0	0	0	0	0	0	0	0	0	0
G5	0	0	0	0	0	0	0	0	0	0	0	0

Note. C1-C4 represent the between-dimension correlation value (C1=0, C2=0.3, C3=0.5, C4=0.7); G1-G5 refer to the pseudo-guessing parameter value (G1=0, G2=0.1, G3=0.2, G4=0.3, G5=0.5).

Table 2. Power of CFI to Reject Unidimensionality in Compensatory Models

	SS = 500				SS = 1,000				SS = 2,000			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
30 Items												
G1	1	1	0.986	0.04	1	1	1	0.022	1	1	1	0.023
G2	1	1	0.911	0.03	1	1	0.983	0.03	1	1	1	0.029
G3	1	1	0.669	0.036	1	1	0.805	0.028	1	1	0.985	0.033
G4	1	1	0.472	0.034	1	1	0.467	0.025	1	1	0.676	0.028
G5	1	1	0.476	0.028	1	1	0.286	0.026	1	1	0.283	0.031
60 Items												
G1	1	1	0.955	0.025	1	1	1	0.03	1	1	1	0.024
G2	1	1	0.678	0.029	1	1	0.982	0.026	1	1	1	0.024
G3	1	1	0.281	0.033	1	1	0.684	0.027	1	1	0.983	0.024
G4	1	1	0.206	0.025	1	1	0.194	0.027	1	1	0.295	0.021
G5	1	1	0.114	0.024	1	1	0.061	0.024	1	1	0.024	0.024

Note. C1-C4 represent the between-dimension correlation value (C1=0, C2=0.3, C3=0.5, C4=0.7); G1-G5 refer to the pseudo-guessing parameter value (G1=0, G2=0.1, G3=0.2, G4=0.3, G5=0.5).

Table 3. Power of TLI to Reject Unidimensionality in Compensatory Models

	SS = 500				SS = 1,000				SS = 2,000			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
30 Items												
G1	1	1	1	0.04	1	1	1	0.022	1	1	1	0.023
G2	1	1	0.964	0.03	1	1	1	0.03	1	1	1	0.029
G3	1	1	0.797	0.036	1	1	0.961	0.028	1	1	1	0.033
G4	1	1	0.472	0.034	1	1	0.691	0.025	1	1	0.798	0.028
G5	1	1	0.476	0.024	1	1	0.458	0.026	1	1	0.472	0.031
60 Items												
G1	1	1	0.955	0.025	1	1	1	0.03	1	1	1	0.024
G2	1	1	0.784	0.029	1	1	1	0.026	1	1	1	0.024
G3	1	1	0.477	0.033	1	1	0.795	0.027	1	1	1	0.024
G4	1	1	0.206	0.025	1	1	0.298	0.027	1	1	0.492	0.021
G5	1	1	0.186	0.024	1	1	0.061	0.024	1	1	0.062	0.024

Note. C1-C4 represent the between-dimension correlation value (C1=0, C2=0.3, C3=0.5, C4=0.7); G1-G5 refer to the pseudo-guessing parameter value (G1=0, G2=0.1, G3=0.2, G4=0.3, G5=0.5).

4.1.1 Without Guessing

Figure 1 shows how RMSEA, CFA, and TLI perform with the baseline conditions (no guessing) at various combinations of sample size, between-dimension correlation, and test length. Same as in tables 1-3, C1, C2, C3, and C4 on the horizontal axis represent the magnitude of between-dimension correlation; the values on the vertical axis, which range from zero to one, represent the statistical power of

DOI: https://doi.org/10.30564/jiep.v1i1.356

applying the cutoff value of a given model fit index under different simulation conditions. For example, the upper left panel displays the statistical power of applying RMSEA > 0.05 with data generated with thirty items, three sample sizes, four levels of between-dimension correlation, and without guessing.

As can be seen, when the correlation between the three dimensions is no greater than 0.5, all three indices have satisfactory power (greater than 0.9) to detect multidimensionality regardless of sample size and test length, with the exception of RMSEA having low power (lower than 0.2) when the sample size is either 500 or 1000 and the test length is 60 items. When the between-dimension correlation increases to 0.7, RMSEA, CFI, and TLI have extremely low power (lower than 0.1) regardless of sample size and test length. Sample size seems to have no considerable effect when the correlation between dimensions is no greater than 0.3, and the power only increases marginally with sample size increase when the correlation between dimensions is 0.5. One interesting pattern is that when the between-dimension correlation is 0.5, the performance of RMSEA is inversely related to the test length: its power with sample sizes of 500 and 1000 drops precipitously when the test length increases from 30 items to 60 items.

4.1.2 Impact of Guessing

As can be seen in Table 1, the power of RMSEA decreases with the increase of guessing magnitude, correlation between dimensions, and test length. Its power increases marginally with the increase of sample size. Compared with RMSEA, neither CFI nor TLI is influenced by the increase of guessing magnitude when the correlation between dimensions is low: both have a power of one when the correlation is 0.3 or lower. When the correlation is 0.5 or higher, however, the power of both decreases considerably with the increase of guessing magnitude regardless of the sample size and test length.

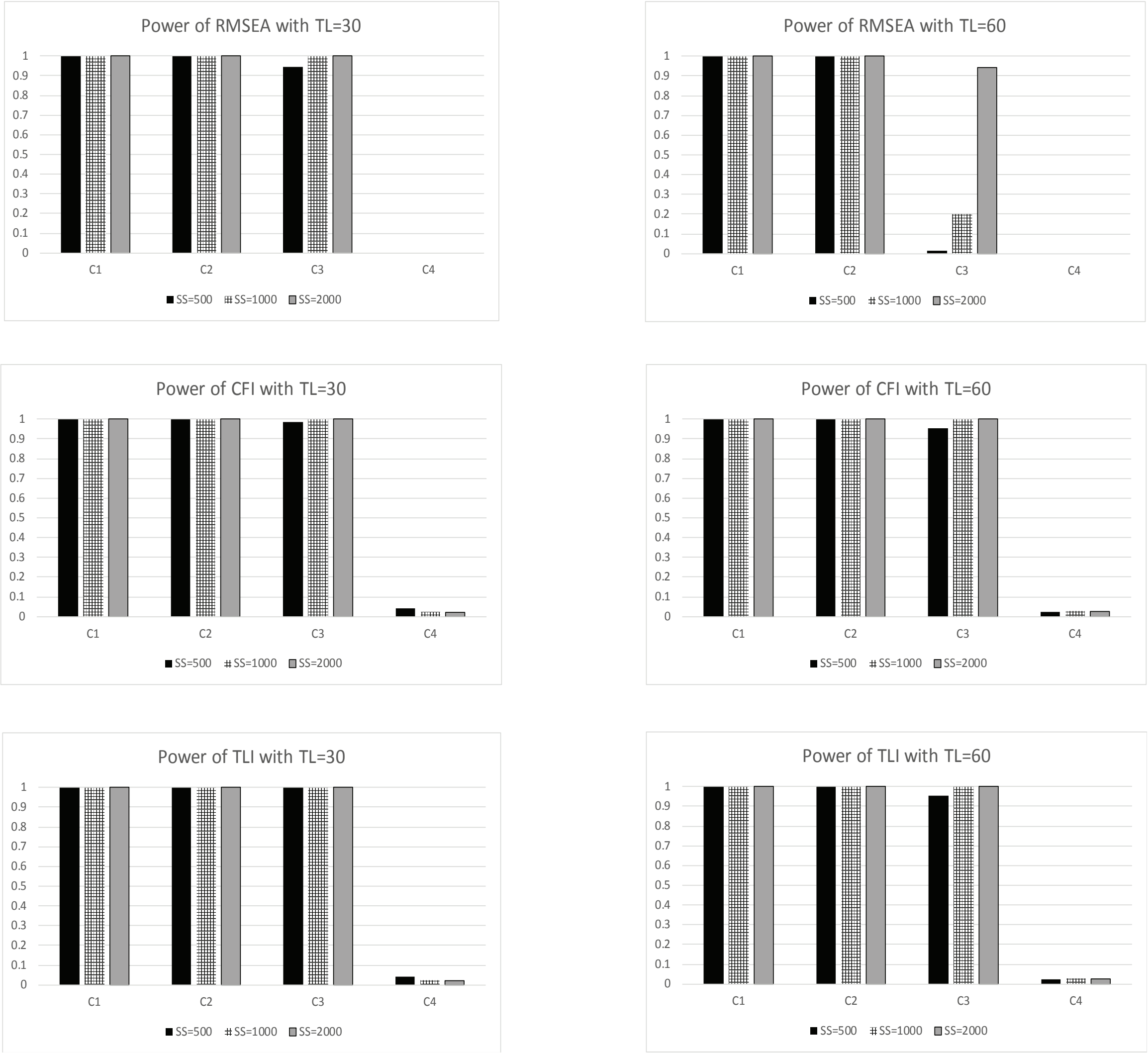


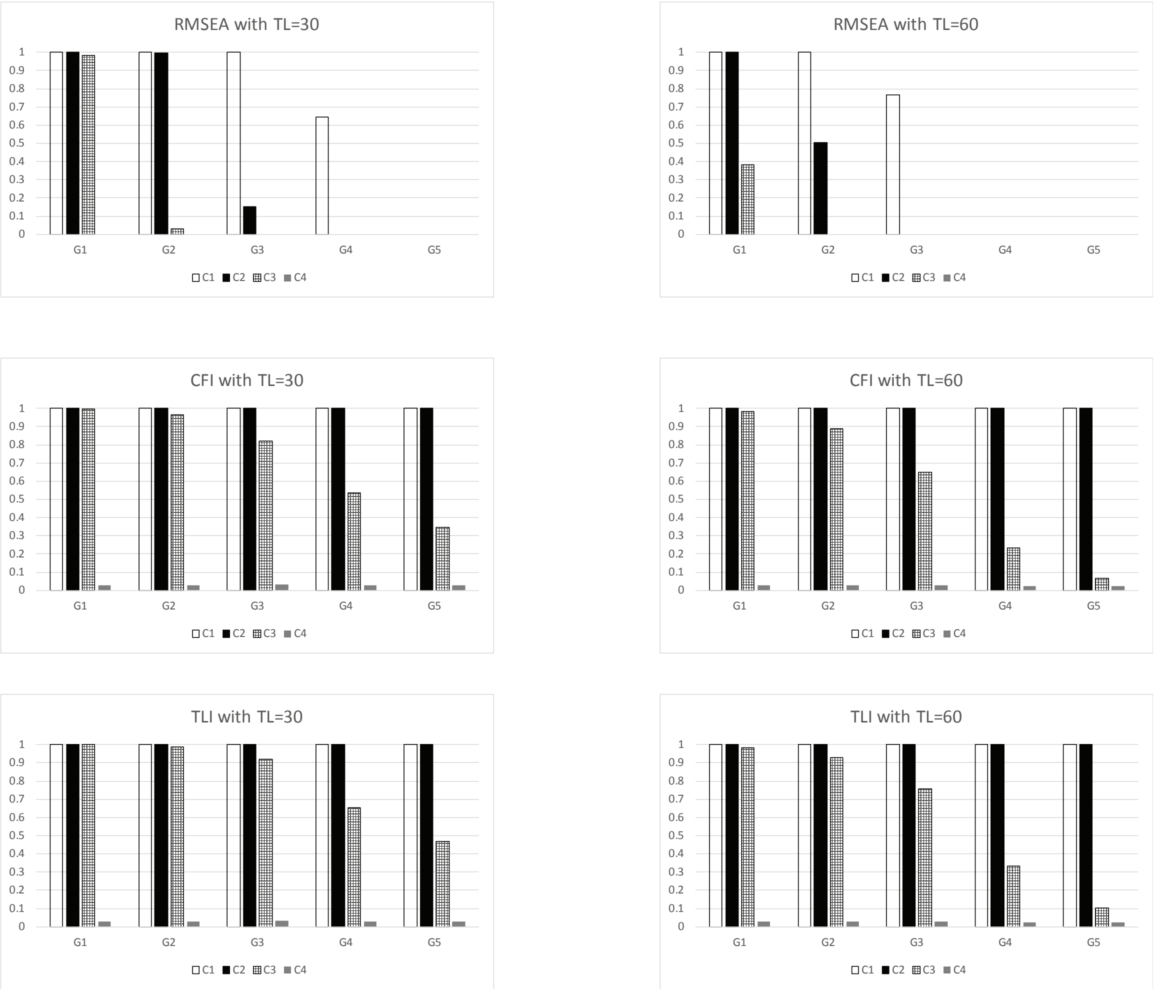
Figure 1. Power of RMSEA, CFI, and TLI with no guessing in compensatory models

DOI: https://doi.org/10.30564/jiep.v1i1.356

To further explore the guessing impact upon the statistical power of RMSEA, CFI, and TLI, in Figure 2 we plot the average power rate of each model fit index across three sample sizes. Same as in tables 1-3, G1, G2, G3, and G4 on the horizontal axis represent the guessing magnitude, and the vertical axis represents the average statistical power across three sample sizes of applying the cutoff value of a given model fit index. For example, the bottom right panel displays the average statistical power of applying TLI < 0.95 across three sample sizes with data generated with sixty items, four levels of between-dimension correlation, and four levels of guessing magnitude.

For RMSEA, when the guessing is no greater than 0.1, the statistical power is close to one with the between-dimension-correlation no greater than 0.3 and the test length equal to 30; if the test length increases to 60, however, the power of RMSEA drops to 0.5 when the between-dimension-correlation is 0.3. When the guessing is 0.2, the statistical power is one only with the between-dimension-correlation is zero and the test length is 30; if the test length increases to 60, however, the power of RMSEA drops to slightly lower than 0.8. When the guessing is 0.3 or 0.4, RMSEA has no satisfactory statistical power regardless of the between-dimension-correlation and test length.

CFI and TLI seem to be robust to the guessing when the between-dimension correlation is no greater than 0.3: their power remains invariably close to one regardless of the guessing value and sample size. When the between-dimension correlation is 0.5, guessing seems to have a systematic influence: the power of both CFI and TLI decreases with the increase of guessing value. Test length also plays a role regarding the power of CFI and TLI when the between-dimension correlation is 0.5: both indices have consistently lower power when the test length is 60 items than when it is 30 items. When the test length is 30 items and the between-dimension correlation is 0.5, both CFI and TLI have satisfactory power (greater than 0.8) when guessing is no greater than 0.2; when the test length is 60 items and the between-dimension correlation is 0.5, CFI


RMSEA with TL=30

RMSEA with TL=60

CFI with TL=30

CFI with TL=60

TLI with TL=30

TLI with TL=60

Figure 2. Power of RMSEA, CFI, and TLI with different guessing levels in compensatory models

and TLI have satisfactory power (greater than 0.8) only when guessing is no greater than 0.1. Similar to RMSEA, neither CFI nor TLI has enough statistical power when the between-dimension-correlation is 0.7 regardless of guessing value, sample size, and test length.

4.2 Noncompensatory Model

As the power of RMSEA to reject unidimensionality are invariably zero across all simulation conditions, we focus on the power of CFI and TLI in the concompensatory cases. Table 4 lists the power of CFI to reject unidimensionality across different simulation conditions, and those of TLI appear in Table 5. Different than the pattern observed in the preceding compensatory model that guessing magnitude, between-dimension correlation, sample size, and test length systematically affect the power of RMSEA, CFA, and TLI, here the only discernable pattern is that the power of these three model fit indices decreases with the increase of between-dimension correlation. In terms of guessing, although the power of RMSEA, CFA, and TLI change with the change of guessing magnitude, the change is not in a systematic pattern as in the compensatory cases. Another difference is that the power of RMSEA, CFA, and TLI observed here seem to be considerably lower than in the compensatory model.

4.2.1 Without Guessing

As RMSEA has no statistical power regardless of the guessing value, the between-dimension-correlation, and test length, we focus on CFI and TLI regarding their performances with the baseline condition (no guessing). Figure 3 plots the mean power rates of these two model fit indices across averaged across four between-dimension correlation values. Regardless of the test length, their statistical power becomes satisfactory (CFI has statistical power slightly lower than 0.8 when the test length is 30) only when the between-dimension-correlation value is zero. In contrast to what has been observed in the compensatory cases where statistical power of RMSEA, CFA, and TLI decreases with the increase of test length, with zero between-dimension correlation CFI and TLI have slightly higher statistical power when the test length is 60 items than when it is 30 items.

4.2.2 Impact of Guessing

As the power of RMSEA remains zero in all simulation conditions, it is not possible to evaluate the effect of guessing upon the performance of RMSEA. In this section we focus on how guessing affects the performances of CFI and TLI in the noncompensatory cases. As can be seen from Tables 4-5, when data were generated with a noncompensatory IRT model, while the power of CFA and TLI seem to decrease with the increase of between-dimension correlation value, there seems to be no discernable patterns regarding how their power change as a result of the change of guessing value, test length, or sample size.

To further explore the guessing impacts upon the statistical power of these two model fit indices, in Figure 4 we

Table 4. Power of CFI to Reject Unidimensionality in Noncompensatory Models

	SS = 500				SS = 1,000				SS = 2,000			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
30 Items												
G1	0.789	0.104	0	0	0.451	0	0	0	0.085	0	0	0
G2	0.482	0.297	0.032	0.054	0.292	0.024	0	0	0.206	0	0	0
G3	0.421	0.275	0.06	0.108	0.438	0.123	0	0	0.21	0	0	0
G4	0.414	0.407	0.194	0.205	0.427	0.203	0.076	0.079	0.27	0.068	0	0
G5	0.411	0.406	0.258	0.266	0.387	0.264	0.19	0.221	0.28	0.184	0.088	0
60 Items												
G1	0.989	0	0	0	0.502	0	0	0	0.189	0	0	0
G2	0.818	0.052	0.03	0	0.806	0.062	0	0	0.804	0.017	0	0
G3	0.677	0.21	0.055	0	0.485	0.051	0	0	0.488	0	0	0
G4	0.651	0.299	0.212	0.061	0.482	0.123	0	0	0.492	0.008	0	0
G5	0.619	0.462	0.279	0.225	0.452	0.296	0.071	0.017	0.475	0.111	0	0

Note. C1-C4 represent the between-dimension correlation value (C1=0, C2=0.3, C3=0.5, C4=0.7); G1-G5 refer to the pseudo-guessing parameter value (G1=0, G2=0.1, G3=0.2, G4=0.3, G5=0.5).

Table 5. Power of TLI to Reject Unidimensionality in Noncompensatory Models

	SS = 500				SS = 1,000				SS = 2,000			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
30 Items												
G1	0.921	0.197	0	0	0.451	0	0	0	0.175	0	0	0
G2	0.513	0.297	0.063	0.054	0.479	0.059	0	0.028	0.297	0	0	0
G3	0.512	0.283	0.1	0.1	0.497	0.112	0.011	0.017	0.292	0.014	0	0
G4	0.53	0.515	0.186	0.201	0.539	0.194	0.099	0.06	0.295	0.051	0	0
G5	0.651	0.543	0.284	0.295	0.56	0.292	0.187	0.219	0.518	0.184	0.061	0
60 Items												
G1	0.989	0	0	0	0.502	0	0	0	0.293	0	0	0
G2	0.819	0.099	0.03	0	0.806	0.106	0	0	0.908	0.027	0	0
G3	0.695	0.206	0.053	0	0.493	0.051	0	0	0.488	0	0	0
G4	0.709	0.493	0.203	0.048	0.522	0.193	0	0	0.499	0.016	0	0
G5	0.764	0.519	0.286	0.322	0.531	0.301	0.104	0.012	0.517	0.1	0	0

Note. C1-C4 represent the between-dimension correlation value (C1=0, C2=0.3, C3=0.5, C4=0.7); G1-G5 refer to the pseudo-guessing parameter value (G1=0, G2=0.1, G3=0.2, G4=0.3, G5=0.5).

DOI: https://doi.org/10.30564/jiep.v1i1.356

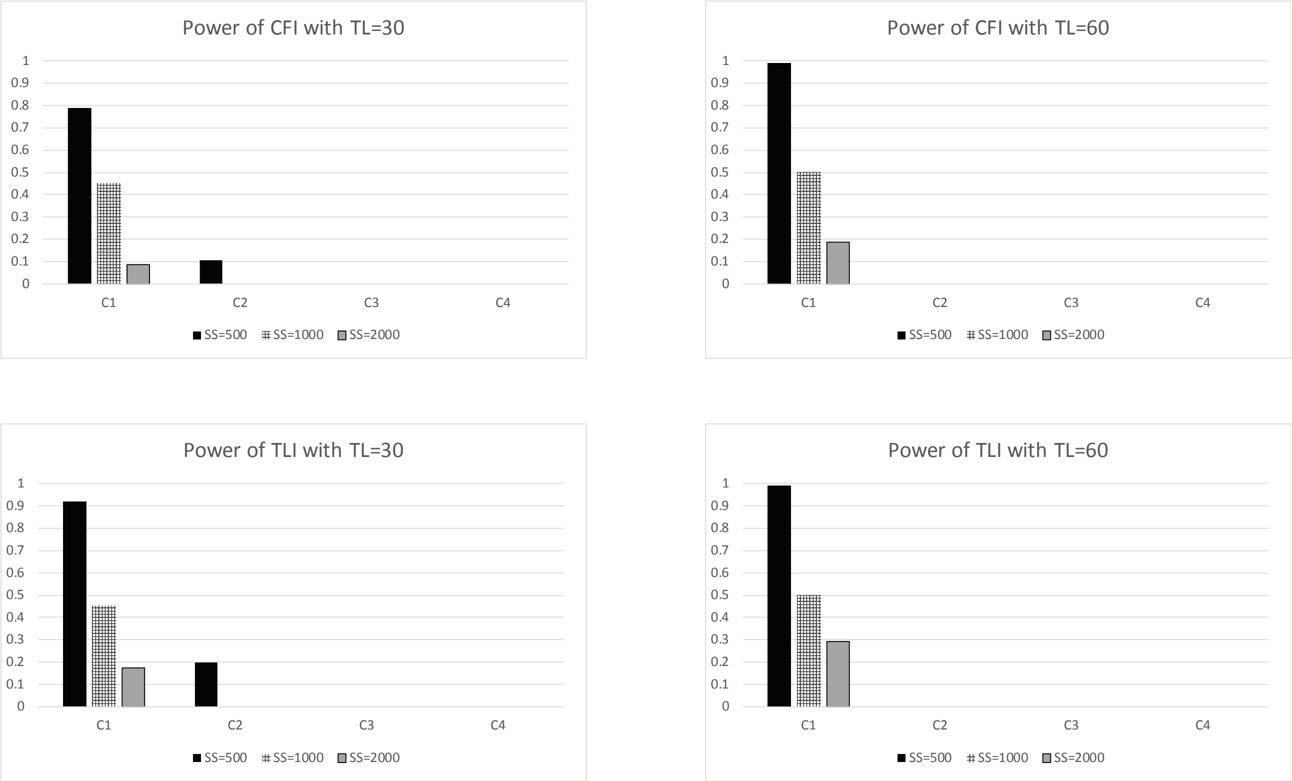


Figure 3. Power of RMSEA, CFI, and TLI with no guessing in noncompensatory models

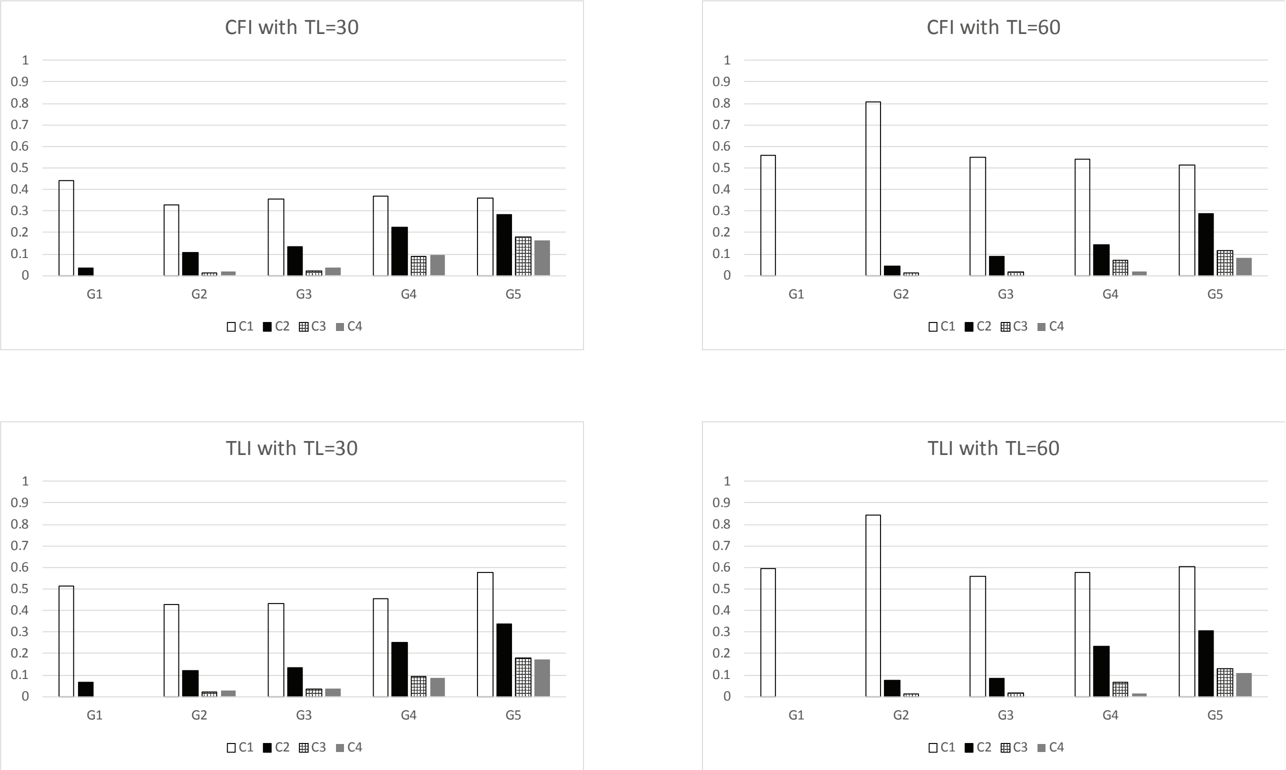


Figure 4. Power of RMSEA, CFI, and TLI with different guessing levels in noncompensatory models

DOI: https://doi.org/10.30564/jiep.v1i1.356

plot the average power rate of CFI and TLI across three sample sizes. As can be seen, CFI and TLI have power greater than 0.8 only when between-dimension correlation is zero and the guessing value is 0.1; however, as far as guessing is concerned, despite the fact that the power of CFI and TLI changes with the change of guessing magnitude, no patterns can be observed that suggest a systematic influence of guessing upon their statistical power to detect multidimensionality.

5. Discussion and Conclusion

It should be noted that due to various warnings and caveats against the use of cutoff values for model fit indices for model fit assessment, methodologists have developed an equivalence testing approach[57][58] that does not rely on common cutoff values but create adjusted cutoff values. Such methodological developments notwithstanding, the use of cutoff values for model fit indices to assess model fit still remains hugely popular among studies across many disciplines[59][60][61][62][63][64].

The purpose of this study was to reiterate the important point that cut off values for model fit indices should never be used indiscriminately for dimensionality assessment. In addition, we explored how guessing and the nature of multidimensionalty, two factors ignored in previous studies, could further negatively affect the performances of cutoff values for model fit indices in dimensionality assessment. Specifically, we systematically investigated how guessing affected the statistical power of commonly used cutoff values for RMSEA, CFI, and TLI to refute unidimensionality with binary data generated with either compensatory or noncompensatory IRT models. It was hypothesized that as all the simulation studies which established the commonly used cutoff values for model fit indices were based on factor analysis models, which do not accommodate guessing, such cutoff values (RMSEA < 0.05; CFI > 0.95; TLI > 0.95) would exhibit poor statistical power with binary data generated with IRT models that include a guessing parameter within.

The simulation results show that when data were generated with a 3PL compensatory multidimensional IRT model, increases of guessing value lead to decreases of the power of RMSEA, and such decreases were exacerbated with the increase of between-dimension correlation. For CFA and TLI, when the between-dimension correlation was no greater than 0.3, they were robust to guessing effect and their power remained constantly one regardless of the guessing magnitude and sample size; the systematic effect of guessing upon the power of CFA and TLI appeared when the between-dimension correlation was 0.5, in that their power decreased with the increase of guessing magnitude, and such decreases became more pronounced with a longer test length. When the between-dimension correlation was 0.7, all three indices had virtually no power to detect multidimensionality. When data were generated with a 3PL noncompensatory multidimensional IRT model, guessing did not have a systematic effect upon the statistical power of the three model fit indices, although it should be noted that a small change of guessing magnitude can result in a considerable change of statistical power for a given model fit index. For example, as can be observed in Table 6, when the sample size was 500 and the between-dimension correlation was zero, the power of TLI dropped from 0.921 to 0.513 when the guessing magnitude changed from zero to 0.1.

We also investigated how the cutoff values performed with the baseline conditions in which the guessing value was zero (the model reduced to a 2PL compensatory/noncompensatory IRT model). In the compensatory case, it was found that when the between-dimension correlation was no greater than 0.5, CFI and TLI exhibited statistical power higher than 0.90 regardless of test length and sample size; RMSEA displayed the same pattern when the test length was 30 items. When the test length was 60 items, RMSEA performed poorly when the between-dimension correlation was 0.5 with sample size equal to 500 or 1000, and its statistical power went up to 0.941 when the sample size was 2000. None of the model fit indices performed satisfactorily when the between-dimension-correlation was 0.7, regardless of sample size and test length. It seems that when such high correlations exist between dimensions, none of RMSEA, CFA, and TLI can statistically differentiate such structures from unidimensional structure. In the noncompensatory case, it was found that RMSEA had no power at all to detect multidimensionality regardless of the sample size, test length, and between-dimension correlation. CFI and TLI displayed unsatisfactory power (less than 0.8) in most conditions with some exceptions: CFI had a power of 0.989 when the sample size was 500, the test length was 60 items, and the between-dimension correlation was zero; TLI had power greater than 0.9 when the test length was 60 items, and the between-dimension correlation was zero.

The well-known advice that model fit indices should not be used indiscriminately is corroborated by the results found in the baseline conditions where no guessing is assumed to exist. Apparently, the power of CFI, TLI, and RMSEA is affected by the test length in that (a) with the same sample size, a longer test results in decreased power of the three model fit indices, and (b) a larger sample size is required for the three model fit indices to perform well in a longer test. Taking the perspective that the degree to which a model is misspecified is determined by the statistical power to detect such misspecifications[30], we conclude that with the same generating model and same model misspecification type, increased test lengths result in amelioration of model misspecification due to the reduced statistical power: it is more difficult to detect model specification of a less misspecified model. Although we only investigated two test lengths in the current study, it is expected that with tests consist of more than 60 items, the power of these three model fit indices will be lower than those presented in Tables 1-3. Another reason that the common cutoff values cannot be generalized is that the magnitude of factor loadings impact their performances. Heene, Hilbert, Draxler, and Ziegler[65] found that the statistical power of RMSEA, SRMR, and CFI changes with the change of the magnitude of factor loadings. In other words, if data were simulated with different item parameters, the results in Tables 1-3 might not be replicated. This is further evidence that "golden rules" are extremely difficult, if not possible, to find.

The findings that the performances of RMSEA, CFA, and TLI are subject to guessing effect are hardly surprising. When guessing effect exists, the measurement quality deteriorates, and as nicely stated by Hancock and Muller[37], "as measurement quality gets poorer, common data-model fit indices-absolute, parsimonious, and/or incremental in nature-paint an increasingly and deceptively favorable picture of the model's latent structure." In other words, the guessing effect introduces noise into data, which can mask the true latent structure. What is surprising, however, is that a small increase of guessing magnitude can result in precipitous decrease of the statistical power of a certain model fit index considered in this study. Take RMSEA as an example: as can be seen in Table 1, when guessing increases from 0 to 0.1, its power drops from 1 to 0.015 with a samples size of 1000 and a test length of 30 items. CFI and TLI do not have such drastic changes of power as RMSEA does, yet an increase of 0.1 of guessing magnitude can still result in a decrease of 0.2 to 0.3 regarding their statistical power.

One piece of advice to practitioners and researchers who are interested in using model fit indices to assess unidimensionality is that the consequent conclusions regarding unidimensionality should be taken with a grain of salt and interpreted cautiously, especially with binary data that represent scores on multiple-choice questions. As shown in this study, existence of guessing decreases the sensitivity of RMSEA, CFA, and TLI to multidimensionality. It is recommended that if model fit indices are used for unidimensionality assessment, other techniques such as DIMTEST and DETECT that can model guessing should be used jointly, although it is possible that different methods might disagree with each other[66]. When facing inconsistence dimensionality assessment results from different methods, we recommend using the bifactor modeling approach[16][67], which, unlike the other unidimensionality assessment approaches that attempt to provide a yes/no answer regarding unidimensionality, provides a detailed picture of the consequence of treating the data as unidimensional and allows one to empirically examine whether and how the model parameter estimates change by fitting a unidimensional structure to a multidimensional data set.

One limitation of the current study is that in the compen-satory cases, we generated data using item parameters that were designed to realistically mimic a simple structure of multidimensionality with each item predominantly measuring one dimension. Although not strictly a simple structure, the generating items are distinct from those used to mimic a complex structure of multidimensionality, and it is expected that the cutoff values of the three model fit indices considered in the current study will perform dif-ferently with items following a complex structure.

Taken together, the results in the present study show that when data follow a compensatory multidimensional structure, guessing systematically decreases the power of the commonly used cutoff values of RMSEA, CFA, and TLI. When data follow a noncompensatory multidimensional structure, these cutoff values do not perform well and guessing does not seem to affect their power in a systematic manner. Such findings point to two directions for possible future research. First, as guessing systematically affects the distribution of model fit indices, which is another reason cutoff values for these model fit indices should not be used, the performances of other more recent methods such as the equivalent testing approach mentioned earlier and the permutation test[68][69], when dealing with data containing guessing, should be investigated. The finding that cutoff values for model fit indices performed poorly in assessing the dimensionality of data generated with noncompensatory models, together with those by Hattie, Krakowski, Rogers, and Swaminathan[42] and Svetina[43], suggest that dimensionality assessment techniques that are based on the compensatory framework do not work well with noncompensatory data. In that regard, methods specifically designed for noncompensatory data are direly needed.

References

[1] Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, 13(2), 113-127.

DOI: https://doi.org/10.30564/jiep.v1i1.356

[2] Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22(4), 249-262.

[3] Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. Applied Psychological Measurement, 25(2), 146-162.

[4] Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. Language Testing, 27(1), 119-140.

[5] Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

[6] Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. Educational and Psychological Measurement, 37(4), 827-838.

[7] Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2), 139-164.

[8] Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: an interrelationship decomposition approach. Educational and Psychological Measurement, 73(4), 581-600.

[9] Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55(2), 293-325.

[10] Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. Psychometrika, 72(2), 141.

[11] Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17(4), 283-296.

[12] Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Testing unidimensionality in polytomous Rasch models. Psychometrika, 67(4), 563-574.

[13] Debelak, R., & Arendasy, M. (2012). An algorithm for testing unidimensionality and clustering items in Rasch measurement. Educational and Psychological Measurement, 72(3), 375-387.

[14] Heene, M., Kyngdon, A., & Sckopke, P. (2016). Detecting violations of unidimensionality by order-restricted inference methods. Frontiers in Applied Mathematics and Statistics, 2, 3.

[15] McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. Multivariate Behavioral Research, 30(1), 23-40.

[16] Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research, 16(1), 19-31.

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

[18] Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64(2), 213-249.

[19] Millsap R. E. (2011). Statistical Approaches to Measurement Invariance. Taylor and Francis Group: New York.

[20] Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

[21] Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. Structural Equation Modeling, 15(1), 136-153.

[22] Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26(3), 457-477.

[23] Luo, Y. (2018). A short note on estimating the testlet model with different estimators in Mplus. Educational and Psychological Measurement, 78(3), 517-529.

[24] Luo, Y., & Dimitrov, D. M. (2018). A short note on obtaining point estimates of the IRT ability parameter with MCMC estimation in Mplus: how many plausible values are needed? Educational and Psychological Measurement.

[25] Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. Psychometrika, 52(3), 393-408.

[26] 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), 1-55.

[27] Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. Structural Equation Modeling, 22(4), 504-516.

[28] Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. Structural Equation Modeling, 12(1), 41-75.

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

[30] Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. Multivariate Behavioral Research, 42(3), 509-529.

[31] Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling, 11(3), 320-341.

[32] Yuan, K. H. (2005). Fit indices versus test statistics. Multivariate Behavioral Research, 40(1), 115-148.

[33] Huggins-Manley, A. C., & Han, H. (2016). Assessing the sensitivity of weighted least squares model fit indexes to local dependence in item response theory models. Structural Equation Modeling, 24(3), 331-340.

[34] Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. Organizational Research Methods, 14(3), 548-570.

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

[36] McNeish, D., An, J., & Hancock, G. R. (2017). The thorny relation between measurement quality and fit index cutoffs in latent variable models. Journal of Personality Assessment.

[37] Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. Educational and Psychological Measurement, 71(2), 306-324.

[38] 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(2), 193-214.

[39] Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. Applied Psychological Measurement, 27(3), 159-203.

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

[41] Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. Applied Psychological Measurement, 27(6), 395-414.

[42] Hattie, J., Krakowski, K., Jane Rogers, H., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. Applied Psychological Measurement, 20(1), 1-14.

[43] Svetina, D. (2013). Assessing dimensionality of noncompensatory multidimensional item response theory with complex structures. Educational and Psychological Measurement, 73(2), 312-338.

[44] Muthén, L., & Muthén, B. (1998-2012). Mplus User's Guide (Seventh Edition). Los Angeles, Ca: Muthén & Muthén.

[45] Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

[46] West, S. G., Taylor, A. B., & Wu, W. (in press). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), Handbook of Structural Equation Modeling. New York: Guilford Press.

[47] Joreskog, K., & Sorbom, D. (1996). User's reference guide. Chicago, IL: Scientific Software International.

[48] Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10(1), 1-19.

[49] Wilson, D.,Wood, R., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). TESTFACT: Test scoring and full information item factor analysis (Version 4.0). Lincoln-wood, IL: Scientific Software International.

[50] Reckase, M. (2009). Multidimensional item response theory (Vol. 150). New York: Springer.

[51] Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.). Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82-89). Minneapolis, MN: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

[52] Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

[53] Zhang, J., & Stout, W. F.(1999). Conditional covariance structure of generalized compensatory multidimensional items. Psychometrika, 64(3), 129-152.

[54] Gessaroli, M. E., & De Champlain, A. F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. Journal of Educational Measurement, 33(2), 157-179.

[55] Gessaroli, M. E., De Champlain, A. F., & Folske. (1997, March). Assessing dimensionality using a likelihood-ratio chi-square test based on a non-linear factor analysis of item response data. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

[56] Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23(2), 267-269.

[57] Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. Structural Equation Modeling: A Multidisciplinary Journal, 23(3), 319-330.

[58] Marcoulides, K. M., & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence test-

ing to assess structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 24(1), 148-153.

[59] Blömeke, S., Dunekacke, S., & Jenßen, L. (2017). Cognitive, educational and psychological determinants of prospective preschool teachers' beliefs. European Early Childhood Education Research Journal, 1-19.

[60] Campbell, P., Hope, K., & Dunn, K. M. (2017). The pain, depression, disability pathway in those with low back pain: a moderation analysis of health locus of control. Journal of Pain Research, 10, 2331-2339.

[61] Cougle, J. R., Summers, B. J., Allan, N. P., Dillon, K. H., Smith, H. L., Okey, S. A., & Harvey, A. M. (2017). Hostile interpretation training for individuals with alcohol use disorder and elevated trait anger: a controlled trial of a web-based intervention. Behaviour Research and Therapy, 99, 57-66.

[62] Drinkwater, K., Denovan, A., Dagnall, N., & Parker, A. (2017). An assessment of the dimensionality and factorial structure of the revised paranormal belief scale. Frontiers in Psychology, 8, 1693.

[63] Firmin, R. L., Lysaker, P. H., McGrew, J. H., Minor, K. S., Luther, L., & Salyers, M. P. (2017). The Stigma Resistance Scale: A multi-sample validation of a new instrument to assess mental illness stigma resistance. Psychiatry Research, 258, 37-43.

[64] Govender, K., Cowden, R. G., Asante, K. O., George, G., & Reardon, C. (2017). Validation of the child and youth resilience measure among South African adolescents. PloS one, 12(10), e0185815.

[65] Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. Psychological Methods, 16(3), 319.

[66] Luo, Y., & Al-Harbi, K. (2016). The utility of the bifactor method for unidimensionality assessment when other methods disagree: an empirical illustration. SAGE Open, 6(4), 2158244016674513.

[67] Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5), 667-696.

[68] Jorgensen, T. D., Kite, B. A., Chen, P. Y., & Short, S. D. (2017). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. Psychological Methods, 23(4), 708-728.

[69] Kite, B. A., Jorgensen, T. D., & Chen, P. Y. (2018). Random permutation testing Applied to measurement invariance testing with ordered-categorical indicators. Structural Equation Modeling: A Multidisciplinary Journal, 25(4), 573-587.

DOI: https://doi.org/10.30564/jiep.v1i1.356