An Assessment of Various Item Response Model and Structural Equation Model Fit Indices To Detect Unidimensionality

Sam Elias University of Western Australia
John Hattie University of Auckland
Graham Douglas University of Western Australia

Address all correspondence to:

John Hattie
Faculty of Education, The University of Auckland
Private Bag 92019
Auckland
New Zealand
j.hattie@auckland.ac.nz
Abstract

A fundamental problem in structural modeling and item response modeling is to determine whether a set items is unidimensional. Both models have a plethora of fit indicators, and the aim of this study is to compare the Stout T-statistic derived within item response modeling with many competing indices used in structural equation modeling to assess unidimensionality. Given that it is rare to find a “perfectly” unidimensional test, a further aim is to ascertain how many items from a second dimension are necessary before the various indices indicate that the item set is no longer “essentially” unidimensional.
Within both item response theory (IRT) and structural equation modeling (SEM) there has been a plethora of indicators aimed to help the user make decisions about the adequacy of fit of data to various models. From the IRT standpoint, one of the more successful indicators of dimensionality, or number of factors, is the Stout T-statistic (described below). It has a theoretical justification, and has been shown to most powerful at detecting one versus more than one dimension/factor. A major aim of the present study is to assess the performance of the Stout T statistic with many of the more defensible fit statistics commonly used within structural modeling, and to suggest that this statistic may be of much value to SEM users. The article also indicates how various attributes of the items (such as item difficulty/threshold, discrimination, and guessing) can affect the performances of many of the commonly used fit indices.

Under the common factor or structural equation models, responses to a set of observed variables by N subjects are summarized by a sample covariance matrix. A hypothesized population covariance matrix summarized by a number of parameters (in CFA the chosen parameters may for example include restrictions on the residual variances, factor loadings, factor variances and covariances) is then compared to the sample covariance matrix for fit. A chi-square based statistic is used to provide the goodness-of-fit test to assess whether the residual differences from the hypothesized and observed covariance matrices are sufficiently small, and whether the hypothesized covariance matrix should be accepted. The problems of this chi-square statistic are well documented (Bentler & Bonett, 1980), with particular reference to the undue influence of the sample size on the magnitude of the statistic.

With few exceptions, most of the currently used goodness-of-fit statistics are ad hoc, and it is rare to find a theoretically justifiable criterion. Exceptions include
the more recent RMSEA or Dk statistic, and the Stout T-statistic. In a study aiming to assess the performance of a large set of goodness-of-fit statistics, Marsh, Balla and McDonald (1980) reported five such indices that appeared to be relatively independent of sample size. These indices are the fitting function incremental type 2 index (FFI2), the likelihood ratio incremental type 2 index (LHRI2), the chi-square incremental type 2 index (CHII2), the Tucker-Lewis index (TLI), and the Cudeck and Brown rescaled Akaike information criterion incremental type 2 index (CAKI2). In a later publication, Marsh and Balla (1994) recommended the use of five further indices, McDonald’s Dk index (also referred to as RMSEA), McDonald’s Mc index, the goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI), and the relative noncentrality index (RNI). These ten indices are based on variations of the basic chi-square statistic, and each index is scaled to lie on an interval in which the end points are defined for a perfect model fit and no model fit. Table 1 presents the ten effective CFA indices, the original investigator(s) of each index, and the formula.

Four of these indices Dk, Mc, GFI and AGFI are considered “stand-alone” indices, as they are based on a hypothesized (target) model fitting the sample data. The remaining are considered Incremental Type 2 Indices as they are based on the difference between a target model and an alternative model such as a null model. A Type 2 index is defined as \(|t-n|/|e-n|\), where \(t\) is the value of the stand-alone index for the target (hypothesized) model, \(n\) is the value of the index for the null model (all the variables are assumed to be uncorrelated), and \(e\) is the expected value of the stand-alone index if the target model is true. The six incremental Type 2 indices are scaled to vary on the interval 0-1, with 1 representing a perfect fit and 0 representing the fit.
of a null model; however, sampling fluctuations may produce values outside this range.

A major purpose of fit indices is to detect model misspecification rather than the adequacy of modeling the null hypothesis. Those who use structural models and factor analysis need indices to assist in detecting well-fitting from badly-fitting models (see Maitai & Mukherjee, 1991). Most of the evaluations of these indices, however, have been related to the sensitivity of these indices to sample size or to the estimation method. There are far fewer evaluations as to their sensitivity to model misspecification. Hu and Bentler (1998) assessed various indices under differing conditions of under-parameterization and concluded that CAK, GFI, and AGFI are not sensitive to misspecification; and TLI, RNI, Mc and RMSEA are moderately sensitive to simple model misspecification (they also found interaction effects with method of estimation). They called for more studies on the performance of fit indices under various types of correct and misspecified models.

**The Stout T-statistic**

The principle of local independence requires that responses to different items are uncorrelated when \( \theta \) (the person’s ability) is fixed, although it does not require that items be uncorrelated over groups in which \( \theta \) varies. As many have demonstrated, the statement of the principle of local independence contains the mathematical definition of latent traits. That is, \( \theta_1, \ldots, \theta_k \) are latent traits, if and only if they are quantities characterizing examinees such that, in a subpopulation in which they are fixed, the ability scores of the examinees are mutually statistically independent. Thus, a latent trait can be interpreted as a quantity that the items
measure in common, since it serves to explain all mutual statistical dependencies among the items. As it is possible for two items to be uncorrelated and yet not be entirely statistically independent, the principle is more stringent than the factor analytic principle that their residuals be uncorrelated. If the principle of local independence is rejected in favor of some less restrictive principle then it is not possible to retain the definition of latent traits, since it is by that principle that latent traits are defined. It is possible, however, to reject or modify assumptions as to the number and distributions of the latent traits and the form of the regression function (e.g., make it nonlinear instead of linear), without changing the definition of latent traits.

A set of items can be said to be unidimensional when it is possible to find a vector of values $\theta = (\theta_i)$ such that the probability of correctly answering an item $k$ is $\pi_{ik} = f_k(\theta_i)$ and local independence holds for each value of $\theta$. This definition is not equating unidimensionality with local independence, because it can further require that it is necessary to condition only on one $\theta$ dimension and that the probabilities $p_{ik}$ can be expressed in terms on only one dimension.

McDonald (1981) outlined ways in which it is possible to weaken the strong principle, in his terminology, of local independence. The strong principle implies that not only are the partial correlations of the test items zero when the latent traits are partialed out, but also the distinct items are then mutually statistically independent and their higher joint moments are products of their univariate moments. A weaker form is to ignore moments beyond the second order and test the dimensionality of test scores by assessing whether the residual covariances are zero (see also Lord & Novick, 1968, pp. 225, 544–545). McDonald (1979, 1981) argued
that this weakening of the principle does not create any important change in anything that can be said about the latent traits, though strictly it weakens their definition.

Stout (1987, 1990) also used this weaker form to develop his arguments for “essential unidimensionality”. He devised a statistical index based on the fundamental principle that local independence should hold \textit{approximately} when sampling from a subpopulation of examinees of \textit{approximately equal ability}. He defined essential unidimensionality as: a test \((U_1, \ldots, U_N)\) of length \(N\) is said to be essentially unidimensional if there exists a latent variable \(\theta\) such that for all values of \(\theta\),

\[
\frac{1}{N(N-1)} \sum_{1 \leq i < j \leq N} \text{Cov}(U_i, U_j | \theta) \approx 0. \tag{1}
\]

That is, \textit{on average}, the conditional covariances over all item pairs must be small in magnitude (for more detail regarding the theoretical developments see Junker, 1991; Junker & Stout 1991; Nandakumar, 1987, 1991; Nandakumar & Stout, 1993; Stout, 1990). Essential unidimensionality thus replaces local independence and is based on assessing only the dominant dimensions.

Stout then developed an empirical notion of unidimensionality to match his weaker form. Either a subjective analysis of item content or an exploratory factor analysis is used to develop a core set of items, and this set is termed by Stout the assessment subtest. The remaining set of items, termed the partitioning subtest, is used to partition examinees into groups for a stratified analysis. When the total set of items is unidimensional, then the assessment and partitioning set are both unidimensional, but when the dimensionality is greater than one, then “the
partitioning subtest will contain many items that load heavily on at least one other dimension not measured by the assessment subtest”.

Stout (1990) defined essential dimensionality ($d_E$) of an item pool as the minimum dimensionality necessary to satisfy the assumption of essential independence. When $d_E = 1$, essential unidimensionality is said to hold. Essential dimensionality only counts the dominant traits that are required to satisfy the assumption of essential independence. Stout introduced a T statistic, which will be small when the assumption of essential independence holds within subgroups and $H_0$ will be accepted (the test is essentially unidimensional). When the assumption of essential independence fails within subgroups, $T$ will be large and $H_0$ will be rejected.

There have been recent reviews of the Stout procedure that have demonstrated the power of the procedure (Hattie et al., 1996), and refinements that have improved some of the steps within the Stout procedure (Stout, 1996, 1997). A more detailed description of these improvements to the Stout method can be found in Elias (1997), Kim, Zhang, & Stout (in press), Nandakumar and Stout (1993), and Nandakumar, Yu, Li, and Stout (in press).

**Purposes of the present study**

The aim of this presentation is threefold. First, it is possible to convert many of the goodness of fit indices developed within the common factor model or structural equation modeling for use in item response theory and vice versa, and then compare them to the Stout T-statistic (see Gessaroli & de Champlain, 1996). Second, an aim is to ascertain how many items from a second dimension are necessary before the various indices determine that the item set is no longer
essentially unidimensional. Third, most simulations merely use items with
unknown/unspecified psychometric properties to assess the effects on the various
indices. Instead, in this study items of varying qualities are added to demonstrate the
importance of considering items with particular qualities, especially demonstrating
their effects on the various indices.

It is noted that, while these goodness of fit indicators can be compared, there
are important distinctions between the IRT and SEM models. The former is
fundamentally a non-linear model, as the probability of getting an item correct is
usually non-linearly related to the underlying ability (θ) (typically a normal ogive
function), and the latter is commonly used as a linear model—although there are
many instances of non-linear SEM models (refs.).

Method

A unidimensional set of 15 items was generated. It is not uncommon for test
developers to use approximately 15 items to measure a single ability, and longer tests
often contain shorter subtests about this length. With 15 items it is possible to define
up to 5 dimensions given the claim that at least three items are needed to define a
dimension (McDonald & Mulaik, 1979).

To generate the data (the core items sets representing one dimension and the
added items representing the second dimension are correlated), a multidimensional
compensatory three parameter model as outlined in Hattie (1984) was used, whereby
the probability of a correct response is given by

\[
P(x_i = 1 | a_i, b_i, \theta_j) = c_i + \frac{1 - c_i}{1 + \exp\{-d \sum_k (a_{ik} \theta_j - b_{ik})\}}
\]
where $P(x_{ij}=1)$ is the probability of a correct response to item $i$ by person $j$; $a_{ik}$ is a vector of discrimination parameters for item $i$ on dimension $k$; $b_{ik}$ is a vector of difficulty parameters for item $i$ on dimension $k$; $c_i$ is the guessing parameter for item $i$; $\theta_{jk}$ is a vector of ability parameters for person $j$ on dimension $k$; $d$ is a scaling factor with a value of 1.7.

Using this model, it is possible to vary the item difficulty, discrimination and guessing parameters, as well as the number of dimensions underlying the simulated data. For this study, data sets were generated with varying levels of item difficulty, guessing and discrimination. The item parameter levels were chosen to be as realistic as possible, so that the results from the simulation can be of practical significance for teacher-made and nationally normed tests.

To the core unidimensional 15 item sets, a further 15 items from a second dimension were sequentially added. With each addition of an item from the second dimension, it was intended that the dominance of the first dimension would progressively decrease, and that this would be reflected in the values of the eleven indices which were calculated at each sequential item addition. The aim of adding these items sequentially was to assess at what point the Stout-T index and other indices indicate that the first unidimensional set was no longer essentially unidimensional.

For the core unidimensional items, two sets of difficulty values were chosen (-1, -.5, 0, .5, 1; and -2, -1, 0, 1, 2), and guessing was either 0, or .15. The 4 parameters, relating to the sequentially added 15 items, were completely crossed to produce 36 generating models: Difficulty (narrow [-1, 1] or wide [-2, 2]) x guessing (0, .15), Discrimination/correlation (.1, .3, .5) x difficulty of the added items (all -1,
0, or 1). For each of these 36 models, item sets using 1000 examinees (assuming ability was normally distributed) and 15 replications were employed to estimate each index (details are in Elias, 1997; Hattie, 1984, 1985). That only 15 replications was used is a limitation of this study, although this number of replications was sufficiently powerful to find reliable differences, and the computer costs of increasing the number of replications was prohibitive (given that the number of runs increased factorially relating to the number of addition items). NOHARM (Fraser, 1981), DIMENSION (Krakowski & Hattie, 1992), and DIMTEST (Stout, et al. 1991) were used.

Results

The first section of the results summarizes the results for the structural model-based indices. In all cases, these indices varied within their expected limits, although in no case did the index vary over the complete potential range nor was there an observable transition point indicating a change from a uni- to two-dimensional set of items. Only for the Tucker-Lewis procedure did the index fall in the range considered indicative of “good fit” (in this case >.90; Bentler & Bonett, 1980, although Bentler, 1999 argued for > .95) when there was a unidimensional set, but it also remained in this range for the two-dimensional set. Only three of the indices increased (or decreased) as expected when the dimensions were close to orthogonal (r = .1), and the LHRI2 and CAK12 did not decrease under any condition.

Five of these indices were able to discriminate in the expected direction under all conditions (RNI, CHI2, MC, Dk, & Stout T). The other indices moved in the opposite direction than expected either when there was a low relationship between
the dimensions, or when easy or average difficulty items were added (Table 2). Only the Stout T index provided an indication of transition from one to two dimensions.

As it is a typical and often highly recommend index within structural equation modeling (and for space reasons), the performance of the Tucker-Lewis index is used as an illustration of the effects of adding items. The following series of graphs show the TLI variation for each parameter level as items are added to the unidimensional set (Figures 1 to 4). The average value of the TLI for the set of 15 items was .994. It increased in value when very difficult items were added, and thus falsely implies that the set is becoming more unidimensional as the number of these items from a second factor is added, and only decreased when easy or medium difficult items were added (Figure 1). TLI also increased as the correlation between factors is .5 or greater but only after at least half the items from the second dimension are added, otherwise it decreased as desired. The changes in slopes were similar when there was and was not guessing but the magnitude of the TLI increased in the presence of guessing. A major problem is that in all conditions the TLI remains above the .9 level, and even when it performs as desired the differences are reflected only at the second and third decimal point.

Table 3 presents the manner in which each index was affected by the addition of certain types of items (note, all except Dk and Stout are expected to decrease as items are added). None of the Type II indices nor the Mc index performed in a satisfactory manner, and tended to perform in the expected direction only when the more difficult items were added. The Cudeck and Browne rescaled Akaike Information Criterion decreased as items from the second dimension were added, but after approximately half the items were added it tend to increase again. While the GFI and AGFI generally performed satisfactory, the range of values was extremely
narrow (.97 to .99) and thus for practical purposes are of little value. Only the McDonald Dk and Stout T statistic performed satisfactorily, and the Dk index had no identifiable transition point(s).

The Stout T-statistic satisfied all criteria, in that it varied within its expected limits, varied over the complete potential range, increased as expected, and there was a clear indication of the transition from unidimensional to multidimensional item sets. The Stout index is expected to have a small unidimensional value close to 0.0, and a large multidimensional upper value. The index values increased as expected, and the lower and upper values of -0.29 and 4.74 representing the 15 item unidimensional set and the 30 item multidimensional set respectively (see Figure 5). Across all conditions of added items, there were minor variations in the performance of the Stout T statistic. As items were added from a second dimension that were more difficult, and correlated less than .5 with the first dimension, then the Stout index was more effective at determining uni- and two-dimensional item sets. Across all conditions, the sharp transition occurred when more than one-third the number of items in the first dimension was added.

Conclusions

Two indices performed in the desired manner, in that they both increased as the number of items from a second dimension were added to the core unidimensional set of items. Although worthy of inclusion as a goodness-of-fit measure, the McDonald Dk or RMSEA index had no transition point such that a user could claim there was evidence for a uni- or two-dimensional data set. The results from these simulations provide strong evidence, however, that the Stout T statistic is an effective index for assessing the degree of unidimensionality of a set of items. The decision to accept a
set of items as unidimensional was not affected by various item parameters such as
difficulty, guessing, or correlation between factors. The Stout index more readily
detects multidimensionality under all conditions, and particularly when: there is no
guessing on the added items; there is a correlation less than .3 between the
dimensions; and when items of low or medium difficulty are added to the core item
set. After approximately one-third more than the number of items in the
unidimensional set are added, then the Stout T statistic indicates the presence of a
second dimension.

Future research could profitably investigate a generalization of the Stout T
index to detect the dimensionality, rather than only essential unidimensionality.

Kim, Zhang, and Stout (in press) have developed DETECT (Dimensionality
Evaluation To Enumerating Contributing Traits) for simultaneously assessing the
latent dimensionality structure and to determine the degree of multidimensionality
present in test data. Similarly, converting the Stout indices for use in structural
equation modeling, and investigating its properties (such as sensitivity to sample
size) could be most fruitful. Most critically, this study has pointed to the importance
of considering the effects of the nature of items and their effects on the performance
of the indices. Most studies that have investigated the properties of SEM based
indices have ignored the importance of the difficulty, discrimination, and guessing.
References


### Table 1. Effective CFA Indices For Assessing Model Fit.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Original Investigators</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stand-alone indices</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dk</td>
<td>McDonald, 1990; McDonald &amp; Marsh, 1990</td>
<td>$\left[ \frac{\chi^2}{(N-1)} \right] - \frac{df}{N}$</td>
</tr>
<tr>
<td>Mc</td>
<td>McDonald, 1990; McDonald &amp; Marsh, 1990</td>
<td>$\exp\left(-\frac{D_k}{2}\right)$</td>
</tr>
<tr>
<td>GFI</td>
<td>Joreskog &amp; Sorbom, 1981; Steiger, 1989, 1990</td>
<td>$\frac{p}{2D_k + p}$</td>
</tr>
<tr>
<td>AGFI</td>
<td>Joreskog &amp; Sorbom, 1981; Steiger, 1989, 1990</td>
<td>$1-(p \times (p+1)/2df) \times (1-GFI)$</td>
</tr>
<tr>
<td><strong>Incremental Type 2 Indices</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFI2</td>
<td>Joreskog &amp; Sorbom, 1981</td>
<td>$\frac{</td>
</tr>
<tr>
<td>LHR12</td>
<td>Joreskog &amp; Sorbom, 1981</td>
<td>$\frac{</td>
</tr>
<tr>
<td>CHII2</td>
<td>Joreskog &amp; Sorbom, 1981</td>
<td>$\frac{</td>
</tr>
<tr>
<td>TLI</td>
<td>Tucker &amp; Lewis, 1973</td>
<td>$\frac{(n2/dfn-t2/dft)}{(n2/dfn-1.0)}$</td>
</tr>
<tr>
<td>CAKI2</td>
<td>Akaike, 1974, 1987; Cudeck &amp; Browne, 1983</td>
<td>$\frac{</td>
</tr>
<tr>
<td>RNI(DkI2)</td>
<td>Bentler, 1990; McDonald &amp; Marsh, 1990</td>
<td>$\frac{(Dkn-Dkt)}{(Dkn-0)}$</td>
</tr>
</tbody>
</table>

where, $\chi^2$ is the chi-square value, df is the number of degrees of freedom, N is the sample size, $p=tr(E^{-1}S)$, E is the population covariance matrix, S is the sample covariance matrix, $FF=\frac{\chi^2}{(N-1)}$; N is the sample size, $LHR=\exp(-.5FF)$, $CAK=FF+2q/N$; q is the number of estimated parameters.
Table 2. Performance of the fit-indices

(y=yes as expected, n=no)

<table>
<thead>
<tr>
<th></th>
<th>Difficulty range</th>
<th>r between dimensions</th>
<th>Difficulty of added item</th>
<th>Guessing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1   1</td>
<td>-2   2</td>
<td>.1  .3  .5</td>
<td>-1.0  .0  1.0</td>
</tr>
<tr>
<td>FFI2</td>
<td>y    y</td>
<td>n    y</td>
<td>y  y  y</td>
<td>n  y  y  y  n  y</td>
</tr>
<tr>
<td>LHRI2</td>
<td>n    n</td>
<td>n    n</td>
<td>n  n  n</td>
<td>y  n  n  n  n</td>
</tr>
<tr>
<td>TLI</td>
<td>y    y</td>
<td>n    y</td>
<td>y  y  y</td>
<td>n  n  y  y  y  y</td>
</tr>
<tr>
<td>CAKI2</td>
<td>n    n</td>
<td>n    n</td>
<td>n  n  n</td>
<td>n  n  n  n  n  n</td>
</tr>
<tr>
<td>RNI</td>
<td>y    y</td>
<td>n    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>CHI2</td>
<td>y    y</td>
<td>y    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>MC</td>
<td>y    y</td>
<td>y    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>GFI</td>
<td>y    y</td>
<td>n    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>AGFI</td>
<td>y    y</td>
<td>n    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>Dk</td>
<td>y    y</td>
<td>y    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>Stout T</td>
<td>y    y</td>
<td>y    y</td>
<td>y  y  y</td>
<td>y  y  y  y  y  y</td>
</tr>
<tr>
<td>Total</td>
<td>8    8</td>
<td>3    8</td>
<td>8  8  8</td>
<td>7  7  8  8  8  8  8</td>
</tr>
</tbody>
</table>
Fit indices for unidimensionality

Table 3
Manner in which the indices varied as items of various types were added.

<table>
<thead>
<tr>
<th>Index</th>
<th>Difficulty of items</th>
<th>r between dimensions</th>
<th>Range of difficulty</th>
<th>Level of guessing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mid/easy</td>
<td>difficult</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FFI2</td>
<td>decrease</td>
<td>stable</td>
<td>decrease</td>
<td>stable</td>
</tr>
<tr>
<td>LHRI2</td>
<td>stable</td>
<td>increase</td>
<td>stable</td>
<td>stable</td>
</tr>
<tr>
<td>TLI</td>
<td>stable</td>
<td>decrease</td>
<td>increase</td>
<td>decrease</td>
</tr>
<tr>
<td>CAKI2</td>
<td>stable</td>
<td>de+increase</td>
<td>de+increase</td>
<td>de+increase</td>
</tr>
<tr>
<td>RNI</td>
<td>stable</td>
<td>decrease</td>
<td>stable</td>
<td>de+increase</td>
</tr>
<tr>
<td>CHII2</td>
<td>stable</td>
<td>increase</td>
<td>stable</td>
<td>stable</td>
</tr>
<tr>
<td>Mc</td>
<td>stable</td>
<td>stable</td>
<td>stable</td>
<td>stable</td>
</tr>
<tr>
<td>GFI</td>
<td>decrease</td>
<td>decrease</td>
<td>stable</td>
<td>decrease</td>
</tr>
<tr>
<td>AGFI</td>
<td>decrease</td>
<td>decrease</td>
<td>stable</td>
<td>decrease</td>
</tr>
<tr>
<td>Dk</td>
<td>increase</td>
<td>increase</td>
<td>increase</td>
<td>increase</td>
</tr>
<tr>
<td>Stout</td>
<td>increase</td>
<td>increase</td>
<td>increase</td>
<td>increase</td>
</tr>
</tbody>
</table>
Figure 1

TLI Variation for Additional Items According to Difficulty of Added Item Levels.
Figure 2

TLI Variation for Additional Items According to Correlation Levels.

0.9905 0.9910 0.9915 0.9920 0.9925 0.9930 0.9935 0.9940 0.9945 0.9950
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

Additional Items

Index

COR=0.1  COR=0.3  COR=0.5
Fit indices for unidimensionality

Figure 3

TLI Variation for Additional Items According to Difficulty Range Levels.
Figure 4

TLI Variation for Additional Items According to Guessing Levels

Fit indices for unidimensionality
Figure 5

Performance of the Stout T-index as items are added to the 15-item core set