DOCUMENT RESUME

ED 397 100 TM 025 221

AUTHOR De Champlain, Andre
TITLE Assessing the Dimensionality of Item Response Matrices Using a Goodness-of-Fit Index Based on Noncentrality.
PUB DATE Apr 96
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chi Square; *Cutting Scores; *Goodness of Fit; *Item Response Theory; *Matrices; *Sample Size; Simulation; Test Length
IDENTIFIERS *Dimensionality (Tests); Law School Admission Test

ABSTRACT

The usefulness of a goodness-of-fit index proposed by R. P. McDonald (1989) was investigated with regard to assessing the dimensionality of item response matrices. The m subscript k index, which is based on an estimate of the noncentrality parameter of the noncentral chi-square distribution, possesses several advantages over traditional tests of hypotheses as well as other descriptive fit indices. This study considered the behavior of the index in simulated conditions across different test lengths, sample sizes, dimensional structures, and other factors. The appropriateness of the recommended model fit cutoff value (0.9) was also studied. Four hundred unidimensional data sets based on the Law School Admission Test were simulated for the study. Results suggest that the index cutoff value recommended by its developers as being indicative of model fit is too high for the data sets simulated. With respect to the simulated unidimensional data sets, results show that none of the manipulated factors had any practical effect on mean m subscript k index values, supporting claims that the index is sample-size and estimation independent. These findings may offer some guidelines to the practitioner interested in using the index to assess the dimensionality of an item response matrix. (Contains 1 figure, 8 tables, and 46 references.) (SLD)

Assessing the Dimensionality of Item Response Matrices Using a Goodness-of-Fit Index Based on Noncentrality

André De Champlain[1]
Law School Admission Council

Paper presented at the meeting of the American Educational Research Association
April, 1996
New York, NY

Running Head: A GOODNESS-OF-FIT INDEX BASED ON NONCENTRALITY

[1] The author wishes to thank Peter Pashley for his helpful comments and Vicki Tompkins for her programming assistance.

Assessing the Dimensionality of Item Response Matrices Using a Goodness-of-Fit Index Based on Noncentrality

The use of item response theory (IRT) models in the fields of educational and psychological measurement has increased significantly over the past few decades. Common IRT models are currently being employed to address a host of measurement issues such as the estimation of the statistical characteristics of items (Mislevy & Bock, 1990), the detection of differentially functioning items (Thissen, Steinberg, & Wainer, 1993) as well as the equating of scores obtained on alternate forms of a test (Lord, 1977; 1980). However, several assumptions must be met in order to legitimately use the majority of IRT models, one of which is unidimensionality of the latent ability space (Hambleton & Swaminathan, 1985). Simply stated, most IRT models assume that the probability of a correct response on a given item can me modeled as a function of a single underlying proficiency or person parameter in addition to item parameters. For example, the three-parameter logistic IRT function (Lord & Novick, 1968) given by,

$$P_i(x_i=1|a_i,b_i,c_i,\theta_j)=c_i+(1-c_i)\ \frac{e^{Da_i(\theta_j-b_i)}}{1+e^{Da_i(\theta_j-b_i)}} \qquad (1)$$

assumes that the probability of a correct response on a given item (x=1) can be modeled as a function of an item discrimination (a), difficulty (b) and lower asymptote (c) parameter as well as a latent trait or proficiency (θ) hypothesized to underlie the item response matrix. Clearly, this assumption of unidimensionality is often violated in practical testing situations where the response to an item is dependent upon several secondary proficiencies in addition to the hypothesized trait. This led researchers to develop a host of descriptive and inferential statistics to assess dimensionality, or perhaps more commonly, departure from the assumption of unidimensionality. At the present time, the two most popular methods with regards to assessing the dimensionality of item response matrices appear to be Stout's DIMTEST procedure as well as statistics based on nonlinear factor analysis (NLFA).

Stout's DIMTEST nonparametric procedure is based on his notions of essential independence and essential dimensionality. Essential dimensionality corresponds to the number of latent traits that are required to satisfy the assumption of essential independence given by,

$$\frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j|\theta)| \approx 0 \qquad N \to \infty \qquad (2)$$

that is, a mean absolute residual covariance value that tends towards zero at fixed levels of the latent trait as the number of items increases towards infinity. The terms in equation (2) can be defined as follows:

N = the number of test items;

U_i = the response to item i for a randomly selected examinee;

U_j = the response to item j for a randomly chosen examinee.

Stout (1987; 1990) and Nandakumar & Stout (1993) proposed several versions of the T statistic to test for this assumption of unidimensionality. T_2, the most powerful version of the statistic, is given by,

$$T_2 = \frac{T_{L,2} - T_b}{\sqrt{2}} .$$

Readers should consult Nandakumar and Stout (1993) to obtain more information on the computational steps associated with the T_2 statistic. The T_2 statistic is asymptotically distributed $N(0,1)$ under the null hypothesis of unidimensionality. In a series of studies carried out by Nandakumar & Stout (1993), the T_2 statistic was found to be quite accurate in correctly determining essential unidimensionality or departure from the assumption with multidimensional data sets except when the test contained few items (less that 25) and the sample sizes were small (less than 750 examinees). Roussos and Stout (1994) have also suggested using cluster analysis in order to select sets of dimensionally distinct items that could subsequently be submitted to DIMTEST runs.

Another promising approach with regards to assessing the dimensionality of item response matrices is the one that treats common IRT models as a special case of a more general NLFA model (NLFA). Research undertaken by Bartholomew (1983), Goldstein and Wood (1989), McDonald (1967) and Takane and De Leeuw (1987) has shown that common IRT models and NLFA models are mathematically equivalent. This led others to suggest that analyzing the residual correlation matrix obtained after fitting an m-factor NLFA model to an item response matrix, where m is the number of factors specified, might be a useful and informative way of assessing dimensionality (Hambleton & Rovinelli, 1986; Hattie, 1985). For example, small residuals obtained after fitting a one-factor NLFA model to an item response matrix would suggest that the data set is unidimensional. Several

descriptive indices and hypothesis tests have been proposed for both limited-information and full-information NLFA models (Hattie, 1984). The estimation of parameters in the former models is restricted to the information contained in the lower-order marginals (e.g., the pairwise relationships between items) whereas the information contained in the higher-order marginals (i.e., in the item response vectors) is utilized to estimate the parameters of the latter models.

PRELIS2/LISREL8 (Jöreskog & Sörbom, 1993) is a comprehensive covariance structure modeling package that enables the user to fit, among other things, a variety of factor analytic models via several estimation procedures. Irrespective of the estimation procedure employed, the parameters of factor analytic models in LISREL8 are estimated so as to minimize the following general fit function,

$$F=(s-\sigma)' W^{-1}(s-\sigma), \qquad (4)$$

where,

s = Sample item covariance matrix;

σ = Population covariance matrix;

W^{-1} = A weight matrix referred to as the correct weight matrix.

A chi-square goodness-of-fit statistic, based on Browne's (1982; 1984) research, is provided in LISREL8 to aid in assessing model fit and is given by,

$$\chi^2=(N-1) * Min(F), \qquad (5)$$

where, N corresponds to the number of examinees in the sample and Min (F) is the minimum value of the fit function given in (4) for a specific model. This statistic is distributed asymptotically as a chi-square distribution with degrees of freedom equal to,

$$.5(p)*(p + 1) - t,$$

where p is equal to the number of items and t is the number of independent parameters estimated in the model.

TESTFACT (Wilson, Wood, & Gibbons, 1991) is a full-information factor analysis package that uses the marginal maximum likelihood (MML) procedure proposed by Bock and Aitkin (1981) to estimate parameters via the EM algorithm of Dempster, Laird and Rubin (1977). The thresholds and factor loadings are estimated so as to maximize the following multinomial probability function,

$$L_m = P(X) = \frac{N!}{r_1!\ r_2!\ \ldots r_s!}\ \tilde{P}_1^{r_1}\ \tilde{P}_2^{r_2}\ \ldots \tilde{P}_s^{r_s}, \qquad (6)$$

where, r_s is the frequency of response pattern s and $\tilde{P}_s$ is the marginal probability of the response pattern based on the item parameter estimates. The function outlined in (6) is commonly referred to as full-information item factor analysis (Bock, Gibbons, & Muraki, 1988). A likelihood-ratio chi-square statistic is provided to help in assessing the fit of a given model as well as a G^2 difference test to aid in determining the adequacy of competing models. The likelihood-ratio chi-square statistic can be defined as,

$$G^2 = 2\sum_{l}^{2^n} r_l \ln \frac{r_l}{N\tilde{P}_l}, \qquad (7)$$

where r_l is the frequency of response vector l and $\tilde{P}_l$ is the probability of response vector l. The degrees of freedom for this statistic are equal to,

$$2^n(m + 1) + m(m-1)/2,$$

where n is the number of items and m, the number of factors. However, this G^2 statistic often poorly approximates the chi-square distribution given the large number of empty cells typically encountered with actual data sets (the number of unique response vectors is equal to 2^n). For that reason, Haberman (1977) suggests using a likelihood-ratio chi-square difference test to assess the fit of alternative models. Research undertaken by Berger and Knol (1990) showed that the likelihood ratio chi-square difference test was generally unable to correctly identify the number of dimensions underlying an item response matrix. However, these results should be interpreted cautiously given the small number of replications (10) performed in the study.

Although useful and informative for the assessment of dimensionality, hypothesis tests possess one important shortcoming: they are inherently sensitive to sample size effects. McDonald (1994) underscored this limitation of all hypothesis tests when he stated:

> "A problem with tests of significance - and this holds equally for asymptotic chi square tests- is that we know a priori that any hypothesis of restricted dimensionality is false, and will be rejected at a sufficiently large sample size" (p.76).

MacCallum (1990) and McDonald (1989) also cautioned against the use of hypothesis tests to assess the fit of an m-factor model given that they are inherently biased with large sample sizes and are dependent upon distributional assumptions. Finally, Kaplan (1990) suggests that the interaction of sample size, violation of distributional assumptions and misspecification of the model all contribute to inflated Type I error rates with chi-square distributed fit statistics.

As an alternative to hypothesis tests, McDonald and Mok (1995) suggest using fit indices, originally proposed within the structural equation modeling (SEM) literature, to assess the dimensionality of item response matrices. McDonald and Mok (1995) argue that the indices proposed within the field of SEM carry over to the assessment of fit, i.e., dimensionality, in IRT models. Indeed, the assessment of unidimensionality can be viewed as an SEM fitting problem, i.e., the extent to which a one-factor model accounts for the item response probabilities estimated on a given test form. The presentation of these fit indices is beyond the scope of the present paper. However, readers interested in obtaining an overview of these statistics should consult Gerbing and Anderson (1993), McDonald and Marsh (1990), Marsh, Balla and McDonald (1988), Mulaik, James, Van Alstine, Bennett, Lind, & Stillwell, (1989) and Tanaka (1993) for thorough reviews.

Tanaka (1993) proposed a taxonomy that allows the practitioner to differentiate fit indices suggested thus far in the SEM literature along six dimensions. In addition, the taxonomy enables the user to be more aware of the respective strengths and limitations of each statistic. Tanaka (1993) states that fit indices can vary as a function of being:

(1) *Population or sample based.* Population based fit indices estimate a known population parameter whereas sample based statistics describe model fit for the sample at hand.

(2) *Simple or complex.* Fit indices are considered to be simple if they penalize overparameterization whereas they are complex if they do not employ such a correction.

(3) *Normed or nonnormed.* Normed fit indices are approximately restricted to lie within a (0,1) range while nonnormed indices are not.

(4) *Absolute or relative.* Relative fit indices are defined with respect to a comparison model whereas absolute fit indices do not make use of such a reference point.

(5) *Estimation free or specific.* Estimation free fit indices provide information that is unrelated to the estimation procedure used while specific indices yield different fit summaries across different estimation methods.

(6) *Sample size independent or dependent.* Sample size-dependent fit indices vary as a function of sample size whereas sample size independent are not influenced by this factor.

McDonald (1989; 1994) proposed a goodness-of-fit index which appears to be promising with respect to assessing the dimensionality of item response matrices. The index, labelled m_k, can be defined as,

$$m_k = e^{[-(1/2)d_k]}, \quad (8)$$

where,

$$d_k = (\chi^2 - df)/N \quad (9)$$

that is, an estimate of the noncentrality parameter of the noncentral chi-square distribution. The degrees of freedom associated with the chi-square statistic are symbolized by df in equation (9) whereas N corresponds to the sample size. McDonald and Mok (1995) suggest that a value of .9 is indicative of "acceptable" fit. However, the authors stress that the latter cutoff is a convenient "rule of thumb" and that model selection should also entail judgment on the part of the researcher, a point previously made by Browne and Cudeck (1993).

According to Steiger and Lind (1980), La Du (1986), and Tanaka (1993), the m_k index possesses several desirable properties:

- It is grounded in a firm statistical framework (chi-square distribution);
- It penalizes overparameterization by subtracting the degrees of freedom from the chi-square value and hence favors a simpler model;
- It is normed to lie in a (0,1) interval which greatly facilitates interpretation for naive users. Nonetheless, the value of the m_k index can exceed unity due to sampling error.
- It is an absolute index and hence does not depend upon a comparison model for interpretation. It is important to point out, however, that Tanaka (1993) has argued that the m_k index is in actuality a relative index in that its value attains unity at the saturated model. Tanaka (1993) states that the one-to-one correspondence noted between the saturated model and the observed data would indicate that m_k is a relative fit index.
- It is estimation method independent and hence the practitioner can feel free to utilize either limited- or full-information factor analytic models to assess the dimensionality of

an item response matrix.

- Its value does not vary as a function of sample size.

Preliminary research undertaken by McDonald (1989), McDonald and Marsh (1990) and McDonald and Mok (1995) shows that, as hypothesized, the m_k index does not vary systematically with sample size nor as a function of estimation procedure. It is important to point out, however, that the analyses undertaken by these authors were quite limited in scope. First, the analyses were based on actual data sets where the true number of dimensions underlying item responses was unknown. Second, a small number of data sets were analyzed in each of the above mentioned studies (at most 12) which considerably limits the extent to which one can generalize their findings to other conditions. Hence, more research needs to be undertaken to systematically examine, in controlled conditions, the extent to which the m_k index value is affected by different sample sizes, estimation procedures and other pertinent factors. In addition, the appropriateness of the suggested cutoff value (.9) should also be carefully examined in a host of conditions.

Purpose

The purpose of this simulation study is three-fold:

1. To determine the extent to which the value of the m_k index varies as a function of sample size, test length and estimation procedure with simulated unidimensional data sets.
2. To ascertain the extent to which the value of the m_k index varies as a function of sample size, test length, latent trait correlation and estimation procedure with simulated two-dimensional data sets.
3. To determine whether the suggested cutoff value of .9 is appropriate for the unidimensional and multidimensional data sets generated in this study.

Methods

Unidimensional data set simulations

In the first part of the study, unidimensional item response vectors were simulated based on the three-parameter logistic IRT model outlined in equation (1). In addition, data sets were generated to vary as a function of two test lengths (20 and 40 items) as well as two sample sizes (2500 and 5000

examinees). The 40-item data sets were comprised of two 20-item tests (i.e., the item parameters for items 21-40 were the same as those for items 1-20). In order to simulate item responses that reflected those typically encountered with actual achievement test data, 20 IRT item parameters were randomly selected from one form of the Law School Admission Test (LSAT) and used in the data generation process. The item parameters that were selected are shown in Table 1.

Insert Table 1 about here

Each cell of this 2 (sample size) x 2 (test length) design was replicated 100 times for a total of 400 unidimensional data sets.

Each of the 400 unidimensional data sets was then analyzed using both TESTFACT (Wilson, Wood, & Gibbons, 1991) as well as PRELIS2/LISREL8 (Jöreskog & Sörbom, 1993).

One- and two-factor models were fit to each simulated unidimensional data set with TESTFACT using all default values. Given that the likelihood ratio chi-square difference test follows a chi-square distribution even in the presence of sparse frequency tables (Haberman, 1977), the latter was selected as the fit statistic for all unidimensional data set analyses. The G^2 difference test was computed in the following fashion,

$$G^2_{diff} = G^2_{1\text{-}F} - G^2_{2\text{-}F}, \qquad (10)$$

where $G^2_{1\text{-}F}$ is the value of the likelihood-ratio chi-square statistic obtained after fitting a one-factor model and $G^2_{2\text{-}F}$ is the value of the likelihood-ratio chi-square statistic obtained after fitting a two-factor model. The degrees of freedom for the difference test are also computed by subtracting those associated with the one- and two-factor model fit statistics.

Once the fit statistics were obtained, McDonald's (1989) m_k index was computed for all data sets using equation (8). The likelihood-ratio chi-square difference test statistic was substituted in equation (9).

The fit of a unidimensional model (i.e., one-factor model) was then ascertained using LISREL8. Initially, the asymptotic covariance matrix was computed for the estimated tetrachoric correlations of the items in each simulated unidimensional data set using PRELIS2. Then, the parameters of the unidimensional model were estimated using a generally weighted least-squares (WLS) procedure which minimizes the following fit function

$$F=(s-\sigma)' W^{-1}(s-\sigma), \quad (11)$$

where,

$\underline{s}$ = Sample estimates of the threshold and tetrachoric correlation values;

$\underline{\sigma}$ = Population threshold and tetrachoric correlation values;

$\underline{W}^{-1}$ = A consistent estimator of the asymptotic covariance matrix of $\underline{s}$, referred to as the correct weight matrix.

The chi-square statistic value given in (5) was computed for all unidimensional data sets after fitting a one-factor model with LISREL8. Then, the $\underline{m}_k$ index value was calculated for each unidimensional data set using equation (8). Also, the chi-square statistic values computed using equation (5) were substituted in equation (9).

In summary, each of the 400 unidimensional data sets had two $\underline{m}_k$ index values computed, that is, one using the likelihood-ratio chi-square difference test provided by TESTFACT and another based on the chi-square fit statistic estimated using LISREL8. $\underline{M}_k$ index values exceeding 1.0 due to sampling error were fixed at 1.0 in all subsequent analyses.

Two-dimensional data set simulations

In the second part of the study, two-dimensional item response vectors were simulated based on a multidimensional extension of the three-parameter logistic IRT model (Reckase,1985) outlined in equation (1). The probability of a correct response on an item based on this multidimensional three-parameter logistic compensatory model (M3PL) is given by,

$$P_i(x_j=1|\underline{a}_i, d_i, c_i, \underline{\theta}_j) = c_i + (1-c_i)\ \frac{e^{\underline{a}_i(\underline{\theta}_j+d_i)}}{1+e^{\underline{a}_i(\underline{\theta}_j+d_i)}}, \quad (12)$$

where,

$\underline{a}_i$ = a vector of discrimination parameters for item i;

d_i = a scalar parameter related to the difficulty of item i;

$\underline{\theta}_j$ = a latent trait vector.

Reckase (1985) states that the multidimensional item discrimination parameter (MDISC) can be estimated using the following equation.

$$MDISC_i = \sqrt{a_{i1}^2 + a_{i2}^2}, \tag{13}$$

where a_{ik} is the discrimination parameter of item i on dimension k (k=1,2, ..., l). Similarly, the multidimensional analogue of item difficulty (MDIF) can also be computed using the following formula,

$$MDIF_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{n} a_{ik}^2}}. \tag{14}$$

Although Reckase (1985) recommends providing direction cosines in addition to the distance outlined in (14) when describing the MDIF value of an item, he does suggest that the distance parameter can be interpreted much like a b parameter would be for a unidimensional IRT model. Past research has shown that a two-factor model appears to underlie the item response probabilities estimated on several forms of the LSAT (Ackerman, 1994; Camilli, Wang, & Fesq, 1995; De Champlain, i press; Roussos & Stout, 1994). More precisely, the first dimension corresponds to deductive reasoning and loads on Analytical Reasoning (AR) items whereas the second factor, which loads on Logical Reasoning (LR) and Reading Comprehension (RC) items, has been labelled as reading/informal reasoning. About 25% of the items on an LSAT form measure deductive reasoning whereas the remaining 75% of the items measure reading/informal reasoning. In an effort to simulate "realistic" two-dimensional data sets, an item parameter structure that resembles that found on a typical form of the LSAT was selected. More precisely, the first dimension (factor) was constrained to load on 25% of the items while the probability of a correct response on the remaining 75% of the items require knowledge of the second latent trait. In addition, the item discrimination parameter values for the first and second dimensions utilized in the simulations were respectively randomly selected from actual LSAT AR and LR/RC items. The unidimensional item difficulty parameter estimates for the selected items were treated as MDIF values in these simulations. The chosen item parameters are shown in Table 2.

Insert Table 2 about here

In addition, as was the case with unidimensional data sets, two-dimensional item response matrices were generated to vary as a function of the same two test lengths (20 and 40 items) and two sample sizes (2500 and 5000 examinees). The 40-item data sets were also comprised of two 20-item tests. In addition, the correlation between the two latent traits was set at either 0.00 or 0.70. Past research has shown that the correlation between reading/informal reasoning and deductive reasoning proficiencies on most LSAT forms is near 0.70, which accounts for the selection of this particular value (Camilli, Wang, & Fesq, 1995; De Champlain, in press). Each cell of this 2 (sample size) x 2 (test length) x 2 (latent trait correlation) design was replicated 100 times for a total of 800 two-dimensional data sets.

Also, the fit of a one- versus a two-factor TESTFACT full-information factor analytic model was ascertained with the likelihood-ratio chi-square difference test. The m_k index was then computed for all multidimensional data sets. Finally, the fit of a unidimensional model to the simulated two-dimensional item response matrices was ascertained using LISREL8 (WLS estimation) after estimating the asymptotic covariance matrix for the estimated tetrachoric correlations of the items in each data set using PRELIS2. Again, the fit of each unidimensional model was assessed using the chi-square test statistic provided in the LISREL8 output. The m_k index was then calculated for all simulated two-dimensional data sets. As was previously the case with the unidimensional data sets, each two-dimensional item response matrix had two m_k index values: one calculated using the TESTFACT fit statistic and the other, computed with the LISREL8 chi-square test statistic. Once more, m_k index values exceeding 1.0 due to sampling error were set at 1.0 in all subsequent analyses.

Analyses: Effects of test length, sample size, latent trait correlation and estimation procedure

For unidimensional data sets, the main effects and interactions of test length, sample size as well as estimation procedure with respect to mean m_k index values were assessed using a 2 x 2 x 2 ANOVA with repeated measures on the last factor. Regarding multidimensional data sets, the main effects and interactions of test length, sample size, latent trait correlation as well as estimation procedure on the mean m_k index value was ascertained using a 2 x 2 x 2 x 2 ANOVA with repeated measures on the last factor. However, it is important to point out that the relatively large sample sizes or in this instance, data sets simulated (400 unidimensional data sets and 800 two-dimensional data sets) would more than likely yield many significant effects due to the large amount of power. However, these significant effects might bear little practical import for the user. In other words,

several small "practically" insignificant effects might turn out to be "statistically" significant. Since the purpose of this paper is to investigate whether "meaningful" fluctuations in mean m_k index values are obtained across different sample sizes, test lengths, latent trait correlation levels and estimation procedures, the emphasis will be placed on "practically" significant differences rather than statistical ones. In order to assess the "practical" significance of the main effects and interactions on mean m_k index values, a measure of effect size (ES; Cohen, 1992) was used in the present study. The ES index that was employed is defined for squared multiple correlations (R^2) and given by,

$$f^2 = R^2 * (1 - R^2), \qquad (15)$$

where,

$$R^2 = SS_{effect} / (SS_{effect} + SS_{error}).$$

In other words, f^2 is indicative of the amount of sums of squares explained by a given effect (SS_{effect}) relative to the unexplained sums of squares in the model ($SS_{effect} + SS_{error}$). Cohen suggests that a small ES corresponds to an f^2 value of .02, whereas f^2 values of .15, and equal to or greater than .35, are indicative of moderate and large ES values, respectively. Only moderate and large ESs were flagged in the current study.

Analyses: Appropriateness of the .9 m_k index value "rule-of-thumb" for acceptable model fit

As previously stated, McDonald and Mok (1995) have suggested that using an m_k index value of .9 might be useful as a "rough" indicator of model fit. Hence, in the present study, one would expect to obtain significantly larger m_k index values for the simulated unidimensional data sets given that the correct (unidimensional) model was fit to the item response matrices. In order to verify this hypothesis, an independent-groups t-test was calculated comparing mean m_k index values for simulated unidimensional and two-dimensional data sets. However, as was the case with the previous ANOVA analyses, a large amount of power is likely to result in a significant t statistic value due to the large sample sizes examined rather than any "practical" difference in mean m_k index values. In order to circumvent this situation, an ES measure will also be computed. The ES used was d (Cohen, 1992) given by,

$$d = \frac{m_A - m_b}{\sigma}, \qquad (16)$$

where,

m_a = mean m_k index value computed for the unidimensional data sets;

m_b = mean m_k index value computed for the two-dimensional data sets;

σ = the standard deviation of m_k index values for all data sets.

Cohen has suggested that d values of .2 are indicative of small ES while values of .5 and .8 respectively suggest moderate and large ESs.

Results

The results will be presented according to the three research purposes. First, findings pertaining to the effects of test length, sample size and estimation procedure on mean m_k index values for simulated unidimensional data sets will be presented. Second, the effects of test length, sample size, latent trait correlation and estimation procedure will be described for the two-dimensional data set analyses. Finally, findings relating to the appropriateness of the m_k index cutoff value of .9 will be outlined.

The effects of test length, sample size and estimation procedure on mean m_k index values for simulated unidimensional data sets

Mean m_k index values for the various test lengths, sample sizes and estimation procedures examined are shown in Table 3.

Insert Table 3 about here

The range of m_k index values was quite small (.009) and suggests that none of the factors had much of an impact on the mean value of the statistic. The results of the 2 x 2 x 2 ANOVA with repeated measures on the estimation procedure factor are shown in Table 4.

Insert Table 4 about here

As expected, all main effects and interactions were statistically significant due to the large

amount of power attributable to the large sample size (400). Since the purpose of these analyses was to focus on the "practical" effects rather than the statistical ones, the ESs were computed for each main effect and interaction of the model and are presented in Table 5.

Insert Table 5 about here

According to the criteria set forth by Cohen (1992), none of the terms in the model would be flagged as showing either moderate or large ESs. The largest f^2 index value (.134), estimated for the main effect of sample size, was below the cutoff value deemed to be indicative of a moderate ES (.15). Hence, neither test length, nor sample size, nor estimation procedure had a "practical" effect on the mean m_k index value.

The effects of test length, sample size, latent trait correlation and estimation procedure on mean m_k index values for simulated two-dimensional data sets

Mean m_k index values for the various test lengths, sample sizes, latent trait correlations and estimation procedures examined are shown in Table 6.

Insert Table 6 about here

The range of m_k index values was again very small (.073). The largest discrepancy in m_k index values occurred between data sets that were analyzed using LISREL8 and simulated to contain 20 items, 5000 examinees as well as no correlation between the latent traits (Mean m_k index = 0.926) and those analyzed with TESTFACT and generated to contain 40 items, 5000 examinees, and a correlation of .70 between the latent traits (Mean m_k index = 1.00). The results of the 2 x 2 x 2 x 2 ANOVA with repeated measures on the estimation procedure factor are shown in Table 7.

Insert Table 7 about here

Again, 12/15 effects were statistically significant due to the large amount power attributable to the large sample size (800 data sets) used in the analysis. The ESs were once more computed for each

main effect and interaction of the model and are presented in Table 8.

Insert Table 8 about here

Using Cohen's (1992) criteria, none of the terms in the model would be flagged as a large ES. However, the f^2 index value (.241) estimated for the "test length x latent trait correlation" interaction would be classified as a moderate ES. The main effect f^2 index values computed for latent trait correlation (.248) and test length (.205) factors are ignored given that they are contained in the two-way interaction term. A plot of the mean m_k index values by test length and latent trait correlation level is shown in Figure 1.

Insert Figure 1 about here

Findings plotted in Figure 1 show that the mean m_k index value was very stable for the 40-item data sets across both levels of latent trait correlation. The mean m_k index value was equal to .9735 with data sets simulated to have zero correlation between latent traits while it was equal to .9759 when the latent trait correlation was set at 0.70. However, this was not the case with the 20-item data sets where the mean m_k index value increased from .9301 ($r_{\theta 1,\theta 2} = 0.00$) to .9817 ($r_{\theta 1,\theta 2} = 0.70$).

Appropriateness of the ".9" m_k index value rule-of-thumb for acceptable model fit

As was previously stated, McDonald and Mok (1995) proposed using an m_k value of .9 as a "rule-of-thumb" to indicate model fit. The results obtained in this study would seem to suggest that this cutoff value is quite helpful for the simulated unidimensional data sets but ineffective for the two-dimensional data sets. With respect to unidimensional data sets, none of the m_k index values computed were below .9 (none were less than .99) which would lead one to conclude that a single latent trait appears to be needed to account for the estimated item response probabilities. However, the m_k values were still large for simulated two-dimensional data sets (none were below .92) and would, according to McDonald and Mok's (1995) proposed rule-of-thumb, suggest that the item response matrices are unidimensional in nature which is clearly not the case, especially for structures generated to have zero correlation between latent traits. It is important to point out that the mean m_k index value computed for simulated unidimensional data sets ($M = 0.9969$) was significantly greater than that

calculated for the two-dimensional item response matrices ($\underline{M} = 0.9653$) from both a statistical ($t(1811.1) = 43.47$, $p<.0001$) as well as a practical standpoint ($\underline{d} = 1.15$). Nonetheless, all m_k index values exceeded the recommended cutoff value of .9.

Discussion

The purpose of this study was to investigate the usefulness of a goodness-of-fit index proposed by McDonald (1989) with regards to assessing the dimensionality of item response matrices. The m_k index, which is based on an estimate of the noncentrality parameter of the noncentral chi-square distribution, possesses several advantages over traditional tests of hypothesis as well as other descriptive fit indices. Among its strengths, it is purported to be population-based, sample size independent as well as estimation method free. However, little research has be undertaken to assess the behavior of the index in controlled, i.e., simulated, conditions as well as the the extent to which its claimed strengths actually hold across different test lengths, sample sizes, dimensional structures and other factors. Finally, the appropriateness of a recommended model fit cutoff value (.9) also needs to be examined.

The results obtained in this study would seem to suggest that the m_k index cutoff value (0.9) recommended by McDonald and Mok (1995) as being indicative of model fit is too high for the data sets simulated. Although m_k index values were, as expected, significantly greater for unidimensional data sets than for two-dimensional item response matrices, the overall mean value for the latter data sets substantially exceeded the "rule-of-thumb" proposed by McDonald and Mok (.9).

With respect to simulated unidimensional data sets, results show that none of the manipulated factors had any practical effect on mean m_k index values. That is, neither sample size, test length, estimation procedure nor any interaction of these effects had any noticeable impact upon mean m_k index values. Hence, these findings would tend to substantiate previous claims made to the effect that the index is sample size and estimation method independent. With regards to simulated two-dimensional item response matrices, findings indicate that the interaction of test length and latent trait correlation had a moderate impact on the mean m_k index value. More precisely, increasing the correlation between the latent traits seems to yield high m_k index values solely for 40-item data sets.

What are the implications of these results for researchers and practitioners interested in using the m_k index as one of several tools for assessing the dimensionality of an item response matrix?

First, m_k index values appear to be intrinsically linked to the item parameter values and

structure of a given data set. In other words, the m_k index values obtained in this investigation are probably indicative of what one might expect with a LSAT type structure. However, these values could vary substantially for other types of tests or assessments. Hence, it seems advisable to develop model fit cutoff values that are specific to the test or data set of interest. In other words, empirical results derived from simulation studies would seem to be a useful and practical way of setting m_k index cutoff values. Second, the m_k index appears to be, as alleged, sample size independent which makes it an attractive alternative to hypothesis tests and other fit indices (e.g., AIC criterion, Akaike, 1987) for the assessment of dimensionality. Third, findings show that the m_k index is estimation method free which provides a greater deal of flexibility to the practitioner with regards to selecting a given factor-analytic approach. Using limited-information factor analytic models might yield m_k index values, and subsequently model fit decisions, which are essentially equivalent to those derived from full-information factor analytic models. Finally, the effect of varying the correlation between latent traits appears to have an impact on mean m_k index values solely for tests that contain over 20 items. Hence, the impact of correlated latent traits with regards to mean m_k index values might be somewhat irrelevant with pretest and other test sections that contain few items.

Although the results obtained in this preliminary study are informative, it is important to point out some of the limitations of this investigation and offer suggestions for future research.

First and foremost, the data sets in the present study were simulated to reflect typical LSAT forms. Therefore, the reader should not generalize the findings obtained in this study to other testing programs without further research. Future research should focus upon examining the extent to which m_k index values vary as a function of item parameters and latent trait structures. Second, it is important to point out that the fit of a simple (unidimensional) model was examined for all simulated data sets. Obviously, the fit of more complex models (two-, three factor models, etc.) should be examined in future investigations in order to determine whether the m_k index value does indeed penalize for overparameterization, as claimed. Third, the m_k index values were computed based solely on two factor analytic estimation procedures (full-information factor analysis using MML estimation and LISREL8 using WLS estimation). It would be interesting to compute index values using other estimation procedures and factor analytic models to examine whether the m_k index value fluctuates noticeably. Also, the behavior of the m_k index value was assessed according to only two test lengths and sample sizes. Future investigations should focus on examining the index with a larger number of sample sizes and test lengths. Finally, given the vast number of SEM goodness-of-fit statistics

reported in the literature over the past few decades, it would be informative to undertake a comparative study emphasizing the strengths and weaknesses of each approach with simulated and real data sets.

Although preliminary, it is hoped that the findings reported in the current study will offer some guidelines to the practitioner interested in using the m_k index to assess the dimensionality of an item response matrix as well as offer some indication of the limitations to be considered when utilizing the procedure with data that resemble typical LSAT item responses. In addition, it is hoped that these initial results will foster future research that will bridge the areas of IRT and SEM with respect to not only goodness-of-fit but other issues of common interest that would benefit from a greater collaboration between both fields.

References

Ackerman, T. (1994, June). Graphical representation of multidimensional IRT analysis. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52, 317-332.

Bartholomew, D.J. (1983). Latent variable models for ordered categorical data. Journal of Econometrics, 22, 229-243.

Berger, M.P.F., & Knol, D.L. (1990, April). On the assessment of dimensionality in multidimensional item response theory models. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Bock, D.R., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. Psychometrika, 4, 443-459.

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

Browne, M.W. (1982). Covariance structures. In D.M. Hawkins (Ed.), Topics in applied multivariate analysis, (pp. 72-141). Cambridge: Cambridge University Press.

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

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. Sociological Methods and Research, 21, 230-258.

Camilli, G., Wang, M.M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. Journal of Educational Measurement, 32, 79-96.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

De Champlain, A. (in press). The effect of multidimensionality on IRT true-score equating results for subgroups of examinees. Journal of Educational Measurement.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1-38.

Gerbing, D.W., & Anderson, J.C. (1993). Monte Carlo evaluations of goodness of fit indices for structural equation models. In K.A. Bollen and J.S. Long (Eds.), Testing structural equation models, (pp. 40-65). Newbury Park, CA: Sage.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cell counts. Annals of Statistics, 5, 1148-1169.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Applied Psychological Measurement, 10, 287-302.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nyjhoff.

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

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Jöreskog, K.G., & Sörbom, D. (1993). PRELIS 2 user's reference guide. Chicago, Il.: Scientific Software, Inc.

Jöreskog, K.G., & Sörbom, D. (1993). LISREL 8 user's reference guide. Chicago, Il.: Scientific Software, Inc.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. Multivariate Behavioral Research, 25, 137-155.

La Du, T.J. (1986). Assessing the goodness of fit of linear structural equation models: The influence of sample size, estimation method, and model specification. Unpublished doctoral dissertation, New York University, New York.

Lord, F.M. (1977). Practical applications of characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

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

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

MacCallum, R.C. (1990). The need for alternative measures of fit in covariance structure modeling. Multivariate Behavioral Research, 25, 157-162.

Marsh, H.W., Balla, J.R., & McDonald, R.P. (1988). Goodness-of-fit in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometrika Monograph No. 15 ,32(4, Pt. 2).

McDonald, R.P. (1989). An index of goodness-of-fit based on noncentrality. Journal of Classification, 6, 97-103.

McDonald, R.P. (1994). Testing for approximate dimensionality. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (eds.), Modern theories in measurement: Problems and issues (pp. 63-86). Ottawa, On.: Edumetrics Research Group.

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

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

Mislevy, R.J., & Bock, R.D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, In: Scientific Software, Inc.

Mulaik, S.A., James, L.R., Van Alstine, J., Bennett, N., Lind, N., & Stillwell, C.D. (1989). An evaluation of goodness of fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

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

Reckase, M.D. (1985). The difficulty of items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Roussos, L., & Stout, W.F. (1994, April). Analysis and assessment of test structure from the multidimensional perspective. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Steiger, J.H. (1980). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 25, 173-180.

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

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

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

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural equation models. In K.A. Bollen and J.S. Long (Eds.), Testing structural equation models. (pp. 11-39). Newbury Park, CA: Sage.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland and H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Wilson, D., Wood, R., & Gibbons, R.D. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis. Mooresville, In.: Scientific Software, Inc.

Table 1

True unidimensional item parameters

Item	*a*	*b*	*c*
1	0.622132	-1.710310	0.119606
2	0.779642	0.470174	0.079124
3	0.806952	0.161454	0.162809
4	0.842712	0.081694	0.140943
5	1.152409	1.679257	0.153869
6	0.558630	-1.387155	0.119606
7	0.341596	-0.599501	0.119606
8	0.878353	1.081976	0.058036
9	0.957605	0.916684	0.196364
10	1.086517	0.693614	0.042316
11	0.751002	-0.696663	0.119606
12	0.551905	-0.315874	0.119606
13	0.630988	1.696784	0.223633
14	0.552291	-1.294931	0.119606
15	0.785618	-0.285280	0.095973
16	0.730466	-0.402966	0.119606
17	0.845300	0.004327	0.188632
18	0.792140	1.138772	0.155819
19	0.822973	1.540107	0.073885
20	0.601753	1.358651	0.111348

Table 2

True two-dimensional item parameters

Item	a_1	a_2	*MDIF*	*c*
1	0.622132	0.000000	-1.710310	0.119606
2	0.806592	0.000000	0.161454	0.162809
3	0.842712	0.000000	0.081694	0.140943
4	0.882054	0.000000	0.854201	0.184434
5	0.904691	0.000000	1.371124	0.242642
6	0.000000	0.644494	-0.892373	0.119606
7	0.000000	0.878353	1.081976	0.058036
8	0.000000	0.957605	0.916684	0.196364
9	0.000000	0.946642	1.520134	0.224578
10	0.000000	0.803943	-1.139963	0.119606
11	0.000000	0.751002	-0.696663	0.119606
12	0.000000	0.551905	-0.315874	0.119606
13	0.000000	0.688839	0.632910	0.145847
14	0.000000	0.808383	0.554415	0.208314
15	0.000000	0.567085	-0.087459	0.119606
16	0.000000	0.783265	0.256477	0.206116
17	0.000000	0.694929	-1.357711	0.119606
18	0.000000	0.543069	-0.608002	0.119606
19	0.000000	0.792140	1.138772	0.155819
20	0.000000	0.773915	0.280484	0.246003

Table 3

Mean m_t index values for simulated unidimensional data sets by test length, sample size and estimation procedure

	Estimation procedure			
	LISREL8		TESTFACT	
Test length	N=2500	N=5000	N=2500	N=5000
20 items	0.99999	1.00000	0.99661	0.99617
40 items	0.99708	1.00000	0.99491	0.99060

Table 4

ANOVA table for unidimensional data sets

Source	DF	SS	MS	F value	Pr > F
Test length	1	0.00127079	0.00127079	81.00	0.0001
Sample size	1	0.00176815	0.00176815	112.71	0.0001
Estimation procedure	1	0.00090098	0.00090098	33.05	0.0001
Test length x sample size	1	0.00157948	0.00157948	100.68	0.0001
Test length x estimation procedure	1	0.00041039	0.00041039	15.06	0.0001
Sample size x estimation procedure	1	0.00075642	0.00075642	27.75	0.0001
Test length x sample size x estimation procedure	1	0.00055767	0.00055767	20.46	0.0001

Table 5

Effect sizes for unidimensional data sets

Source	Effect size
Test length	0.10533
Sample size	0.13374
Estimation procedure	0.12348
Test length x sample size	0.12363
Test length x estimation procedure	0.06625
Sample size x estimation procedure	0.10864
Test length x sample size x estimation procedure	0.08558

Table 6

Mean m_k index values for simulated two-dimensional data sets by test length, sample size, latent trait correlation and estimation procedure

		Estimation procedure			
		LISREL8		TESTFACT	
Latent trait correlation	Test Length	N=2500	N=5000	N=2500	N=5000
$r\theta_1\theta_2 = 0.00$	20 items	0.92808	0.92649	0.93017	0.93568
	40 items	0.94645	0.95138	0.99740	0.99865
$r\theta_1\theta_2 = 0.70$	20 items	0.98447	0.98442	0.97721	0.98056
	40 items	0.94932	0.95443	0.99999	1.00000

Table 7

ANOVA table for two-dimensional data sets

Source	DF	SS	MS	F value	Pr > F
Test length	1	0.14146572	0.14146572	481.46	0.0001
Sample size	1	0.00216577	0.00216577	7.37	0.0067
Latent trait correlation	1	0.29161156	0.29161156	992.46	0.0001
Estimation procedure	1	0.23647611	0.23647611	999.99	0.0001
Test length x sample size	1	0.00010801	0.00010801	0.37	0.5444
Test length x latent trait correlation	1	0.24122229	0.24122229	820.97	0.0001
Sample size x latent trait correlation	1	0.00001574	0.00001574	0.05	0.8170
Test length x estimation procedure	1	0.23577316	0.23577316	999.99	0.0001
Sample size x estimation procedure	1	0.00002028	0.00002028	999.99	0.0001
Latent trait correlation x estimation procedure	1	0.00374214	0.00374214	999.99	0.0001
Test length x sample size x latent trait correlation	1	0.00000085	0.00000085	0.00	0.9571
Test length x sample size x estimation procedure	1	0.00230066	0.00230066	999.99	0.0001
Test length x latent trait correlation by estimation procedure	1	0.00258019	0.00258019	999.99	0.0001
Sample size x latent trait correlation x estimation procedure	1	0.00015715	0.00015715	999.99	0.0001
Test length x sample size x latent trait correlation x estimation procedure	1	0.00003578	0.00003578	999.99	0.0001

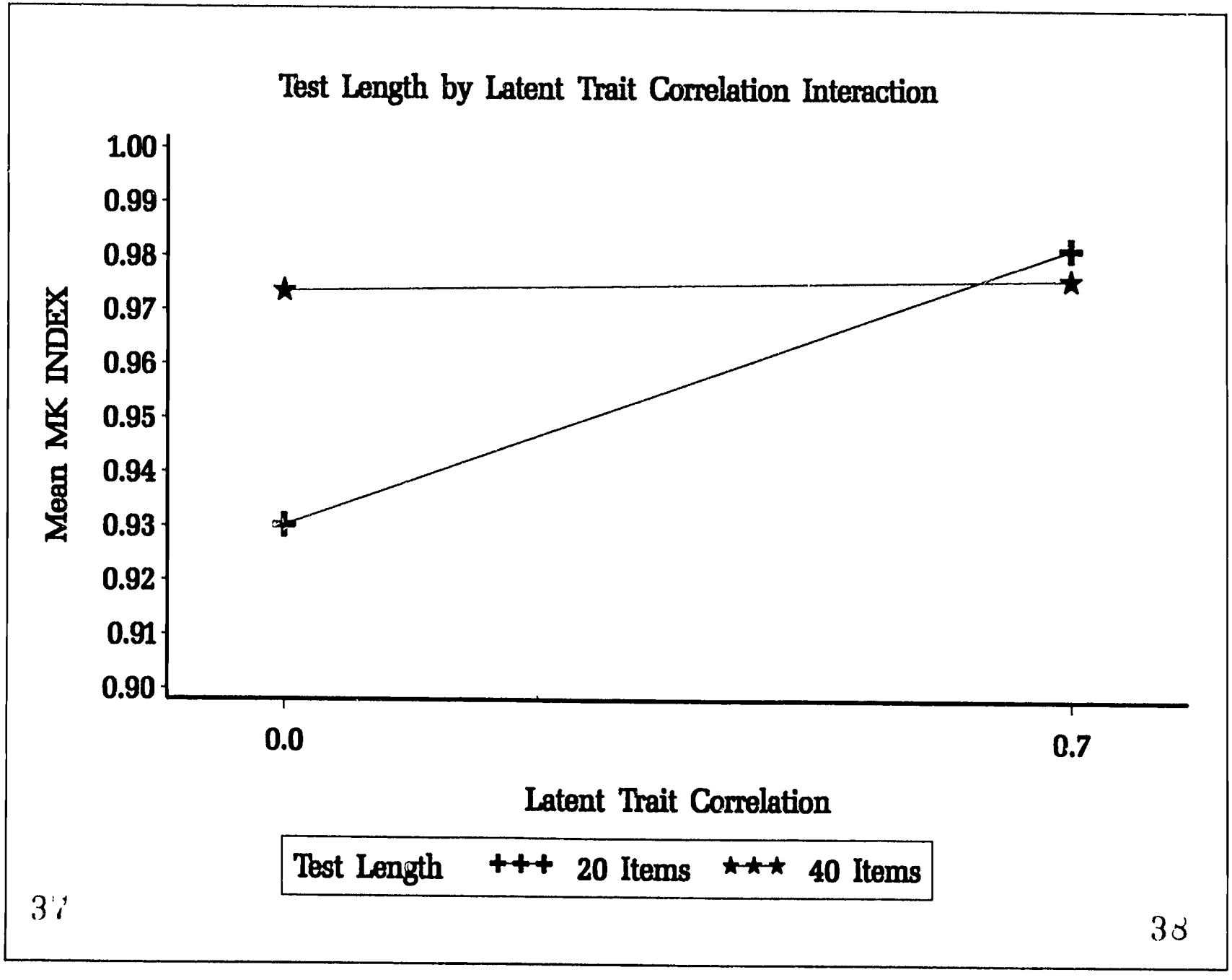
Table 8

Effect sizes for two-dimensional data sets

Source	Effect size
Test length	0.20493
Sample size	0.00611
Latent trait correlation	0.24791
Estimation procedure	0.00001
Test length x sample size	0.00031
Test length x latent trait correlation	0.24151
Sample size x latent trait correlation	0.00004
Test length x estimation procedure	0000001
Sample size x estimation procedure	0.00005
Latent trait correlation x estimation procedure	0.00001
Test length x sample size by latent trait correlation	0.00001
Test length x sample size x estimation procedure	0.00001
Test length x latent trait correlation x estimation procedure	0.00001
Sample size x latent trait correlation x estimation procedure	0.00001
Test length x sample size x latent trait correlation x estimation procedure	0.00003

Figure Captions

Figure 1. Test length by latent trait correlation interaction for simulated two-dimensional data sets


Test Length by Latent Trait Correlation Interaction
Mean MK INDEX
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
0.0
0.7
Latent Trait Correlation
Test Length
20 Items
40 Items