DOCUMENT RESUME

ED 329 584 TM 016 218

AUTHOR Berger, Martijn P. F.; Knol, Dirk L.
TITLE On the Assessment of Dimensionality in Multidimensional Item Response Theory Models. Research Report 90-8.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Dec 90
NOTE 52p.
AVAILABLE FROM Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Computer Simulation; *Data Analysis; Equations (Mathematics); Factor Analysis; Foreign Countries; *Item Response Theory; *Mathematical Models
IDENTIFIERS *Dimensionality (Tests); Eigenvalues; *Multidimensional Models

ABSTRACT

The assessment of dimensionality of data is important to item response theory (IRT) modelling and other multidimensional data analysis techniques. The fact that the parameters from the factor analysis formulation for dichotomous data can be expressed in terms of the parameters in the multidimensional IRT model suggests that the assessment of the dimensionality of the latent trait space can also be approached from the factor analytical viewpoint. Some problems connected with the assessment of the dimensionality of the latent space are discussed, and the conclusions are supported by simulated results for sample sizes of 250 and 500 on a 15-item test. Five tables contain data from the simulation; and 48 graphs illustrate eigenvalues and plotted mean residuals. A 46-item list of references is included. (Author/SLD)

ED329584

On the Assessment of Dimensionality in Multidimensional Item Response Theory Models

Research Report 90-3

Martijn P.F. Berger
Dirk L. Knol

BEST COPY AVAILABLE

ent of
CATION

Division of Educational Measurement and Data Analysis

University of Twente

Project Psychometric Aspects of Item Banking No. 52

Colofon:
Typing: L.A.M. Bosch-Padberg
Cover design: Audiovisuele Sectie TOLAB
Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling
Oplage: 125

On the Assessment of Dimensionality in Multidimensional Item Response Theory Models

Martijn P.F. Berger

Dirk L. Knol

On the assessment of dimensionality in multidimensional item response theory models , Martijn P.F. Berger & Dirk L. Knol - Enschede : University of Twente, Department of Education, December, 1990. - 44 pages

Abstract

The assessment of dimensionality of the data is not only important to item response theory (IRT) modelling, but has also been important to other multidimensional data analysis techniques. The fact that the parameters from the factor analysis formulation for dichotomous data can be expressed in terms of the parameters in the multidimensional IRT model, suggests that the assessment of the dimensionality of the latent trait space can also be approached from the factor analytical point of view. In this paper some problems connected with the assessment of the dimensionality of the latent space are discussed and the conclusions are supported by simulated results.

On the Assessment of Dimensionality in Multidimensional Item Response Theory Models

The dimensionality of a set of test items has been a source of debate in educational and psychological literature. The development of the well-known one-, two-, and three-parameter item response theory (IRT) models has increased the need of adequate tests for the unidimensionality of the latent trait space. See Hattie (1985) for a review of methods to assess unidimensionality. In many situations the assumption of a unidimensional latent trait is untenable and this has lead to the development of multidimensional latent trait models. As these models gained currency in educational research and their application became feasible by the development of computer programs like TESTFACT (Wilson, Wood & Gibbons, 1984), NOHARMII (Fraser, 1988) and MAXLOG (McKinley & Reckase, 1983), the need for a measure to assess the adequacy of the model i.e. the assessment of the dimensionality of the latent trait space, increased. It is well-known that misspecifications of the dimensionality of the model will severely change the estimates of the item parameters. Unfortunately no widely accepted index to identify the dimensionality of items is yet available.

The dimensionality of data is not only an important issue in IRT modelling, but has also been important to other multidimensional techniques, such as factor analysis. Recently it has become known, that there is an important relation between the multidimensional IRT model and the

factor analysis model for dichotomous data. Takane and de Leeuw (1987) showed the formal equivalence of the marginal likelihood of the multidimensional two-parameter normal ogive model (Bock and Aitkin, 1981) and the likelihood of the factor analysis model for binary data (Muthén, 1978). This means that the IRT parameters can be expressed in terms of the parameters of the factor analysis formulation. An empirical comparison done by Knol and Berger (1990) showed that for the simulated conditions, common factor analysis estimation procedures performed equally well as the marginal maximum likelihood estimation of the IRT parameters proposed by Bock and Aitkin (1981). The estimates of the item parameters obtained from MINRES (Harman & Jones, 1966; Zegers & Ten Berge, 1983) computed on tetrachoric correlations among the items, for example, were quite comparable with the estimates obtained from TESTFACT. These results suggest that the problem of assessing the dimensionality in multidimensional IRT models for dichotomous data can also be approached from the factor analytical point of view.

A number of empirical studies have been devoted to this problem. Hambleton and Rovinelli (1986), Tucker, Humphreys, Lloyd and Roznowski (1986) and McDonald (1985) among others, considered various measures and Muthén (1978) and Bock and Aitkin (1981) provided formal statistical tests.

The purpose of the present study is to discuss some of the problems connected with the assessment of the dimensionality of a latent space and to compare some factor analytical methods for the assessment of dimensionality with

procedures available from multidimensional IRT models. First, however, a short description of dimensionality will be given.

The Dimensionality of the Latent Trait Space

Multidimensional IRT models assume that an examinee is characterized by more than one latent trait. Let $\underline{\theta}$ be the vector of m latent traits of an examinee, i.e. $\underline{\theta} = [\theta_1, \theta_2, \ldots, \theta_m]$. An examinee can be represented as a point in an m-dimensional latent trait space. If each of the traits influences the performance of the examinee on at least two of the n items in a test, then m is the dimension of the latent space. This definition of dimensionality is connected with the principle of local independence. The formal requirement of the independence of a set of item responses $\underline{X} = [X_1, X_2, \ldots, X_n]$, is that the joint distribution of the responses given a latent trait vector $\underline{\theta}$ is equal to the product of the marginal distributions of the items given $\underline{\theta}$, i.e.:

$$P(\underline{X}=\underline{x}|\underline{\theta}) = \prod_{i=1}^{n} P(X_i=x_i|\underline{\theta}). \qquad (1)$$

This means that if a population of examinees is characterized by m latent traits which completely span the latent space, then the responses of a 'sub' population of examinees with fixed values for $\underline{\theta}$ are mutually statistically independent.

If, however, a model specifies a number of latent traits, which do not completely span the latent space, then there will still remain mutual dependencies among the items for fixed values of $\underline{\theta}$. This is why McDonald (1989) concluded that ocal independence is not an assumption as such, but merely defines the dimensionality of the latent traits. In other words, the local independence principle is formally trivial, if no further conditions are added. In addition to locally independent models, Jannarone (1986) presented conjunctive locally dependent models. Although Suppes and Zanotti (1981) presented a process in which any locally dependent model can be replaced by an equivalent locally independent model and it may be concluded that all latent trait models are in fact locally independent models, this does not mean that the dependent models discussed by Jannarone (1986) do not provide meaningful alternatives to the more traditional independent models. However, dependent models will not be considered in this paper.

The above definition of dimensionality of the latent trait space does not make a distinction between dominant and nondominant dimensions. For example, if some of the latent traits guide the performance on only a few items in a test, then these traits are not as important for the explanation of the total variance as traits that influence all items in that test. An example of such a dominant trait is a 'reading' trait for a verbal test composed of items on various topics. One of the problems encountered in assessing the dimensionality of the items is the existence of such dominant

(major) dimensions together with less dominant (minor) dimensions and this distinction may explain some of the conflicting results found in a number of studies on the assessment of dimensionality.

Another problem with the traditional definition of dimensionality, is that the principle of local independence is really a very strong principle. Goldstein (1980) doubts whether this condition is actually met in real life data. To assess the dimensionality of items all higher order cross-product moments should be considered. Instead, the assessment of dimensionality is often restricted to the inspection of only first order correlations. McDonald (1981,1985) refers to this as a 'weak' form of the principle of local independence. Since multivariate normal variables are mutually independent if, and only if all covariances are zero, this 'weak' form implies the 'strong' form of local independence under multivariate normality.

Both above mentioned problems with the traditional definition have led Stout (1987, 1989) to relax the definition and to propose a more practical term, namely essential dimensionality which is based on the concept of essential independence. His definition of essential independence is as follows. Given a vector of latent traits $\underline{\theta}$, the n responses in $\underline{X}$ of a sample of N examinees are essentially independent, if $\underline{X}$ satisfies

$$D_n(\underline{\theta}) \equiv \sum_{i<j\leq n} \frac{|\ \mathrm{Cov}(x_i, x_j | \underline{\Theta}=\underline{\theta})\ |}{\binom{n}{2}} \to 0 \quad \text{as } n \to \infty \qquad (2)$$

If the average covariance of the responses is small in magnitude, then the responses of a 'sub'population with fixed $\underline{\theta}$'s are said to be essentially independent, asymptotically. The essential dimensionality of an item pool is the minimal dimensionality necessary to satisfy the essential independence principle. This definition of dimensionality is easier to implement than the traditional one. Although all common measures for the assessment of dimensionality are implicitly based on this definition, only the measures based on "residuals" are a direct implementation of this principle.

Generally procedures to assess dimensionality can be grouped into the following four categories:

- Procedures based on formal statistical tests for the fit of a model.
- Procedures based on information theoretic model selection criteria.
- Procedures based on the total amount of explained variance.
- Procedures based on "residuals" after a model has been fitted to the data.

In the following sections each of these procedures and the problems connected with their use, will be described. Finally, some results based on simulated data will be given for each of these procedures.

Formal Statistical Test for the Fit of a Model

Bock and Aitkin (1981) proposed to apply the multiple factor analysis model to dichotomous data. They assumed that an unobserved response process y_{ij}, for person i and item j is a linear function of m normally distributed latent variables $\underline{\theta}_i = [\theta_{i1}, \theta_{i2}, \ldots, \theta_{im}]$ and factor loadings $\underline{\lambda}_j = [\lambda_{j1}, \lambda_{j2}, \ldots \lambda_{jm}]$, i.e.:

$$y_{ij} = \lambda_{j1}\theta_{i1} + \lambda_{j2}\theta_{i2} + \ldots + \lambda_{jm}\theta_{im} + \upsilon_{ij}. \qquad (3)$$

θ_{ik} is the ability of person i on ability dimension k, λ_{jk} is the loading of item j and dimension k and υ_{ij} are error terms; $\upsilon_{ij} \sim MVN(\underline{0}, \Gamma^2)$, where Γ^2 is diagonal matrix with positive elements σ_j^2. In contrast to classical factor analysis for continuous data the unobservable process y_{ij} is connected to the dichotomous response variables x_{ij} by the following mechanism:

$$\begin{cases} \text{if } y_{ij} \geq \gamma_j, \text{ then } x_{ij} = 1 \\ \text{otherwise} \qquad\quad x_{ij} = 0 , \end{cases}$$

where γ_j is a threshold parameter connected with item j.
The probability of answering an item j correctly, given the ability vector $\underline{\theta}_i$ for person i is

$$P(x_{ij}=1|\theta_i) = \Phi\left((\gamma_j - \sum_{k=1}^{m} \lambda_{jk}\theta_{ik})/\sigma_j\right) , \qquad (4)$$

where Φ is the cumulative standard normal distribution. To obtain the parameter estimates, Bock and Aitkin (1981) developed an iterative marginal maximum likelihood estimation procedure based on the EM algorithm due to Dempster, Laird and Rubin (1977). The procedure was implemented in TESTFACT by Wilson, Wood and Gibbons (1984).

If the sample consists of N persons with responses on n items, then there are $s \leq \min(2^n, N)$ distinct response patterns. The estimated joint probability of the response patterns is:

$$L_m = N! \prod_{l=1}^{s} \frac{\tilde{P}_l^{r_l}}{r_l!} , \qquad (5)$$

where r_l is the frequency of response pattern l and $\tilde{P}_l$ is the estimated marginal probability of the response pattern computed from the item parameter estimates. The well known likelihood ratio statistic is:

$$G_m^2 = 2\left[\sum_{l=1}^{s} r_l \ln \frac{r_l}{N} - \sum_{l=1}^{s} r_l \ln \tilde{P}_l\right] , \qquad (6)$$

with $df_m = s - n(m+1) + m(m-1)/2$ degrees of freedom.

By means of this statistic two different test procedures can be carried out to decide the smallest number of factor such that the model fits the data.

The first procedure is based on a sequence of tests of the hypothesis H_{m*}: $m = m^*$ against the alternative $m > m^*$, for a stepwise increase of the number of factors $m^* = 1, \ldots, n$. The procedure stops when

$$G_m^2 \leq \chi^2_{1-\alpha, df_m}. \tag{7}$$

The second procedure is based on Haberman's (1977) result that the difference of two statistics for alternative models is asymptotically chi-square distributed with degrees of freedom equal to the difference of the degrees of freedom corresponding to each of the two statistics. The sequence of tests of the hypothesis H_{m*}: $m = m^*$ against the alternative $m = m^* + 1$ for $m^* = 1, \ldots, n-1$, stops when:

$$G^2_{diff} = G^2_m - G^2_{m+1} \leq \chi^2_{1-\alpha, (n-m)}. \tag{8}$$

This procedure is especially useful when some of the expected cell frequencies are almost zero. This may often be the case when $N < 2^n$.

Both procedures can be carried out by the TESTFACT program and Bock, Gibbons and Muraki (1988) concluded from an empirical comparison that these procedures seem adequate in

selecting the number of factors. There are, however, some problems connected with these test procedures.

The first problem is that of the asymptotic distribution. Kendall (1977) has pointed out that Pearson's chi-square and the likelihood ratio statistic are often regarded as equivalent because of their asymptotic properties. In practice, however, the results may be very different. Although maximum likelihood estimators of parameters retain standard large sample properties under various conditions, the large sample approximation of chi-square statistics are often unacceptable. Moreover, the asymptotic test for the fit of a model is generally very sensitive to sample size. For relatively small samples the procedure is unreliable, because of the small expected cell frequencies and for large samples the procedure almost always favours the alternative hypothesis (McDonald, 1989).

A second problem is the choice of the level of significance. The adequate selection of the number of factors is in fact a multiple decision problem. In general, a whole family of hypotheses $\Omega \equiv \{ H_i | i \in I \}$ can be considered, where I is an index set. Corresponding with each hypothesis H_i there is a real valued statistic G_i^2. The family of statistics $G \equiv \{ G_i^2 | i \in I \}$ and the family of hypotheses Ω together form a a so-called testing family $[\Omega , G]$. The overall probability α of making type I errors is unknown, and some sort of simultaneous test procedure (STP) will be needed to control α for the whole testing family. Although a likelihood ratio based STP is available for the testing of association in a

contingency table (Gabriel, 1969), there is no STP available yet for the fit of a multidimensional IRT model. Some sort of correction, like the one proposed by Aitkin (1979) may be useful. His correction is based on the fact that the upper limit of the overall α for n tests with level of significance α_i is $\{1 - (1-\alpha_i)^n\}$.

A third problem is the coherence of the procedure. The hypotheses H_i for the testing of the dimensionality of the latent trait space are ordered. A two-dimensional model, for example, will explain more variance than a one-dimensional model. If the hypothesis of unidimensionality of the latent space H_1 is accepted, a coherent test procedure will also accept the hypothesis H_2 that there are two dimensions. Gabriel (1969) proved that for likelihood ratio based statistics G_i^2 the testing family is monotone, i.e. whenever two hypotheses H_i and H_j are ordered (nested) such that $i<j$, the numerical relation $G_i^2 \geq G_j^2$ holds for $\{i,j \in I\}$. Increasing the number of factors in the model will decrease the values of the likelihood ratio statistics for the fit of the model. Note that tests based on the union intersection principle are also monotone, but that Pearson's chi-square statistic is not (Gabriel,1969).

The monotonicity of the testing family, however, does not necessarily indicate that the testprocedure is coherent. Coherence prevents the contradiction of rejecting a model with say m factors without also rejecting all models with less than m factors.

To illustrate the above mentioned problems the results for the Bock and Lieberman (1970) LSAT-7 data are given in Table 1. It seems difficult to draw conclusions. For a significance level $\alpha_i = 0.10$, the results give evidence of a two factor model. If the relative large sample size is taken into account and a smaller α_i is selected, $\alpha_i = 0.01$, then a one factor model seems more appropriate. The non-coherence of the test procedure is demonstrated by the fact that the p-values do not always increase as the number of factors increases. Although this example does not actually show an incoherent decision, it can be seen that the possibility exists.

Insert Table 1 about here

A Model Selection Criterion

Since we are dealing with a model identification problem and inferences are based on maximum likelihood, it is suggested to assess the dimensionality of the latent trait space by a model selection criterion which does not have the above mentioned disadvantages of the formal test procedures.

Akaike (1974) developed an information theoretic criterion for the identification of optimal and parsimonious

models in data analysis. Akaike's information criterion for model (4) takes the form of a maximized likelihood and is defined as:

$$AIC(m) = -2 \ln[\, L_m(\hat{\underline{\theta}}, \hat{\Lambda}, \hat{\sigma}_j, \hat{\gamma}_j)\,] + 2\, k_m, \qquad (9)$$

where $L_m(\hat{\underline{\theta}}, \hat{\Lambda}, \hat{\sigma}_j, \hat{\gamma}_j)$ is the maximized likelihood and k_m is the number of independent parameters in the model. (-AIC(m)/2) is an unbiased estimate of the mean expected log likelihood. The first term of AIM(m) is a measure of badness of fit. The term $2k_m$ is the penality term correcting for overfitting due to the increasing bias in the first term as the number of parameters in the model increases. The model with the minimum AIC(m) value is chosen to be the best fitting model. Bozdogan and Ramirez (1988) and Ichikawa (1988) showed that the criterion worked quite good in assessing the number of factors in ML factor analysis. A review of the general theory is given by Akaike (1987) and Bozdogan (1987).

To assess the dimensionality, it will be convenient to formulate a criterion comparing the likelihood of the model with m factors with that of the saturated model. The log of the maximized likelihood for the normal ogive model is:

$$\ln (L_m) = \ln N! + \sum_{l=1}^{s} r_l \ln \tilde{P}_l - \sum_{l=1}^{s} \ln (r_l!). \qquad (10)$$

The number of independent estimated parameters is $n(m+1)-m(m-1)/2$. Subtraction of this expression from that of the log of the likelihood for the saturated model with $s \leq \min(2^n, N)$ different response patterns, results in the following AIC expression:

$$AIC(m) = 2\left[\sum_{l=1}^{s} r_l \ln \frac{r_l}{N} - \sum_{l=1}^{s} r_l \ln \tilde{P}_l\right] - 2(df_m). \quad (11)$$

A minimum value for the AIC(m) criterion will indicate the 'true' dimensionality. In Table 1 the AIC(m) values are given for the LSAT-7 data. These values show that a 2-factor normal ogive model would probably be the best fitting model.

Procedures Based on the Proportion of Explained Variance

The determination of the number of components in principal component analysis and in factor analysis is often based on the amount of explained variance, i.e. based on criteria formulated on the eigenvalues of a correlation matrix. Among these procedures are the well known Kaiser's (1960) eigenvalue greater than 1.0 rule, the scree test of Cattell (1966) and the parallel analysis method (Horn, 1965). A huge amount of research has been done on these methods. Although there are sometimes conflicting conclusions, the main trend seems to be that parallel analysis and the scree test perform quite good for continuous data (Zwick & Velicer, 1986).

Similar methods have also been proposed for dichotomous data. Collins, Cliff, McCormick and Zatkin (1986) investigated the modified scree test on phi and tetrachoric correlations and concluded that both procedures perform poorly. Although they prefer the analysis with phi-correlations, Green (1983) and Hambleton and Rovinelli (1986) among others, found that phi coefficients produce spurious factors based on the difficulty level of the binary items. On the other hand, Drasgow and Lissak (1983) found that a modified parallel analysis was quite good in detecting the unidimensionality of dichotomous data. Tucker, Humphreys, Lloyd and Roznowski (1986) compared some eigenvalue indices and found that these indices did not work very well. Recently Bernstein and Teng (1989) questioned the application of eigenvalue criteria to categorical data. They concluded that false evidence of multidimensionality is often found. These results do not seem to be very encouraging. Moreover, there are some problems in applying eigenvalue criteria to dichotomous data.

The first problem with these criteria is the choice of correlation coefficient. Phi coefficients generally produce a positive definite correlation matrix and tend to avoid the problem of Heywood cases. On the other hand, phi coefficients tend to overestimate the number of underlying dimensions. The use of tetrachoric correlations will produce a more reliable estimate of the underlying dimensions, but the sample based estimate of the correlation matrix is often not positive definite and tends to produce more Heywood cases. Although

fast and generally accurate computational algorithms exist (Divgi, 1979) the coefficients will become unstable when extreme values are reached. Moreover, the use of tetrachoric correlations is inappropriate when ability distributions are not normal and the item response function is not normal ogive (Lord, 1980). Of course, the problem of non positive definiteness can be avoided by smoothing the correlation matrix. But this may change the pattern of the eigenvalues, especially when n is relatively small.

Another problem which is not generally recognized, is that the pattern of the eigenvalues depends on the size of the discrimination parameters of the items. Under the assumptions for model (3), that the abilities $\underline{\theta}_i \sim MVN(\underline{0}, I)$ and the errors $\underline{v}_j \sim MVN(\underline{0}, \Gamma^2)$, the covariance matrix of the unobservables y_{ij} is:

$$\Sigma_m = \Lambda\Lambda' + \Gamma^2. \qquad (12)$$

The nxm matrix of factor loadings is Λ and Γ^2 is a diagonal matrix with unique variances σ_j^2. Under multivariate normality the population correlation matrix is:

$$R = [Diag(\Lambda\Lambda')+\Gamma^2]^{-1/2}\,[\Lambda\Lambda'+\Gamma^2]\,[Diag(\Lambda\Lambda')+\Gamma^2]^{-1/2}. \qquad (13)$$

Equation (13) can also be written as:

$$R = [I+ Diag(AA')]^{-1/2}[I+ AA']\,[I+Diag(AA')]^{-1/2}, \qquad (14)$$

where A is an nxm discrimination parameter matrix and relates to Λ by $A = \Gamma^{-1}\Lambda$. It can be seen, that an increase of values of the discrimination parameters will generally increase the correlations.

There is no analytical expression available for the eigenvalues of an arbitrary correlation matrix, but from a theorem due to Gersgorin (see Pullman, 1976) an upper limit of the largest eigenvalue e_{max} of an arbitrary positive definite matrix can be derived:

$$e_{max} \leq \max\left(\sum_{j=1}^{n} |r_{ij}| \ , \ i=1,\ldots,n\right) \ , \qquad (15)$$

where r_{ij} is an element of R. This upper limit is generally a good approximation of the largest eigenvalue of a correlation matrix (Morrison, 1976). Increasing the discrimination values will increase the correlations among the items and will increase the largest eigenvalue. Since $\Sigma_i e_i = n$, increase of the largest eigenvalue will generally decrease the other n-1 eigenvalues.

In conclusion, the pattern of the eigenvalues of the correlation matrix is influenced by the size of the discrimination parameters of the items and a procedure based on the pattern of the eigenvalues may lead to erroneous conclusions about the dimensionality of the latent trait space.

Procedures Based on Residuals after Fitting the Model

One of the conclusions drawn by Hattie (1984) and Hambleton & Rovinelli (1986) was that residuals obtained from nonlinear factor analysis could very well determine the correct dimension of a latent trait space and McDonald (1981,1989) recommended the use of the mean (absolute) residuals.

If Λ_m is the nxm estimated matrix of factor loadings from a solution with m estimated common factors, and R is the tetrachoric correlation matrix, then the off-diagonal elements of the matrix:

$$R^* = R - \Lambda_m \Lambda_m' \quad , \tag{16}$$

are the residuals r_{ij}^*. The mean squared and mean absolute residuals are:

$$f_1 = 2[n(n-1)]^{-1} \sum_{i<j} \sum (r_{ij}^*)^2$$

and

$$f_2 = 2[n(n-1)]^{-1} \sum_{i<j} \sum |r_{ij}^*| \quad , \tag{17}$$

respectively. Since the squared residuals f_1 are generally more sensitive to outliers than absolute residuals f_2, only the results for f_2 will be reported in this study.

It has already been mentioned, that the mean absolute residuals is an implementation of the definition of essential dimensionality by Stout (1987, 1989). The only problem in

applying this criterion is that it may not be clear when the value of the criterion is small enough. A possibility is to compare the f_2 criterion after the fit of an m-dimensional model with values of this criterion from random data.

Conditions for the Simulation Study

To compare the performances of the different measures binary data matrices were generated for known difficulty and discrimination parameters. Though several values for these parameters were considered, the results of only two matrices with different discrimination parameter values will be given. The two matrices are:

$$A_I = \begin{bmatrix} 1.0 & 0.5 & 0.0 \\ 1.0 & 0.0 & 0.5 \\ 0.0 & 0.5 & 1.0 \end{bmatrix} \begin{matrix} (5x) \\ (5x) \\ (5x) \end{matrix} \qquad A_{II} = \begin{bmatrix} 2.0 & 2.0 & 0.0 \\ 0.0 & 2.0 & 2.0 \\ 2.0 & 0.0 & 2.0 \end{bmatrix} \begin{matrix} (5x) \\ (5x) \\ (5x) \end{matrix}$$

The notation in brackets (5x) indicates that groups of 5 items have the same pattern. The sample sizes were set equal to N = 250 and N = 500, and the number of items was n = 15. Each group of 5 items with the same discrimination parameter values has difficulty parameter values -2, -1, 0, 1 and 2. Generation of data from a one-dimensional and a two-dimensional latent trait model was done by using the first and the first two columns of matrices A_I and A_{II}, respectively. The actual generation of the binary data for the multidimentional IRT model was done by using the random number generator of the NAG (1984) program library and the

sample tetrachoric correlations were computed by a modification of the Divgi (1979) procedure.

Results

For almost all conditions of the simulation study the G^2_m test procedure erroneously favoured the alternative hypotheses. The results for the G^2_{diff} test procedure are given in Tables 2 and 3.

These results indicate that the test procedure is often not capable of locating the correct dimension of the latent space. The results in Table 2 show that the test procedure is not capable of locating the minor third dimension. It must be emphasized that these results are based on only 10 simulated data matrices per condition and that more simulations will be needed to draw firm conclusions. The huge amount of computer time needed by TESTFACT prevented us from expanding these simulations.

Insert Tables 2 and 3 about here

The results for the AIC criterion are given in Tables 4 and 5. Although there are misspecifications, it can be seen that the AIC criterion tends to locate the correct dimension of the latent space somewhat better than the χ^2_{diff} procedure.

Again it must be emphasized, that the number of simulated runs is small, due to the large amount of CPU time needed to run TESTFACT.

Insert Tables 4 and 5 about here

Figures 1 and 2 show the eigenvalue patterns for the two discrimination parameter matrices A_I and A_{II}. The points in the Figures indicate the means of the eigenvalues for 50 runs and the vertical lines from the plotted points indicate the range of the eigenvalues. The almost horizontal line is the eigenvalue pattern for random data matrices. The Figures show that only parallel analysis seems adequate, although the minor third dimension in A_I is not recognized.

Insert Figures 1 and 2 about here

Residuals have been obtained from various estimation procedures. In the Figures 3 through 8 the results are given for the common factor analysis procedure MINRES performed on the matrix of tetrachoric correlations, the residuals obtained from TESTFACT and the residuals obtained after an unweighted least squares approximation of pairwise proportions due to McDonald (1985) and computed by NOHARMII.

The computer program MAXLOG, with joint ML estimation was also considered. The residuals from MAXLOG, however, were not very adequate, probably because of the drift of the discrimination parameters.

Insert Figures 3 through 8 about here

The pattern of the residuals show a relatively large drop of f_2 after a model with the correct dimension of the latent traits is fitted to the data. Comparison of the f_2 values with those from random data indicate that all procedures perform quite well and that residuals from NOHARMII even were able to discover the minor third dimension in A_I.

Summary and Conclusions

The assessment of dimensionality of the latent traits is a difficult problem and each of the discussed measures has it own disadvantages. In this paper the assessment of dimensionality is approached from the factor analytical point of view. Although the results from this study do not indicate that one of the measures performs best in assessing the dimensionality of a latent space, the results seem to

indicate that some of the measures are not very good. The following conclusions may be drawn from this study:

- Asymptotic χ^2 tests for the fit of a latent trait model are not very reliable. The AIC criterion seems to perform somewhat better, but more research will be needed to draw definite conclusions.
- The structure of the matrix with discrimination parameters is crucial for the performance of the eigenvalue measures, because the size of the discrimination parameter values determines the explained variance of each factor in the model.
-The disadvantages of the eigenvalue criteria and their performance in this study lead to the conclusion that these measures should be avoided. Parallel analysis, however, did perform quite good.
- The "residual" measures from TESTFACT, MINRES and NOHARMII performed quite good, and even a minor dimension could be located by NOHARMII.

Finally, it must be noted, that the conditions of this study were limited and that the results from TESTFACT were based on a small number of runs. More simulations will be needed to drawn more definite conclusions.

References

Aitkin, M. (1979). Simultaneous test procedure for contingency table models. Applied Statistics, 28, 233-242.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on Automatic Control, AC-19, 716-723.

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

Bernstein, I.H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. Psychological Bulletin, 105, 467-477.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The genral Theory and its analytical extensions. Psychometrika, 52, 345-370.

Bozdogan, H., & Ramirez, D.E. (1988). FACAIC: Model selection algorithm for the orthogonal factor model using AIC and CAIC. Psychometrika, 53, 407-415.

Cattell, R.B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Collins, L.M., Cliff, N., McCormick, D.J., & Zatkin, J.L. (1986). Factor recovery in binary data sets. A simulation. Multivariate Behavioral Research, 3, 377-392.

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

Divgi, D.F. (1979). Calculation of tetrachoric correlation coefficient. Psychometrika, 40, 5-32.

Drasgow, F., & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Fraser, C. (1988). NOHARMII. A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioural Studies.

Gabriel, K.R. (1969). Simultaneous test procedures - Some theory of multiple comparisons. The Annals of Mathematical Statistics, 40, 221-250.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

Green, S.B. (1983). Identifiability of spurious factors using linear factor analysis with binary data. Applied Psychological Measurement, 7, 139-147.

Haberman, S.J. (1977). Log-linear models and frequency tables with small expected cells 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.

Harman, H.H., & Jones, W.H. (1966). Factor analysis by minimizing residuals (MINRES). Psychometrika, 31, 351-368.

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

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

Horn, J.L. (1965). A rational and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Ichikawa, M. (1988). Empirical assessments of AIC procedure for model selection in factor analysis. Behaviormetrika, 24, 33-40.

Jannarone, R.J. (1986). Conjunctive item response theory kernals. Psychometrika, 51, 357-373.

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

Kendall, M.G. (1977). Multivariate contingency tables and some further problems in multivariate analysis. In: P.R. Krishnaiah (Ed.). Multivariate Analysis IV. North Holland Publ. Comp., 483-494.

Knol, D.L., & Berger, M.P.F. (1990). Empirical Comparison between factor analysis and multidimensional item response models. Multivariate Behavioral Research. To appear.

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

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

McDonald, R.P. (1985). Unidimensional and multidimensional models for item response theory. In D.J. Weiss (ed.), Proceedings of the 1982 Computorized Adaptive Testing Conference. Minn: University of Minnesota, 127-148.

McDonald, R.P. (1989). Future directions for item response theory. International Journal of Educational Research, 13, 2, 205-220.

McKinley, R.L., & Reckase, M.D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. Behavior Research Methods & Instrumentation, 15, 389-390.

Morrison, D.F. (1976) Multivariate statistical methods, New York: McGraw-Hill.

Muthén, B. (1978). Contributions to factor analysis of dichotomized variables. Psychometrika, 43, 551-560.

NAG (1984) Library (Mark 11) Oxford, UK:Numerical Algorithms.

Pullman, N.J. (1976). Matrix Theory and its Applications. New York: Marcel Dekker, Inc.

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

Stout, W.F. (1989). A Nonparametric multidimensional IRT approach with applications to ability estimation. Psychometrika.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? Synthese, 18, 191-199.

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

Tucker, L.R., Humphreys, L.S., Lloyd, G., & Roznowski, M.A. (1986). Comparative accuracy of five indices of dimensionality. Urbana: University of Illinois, Department of Psychology.

Wilson, D.T., Wood, R., & Gibbons, R.T. (1984). TESTFACT: Test scoring, item statistics, and factor analysis. Mooresville, IN:Scientific Software.

Zegers, F.E., & Ten Berge, J.M.F. (1983). A fast and simple computational method of minimum residual factor analysis. Multivariate Behavioral Research, 18, 331-340.

Zwick, W.R., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

Table 1

Chi-square statistics and AIC criterion for LSAT-7 data

(N=1000, n=5, quadrature points=7)

m	G^2_m	df	p	G^2_{diff}	df	p	AIC(m)/2
1 factor	31.72	21	0.064				10.72
2 factors	22.76	17	0.157	8.96	4	0.062	5.76
3 factors	21.45	14	0.090	1.31	3	0.727	7.45

Table 2

Percentages of decisions about the dimensionality of the latent trait space for A_I with the G^2_{diff} procedure (α_i=0.05)

N=250

decision	Dimension of latent trait space: 0	1	2	3
0	-	-	-	-
1	90	80	60	-
2	10	20	40	100
3	-	-	-	-

N=500

decision	Dimension of latent trait space: 0	1	2	3
0	-	-	-	-
1	80	60	20	-
2	20	40	80	90
3	-	-	-	10

Table 3

Percentages of decisions about the dimensionality of the latent trait space for A_{II} with the G^2_{diff} procedure (α_i=0.05)

N=250

decision	Dimension of latent trait space 0	1	2	3
0	-	-	-	-
1	80	100	10	-
2	20	-	80	10
3	-	-	10	90

N=500

decision	0	1	2	3
0	-	-	-	-
1	90	100	20	-
2	10	-	80	60
3	-	-	-	40

Table 4

Percentages of decisions about the dimensionality of the latent trait space for A_I with the AIC(m) criterion

		Dimension of latent trait space			
N=250		0	1	2	3
decision	0	-	-	-	-
	1	90	80	30	-
	2	10	20	50	40
	3	-	-	20	60
N=500					
decision	0	-	-	-	-
	1	100	80	20	-
	2	-	20	70	40
	3	-	-	10	60

Table 5

Percentages of decisions about the dimensionality of the latent trait space for A_{II} with the AIC(m) criteria

N=250

		Dimension of latent trait space			
		0	1	2	3
decision	0	-	-	-	-
	1	100	90	-	-
	2	-	10	90	-
	3	-	-	10	100

N=500

decision	0	-	-	-	-
	1	90	80	10	-
	2	10	20	80	40
	3	-	-	10	60

Figure Captions

Figure 1. Eigenvalue plots for matrix of discrimination parameters A_I.

Figure 2. Eigenvalue plots for matrix of discrimination parameters A_{II}.

Figure 3. Plotted mean absolute residuals for A_I and MINRES.

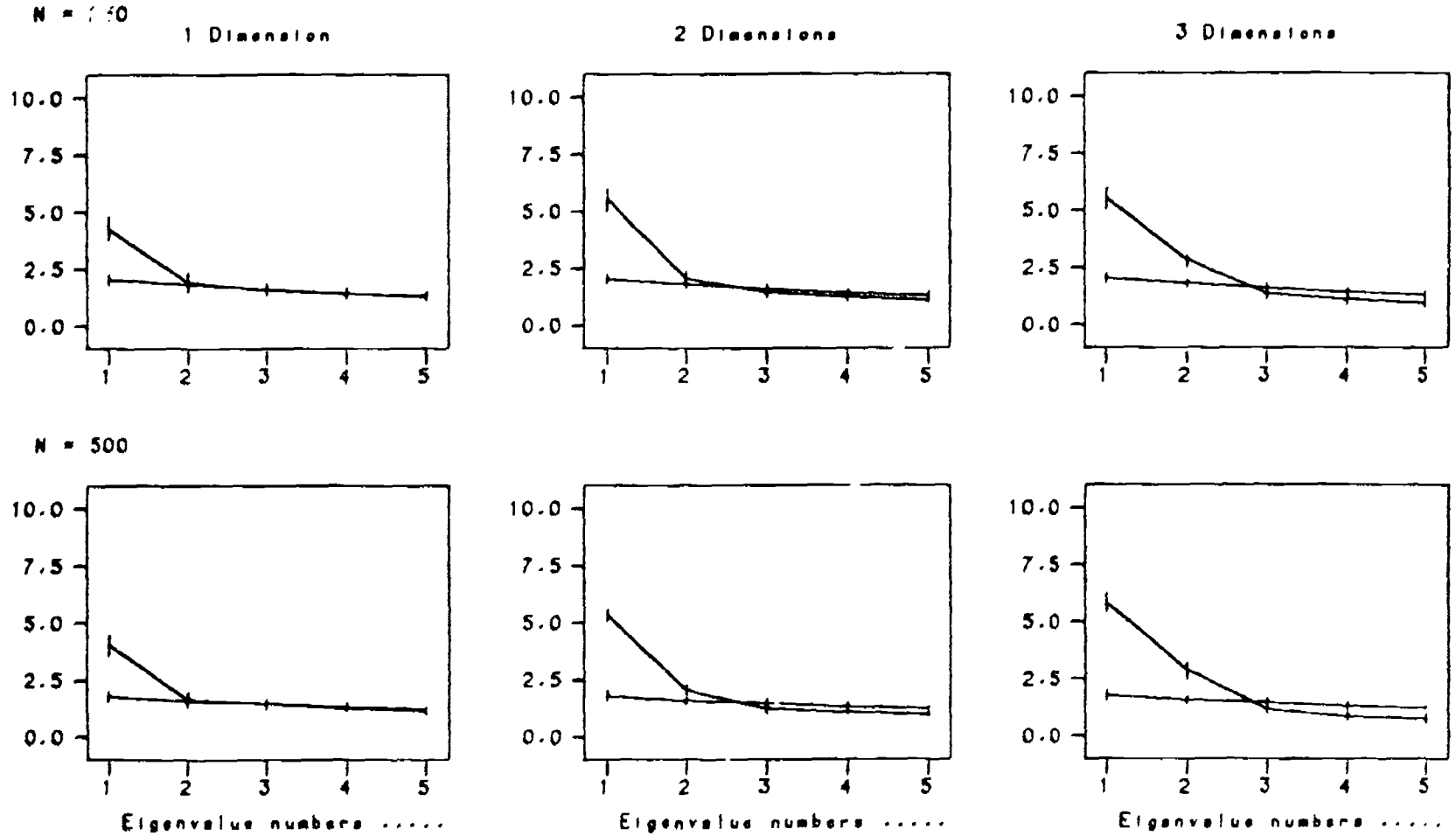
Figure 4. Plotted mean absolute residuals for A_{II} and MINRES.

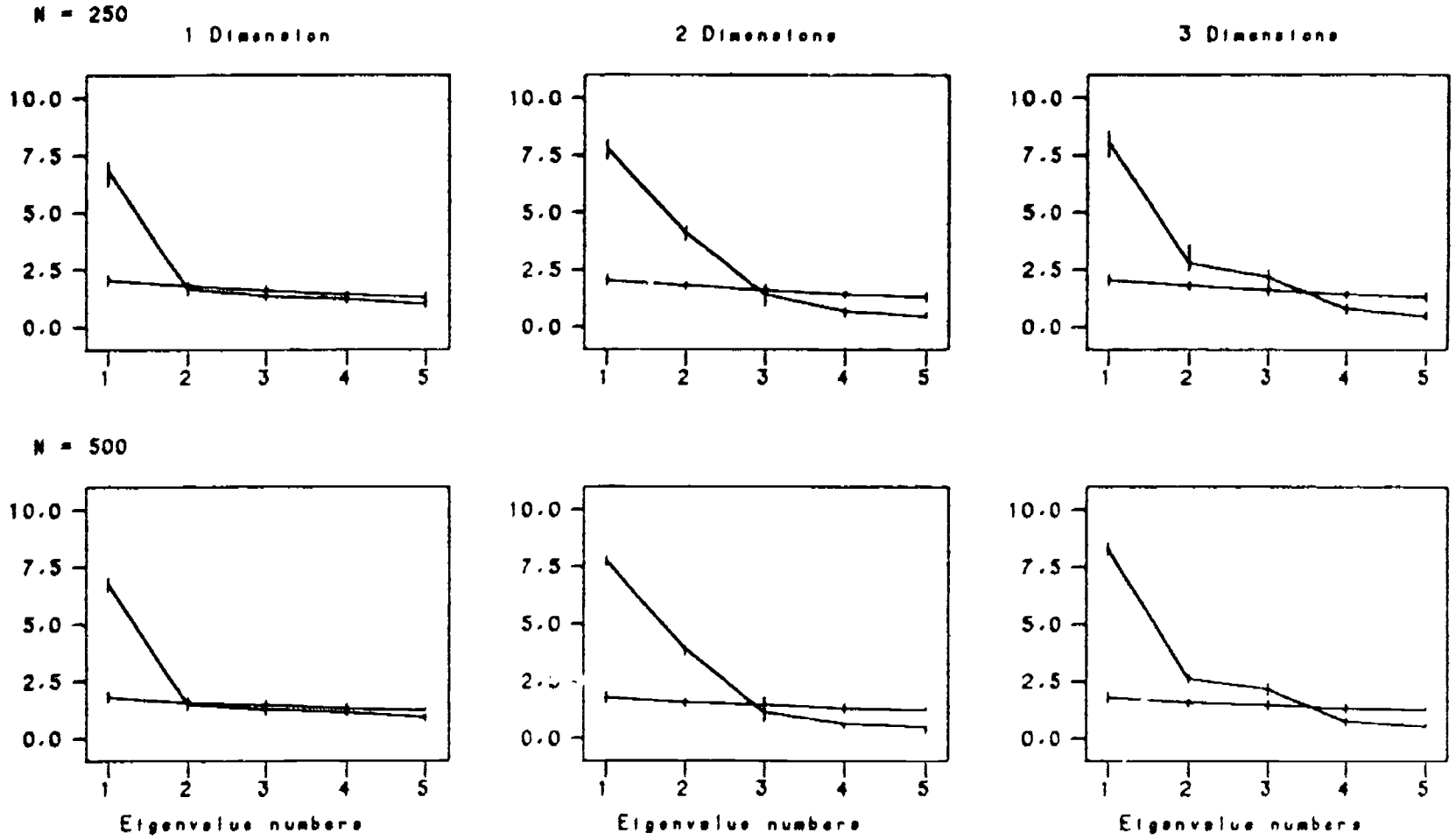
Figure 5. Plotted mean absolute residuals for A_I and NOHARMII.

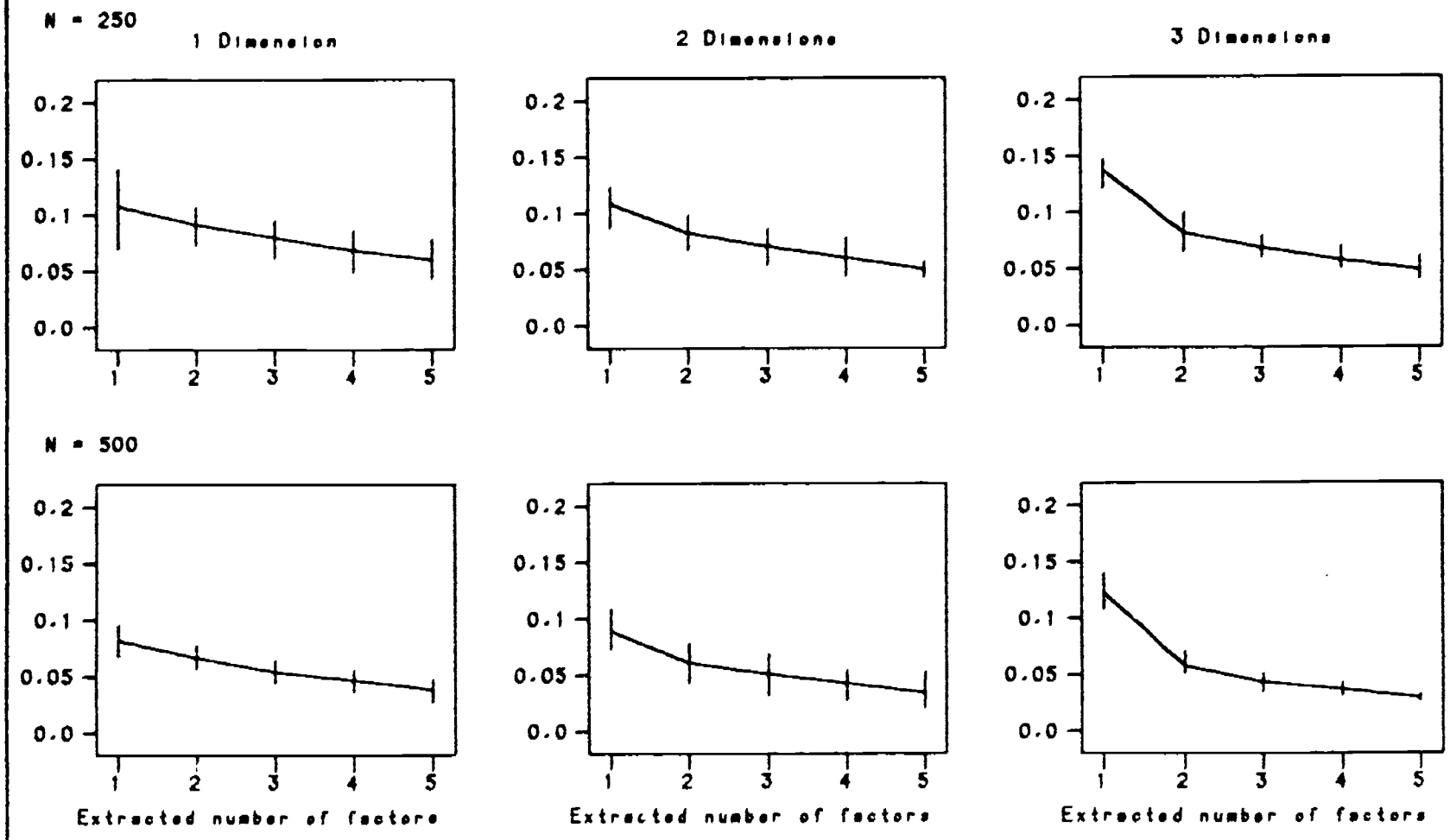
Figure 6. Plotted mean absolute residuals for A_{II} and NOHARMII.

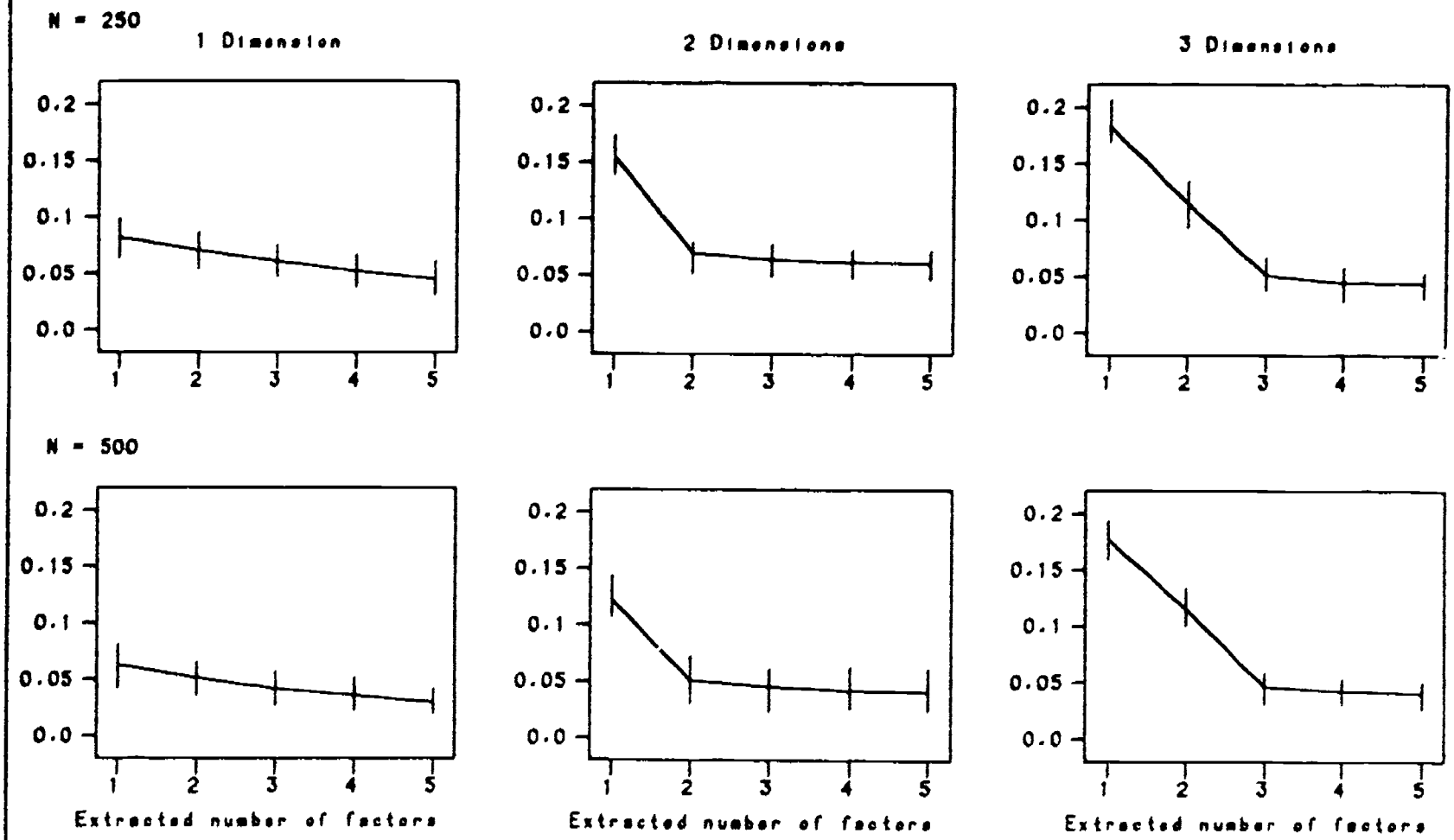
Figure 7. Plotted mean absolute residuals for A_I and TESTFACT.

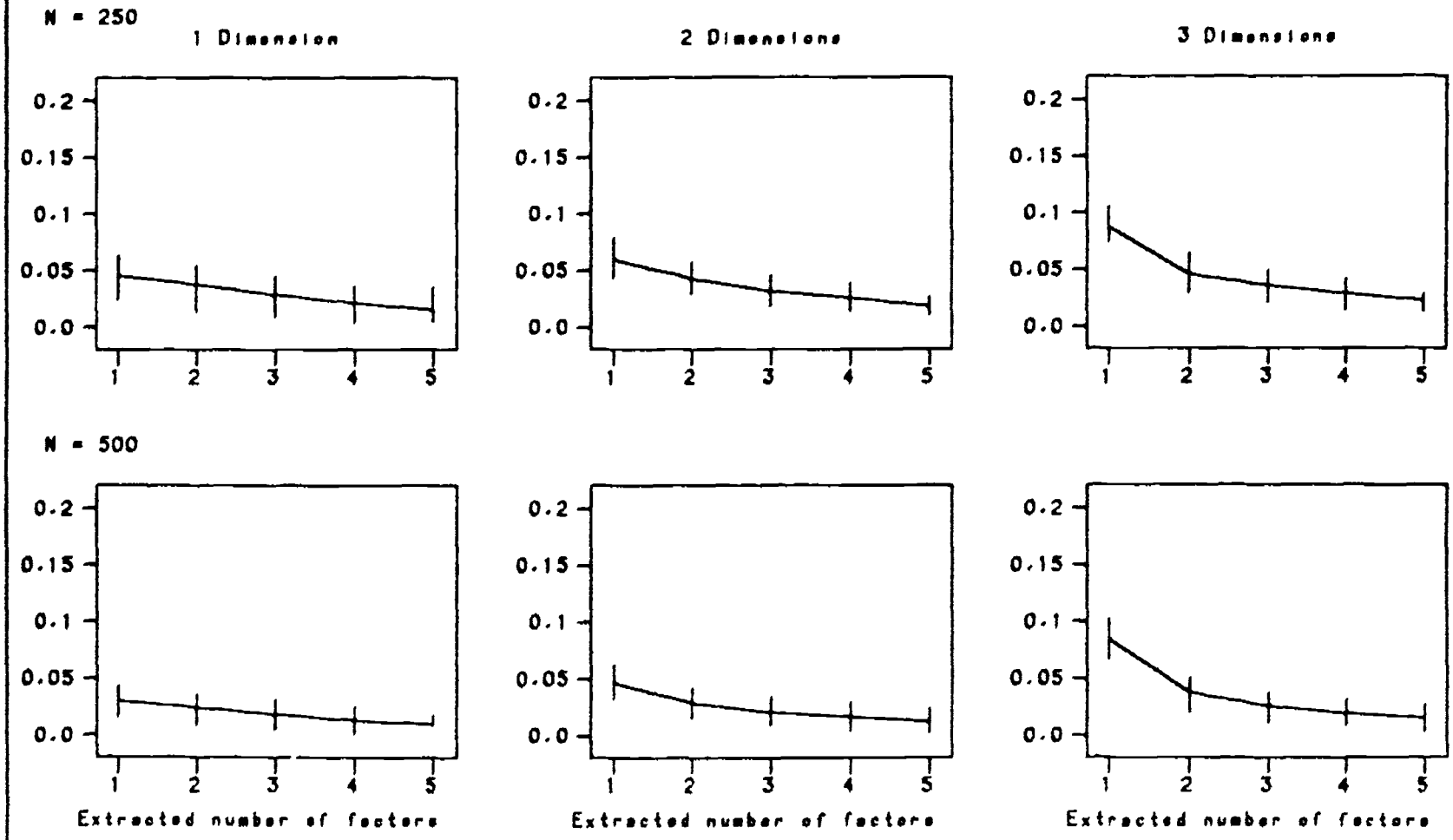
Figure 8. Plotted mean absolute residuals for A_I and TESTFACT.

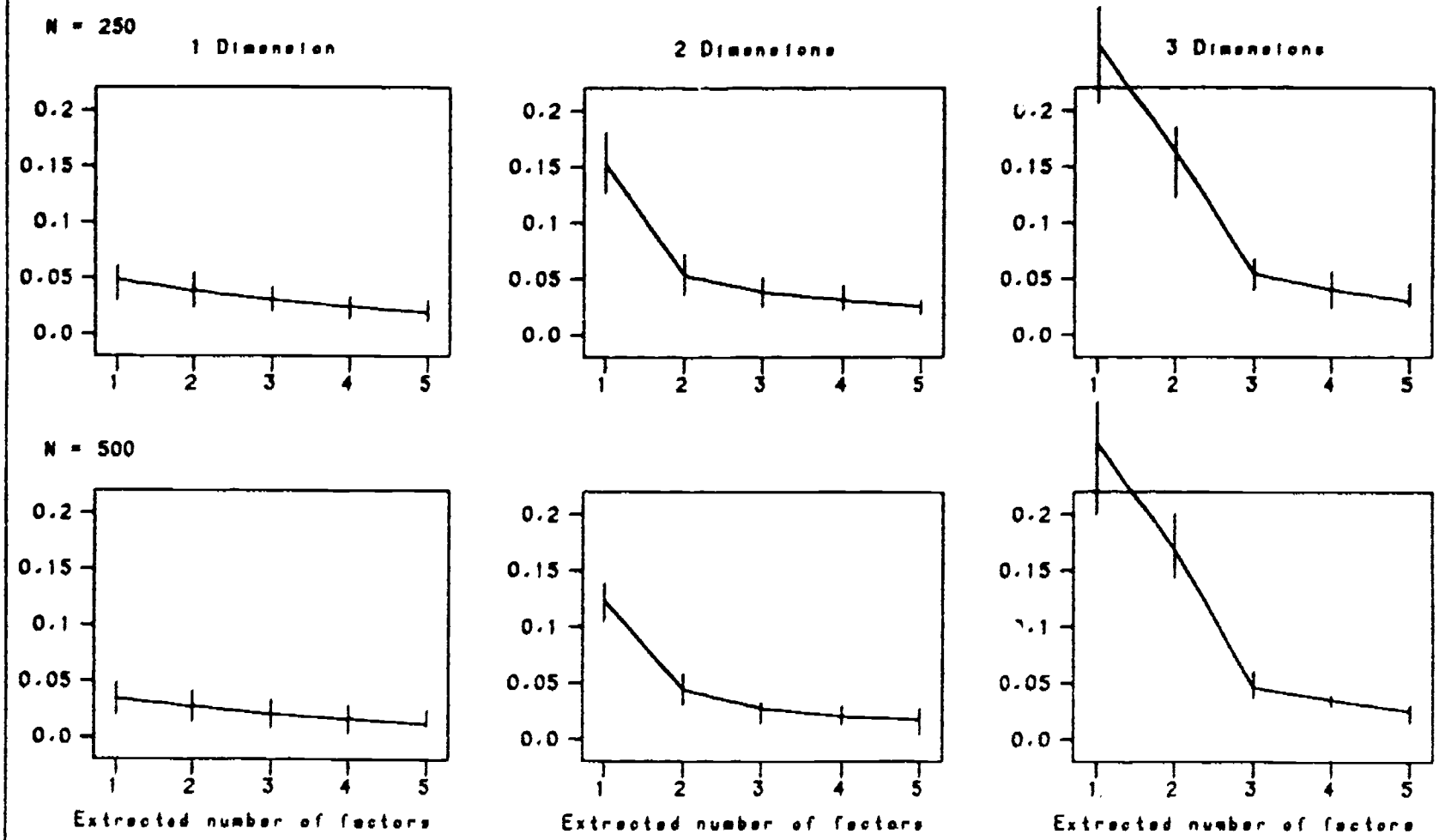

N = 150
1 Dimension
2 Dimensions
3 Dimensions
N = 500
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Eigenvalue numbers .....

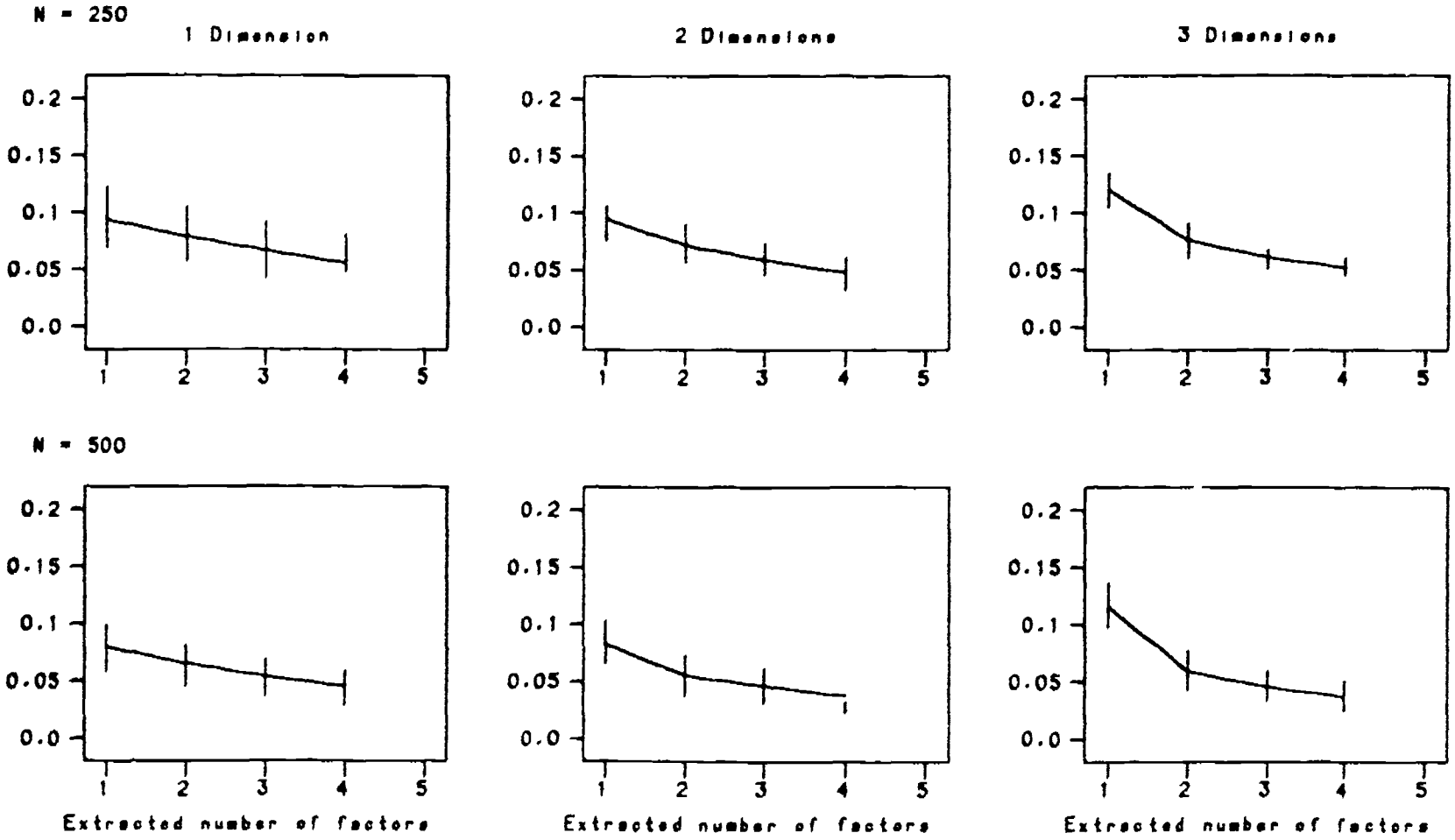

N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Eigenvalue numbers

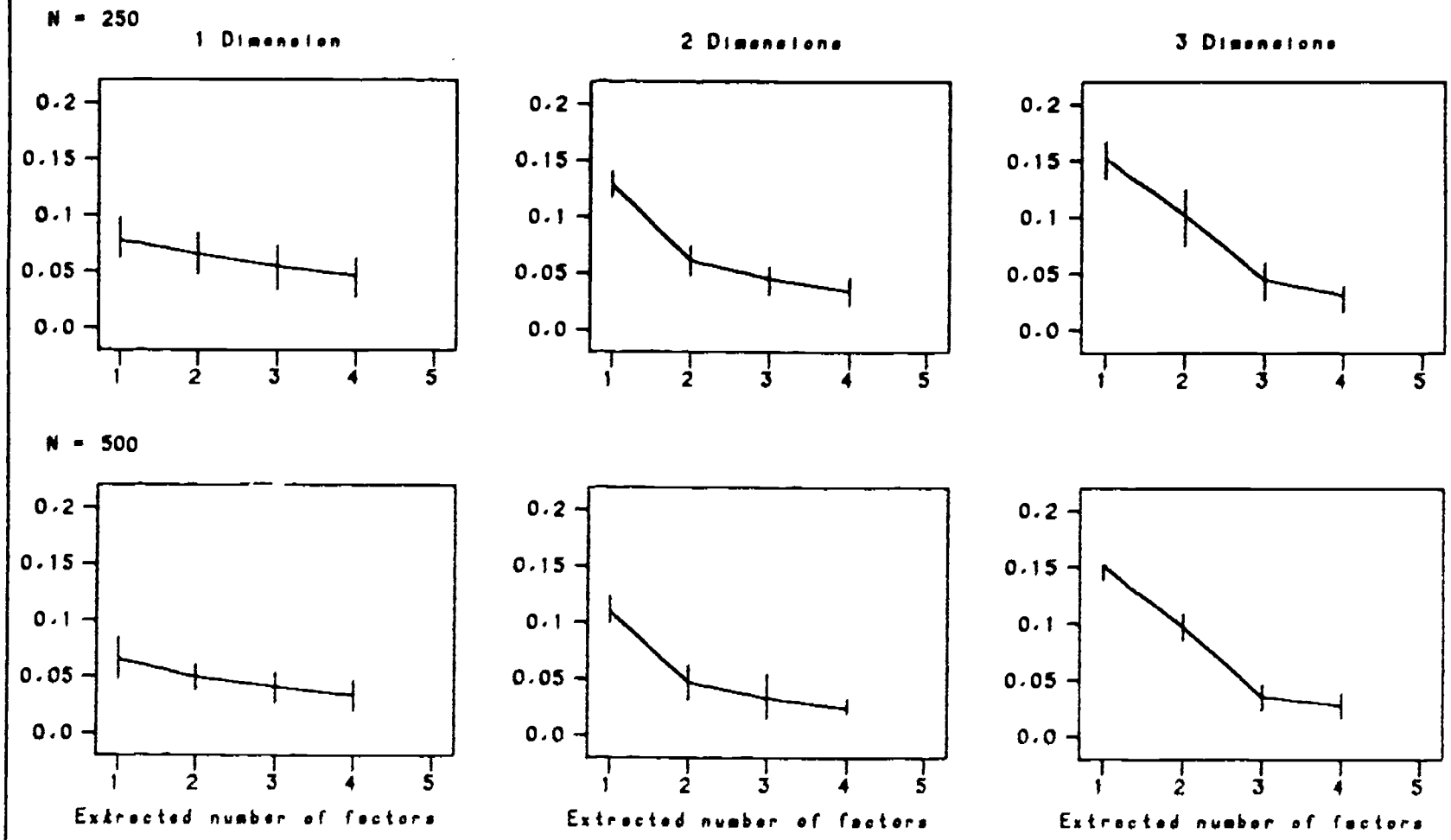

N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors


N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors


N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors


N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors


N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors


N = 250
1 Dimension
2 Dimensions
3 Dimensions
N = 500
0.2
0.15
0.1
0.05
0.0
1
2
3
4
5
Extracted number of factors

Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-3 H.J. Vos, *Simultaneous Optimization of Classification Decisions Followed by an End-of-Treatment Test*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

RR-89-6 J.J. Adema, *Implementations of the Branch-and-Bound method for test construction problems*

RR-89-5 H.J. Vos, *A simultaneous approach to optimizing treatment assignments with mastery scores*

RR-89-4 M.P.F. Berger, *On the efficiency of IRT models when applied to different sampling designs*

RR-89-3 D.L. Knol, *Stepwise item selection procedures for Rasch scales using quasi-loglinear models*

RR-89-2 E. Boekkooi-Timminga, *The construction of parallel tests from IRT-based item banks*

RR-89-1 R.J.H. Engelen & R.J. Jannarone, *A connection between item/subtest regression and the Rasch model*

RR-88-18 H.J. Vos, *Applications of decision theory to computer based adaptive instructional systems*

RR-88-17 H. Kelderman, *Loglinear multidimensional IRT models for polytomously scored items*

RR-88-16 H. Kelderman, *An IRT model for item responses that are subject to omission and/or intrusion errors*

RR-88-15 H.J. Vos, *Simultaneous optimization of decisions using a linear utility function*

RR-88-14 J.J. Adema, *The construction of two-stage tests*

RR-88-13 J. Kogut, *Asymptotic distribution of an IRT person fit index*

RR-88-12 E. van der Burg & G. Dijksterhuis, *Nonlinear canonical correlation analysis of multiway data*

RR-88-11 D.L. Knol & M.P.F. Berger, *Empirical comparison between factor analysis and item response models*

RR-88-10 H. Kelderman & G. Macready, *Loglinear-latent-class models for detecting item bias*

RR-88-9 W.J. van der Linden & T.J.H.M. Eggen, *The Rasch model as a model for paired comparisons with an individual tie parameter*

RR-88-8 R.J.H. Engelen, W.J. van der Linden, & S.J. Oosterloo, *Item information in the Rasch model*

RR-88-7 J.H.A.N. Rikers, *Towards an authoring system for item construction*

RR-88-6 H.J. Vos, *The use of decision theory in the Minnesota Adaptive Instructional System*

RR-88-5 W.J. van der Linden, *Optimizing incomplete sample designs for item response model parameters*

RR-88-4 J.J. Adema, *A note on solving large-scale zero-one programming problems*

RR-88-3 E. Boekkooi-Timminga, *A cluster-based method for test construction*

RR-88-2 W.J. van der Linden & J.J. Adema, *Algorithmic test design using classical item parameters*

RR-88-1 E. van der Burg & J. de Leeuw, *Nonlinear redundancy analysis*

Research Reports can be obtained at costs from Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

[redacted]

[redacted]tment of
[redacted]UCATION

[redacted]cation by
the Department of Education
of the University of Twente

[redacted]de
[redacted]nds