British Journal of Mathematical and Statistical Psychology (2000), **53**, 1–16 *Printed in Great Britain*

Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima

Marieke E. Timmerman* and **Henk A. L. Kiers**

Heymans Institute of Psychology, University of Groningen, The Netherlands

A method that indicates the numbers of components to use in fitting the three-mode principal components analysis (3MPCA) model is proposed. This method, called DIFFIT, aims to find an optimal balance between the fit of solutions for the 3MPCA model and the numbers of components. The achievement of DIFFIT is compared with that of two other methods, both based on two-way PCAs, by means of a simulation study. It was found that DIFFIT performed considerably better than the other methods in indicating the numbers of components.

The 3MPCA model can be estimated by the TUCKALS3 algorithm, which is an alternating least squares algorithm. In a study of how sensitive TUCKALS3 is at hitting local optima, it was found that, if the numbers of components are specified correctly, TUCKALS3 never hits a local optimum. The occurrence of local optima increased as the difference between the numbers of underlying components and the numbers of components as estimated by TUCKALS3 increased. Rationally initiated TUCKALS3 runs hit local optima less often than randomly initiated runs.

1. Introduction

Three-way data are data that can be classified in three ways. An example is scores of a number of subjects on different variables measured on different occasions. Three-mode principal components analysis (3MPCA) (Tucker, 1966) is a method for summarizing three-way data, and is a generalization of standard two-way principal components analysis (PCA). In two-way PCA the data are decomposed into two matrices, namely the component scores matrix and the component loading matrix. In 3MPCA, the three-way data are decomposed into three component matrices, where the numbers of components to be used are not necessarily equal for each component matrix. When the numbers of components are not suggested by the nature of the data, a method is needed to indicate these numbers. In order to choose the numbers of components, Tucker (1966) proposed the application of a method ordinarily used in two-way PCA. However, it is not clear that this method is suitable for use in three-way problems. Therefore, a new method is proposed for indicating the numbers of components in 3MPCA, and this method is compared to two methods ordinarily used in two-way PCA by means of a simulation study.

* Requests for reprints should be addressed to Marieke E. Timmerman, Department of Psychology. University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.

The 3MPCA model is usually fitted to the data by TUCKALS3 which is an alternating least squares algorithm. Unfortunately, this kind of algorithm may end in a local optimum. At the cost of (sometimes much) computational effort, the possibility of missing the global optimum can be reduced by using multiple 'starts' for a single 3MPCA model. Since the new method of determining the numbers of components requires a large number of 3MPCAs, it is useful to examine the necessity of using multiple starts. Therefore, we investigate how sensitive TUCKALS3 is to hitting local optima under different conditions.

The 3MPCA model is defined by

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}, \tag{1}$$

where x_{ijk} denotes the elements of the $I \times J \times K$ three-way array $\underline{\mathbf{X}}$, a_{ip}, b_{jq} and c_{kr} denote the elements of the component matrices $\mathbf{A}$ ($I \times P$), $\mathbf{B}$ ($J \times Q$) and $\mathbf{C}$ ($K \times R$), respectively, g_{pqr} denotes the elements of the so-called core array $\underline{\mathbf{G}}$ ($P \times Q \times R$), and e_{ijk} denotes the elements of the error array $\underline{\mathbf{E}}$ ($I \times J \times K$); $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $p = 1, \ldots P$, $q = 1, \ldots, Q$, $r = 1, \ldots, R$. The element g_{pqr} of the core $\underline{\mathbf{G}}$ denotes the relationship between the components p, q, and r in $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively, $p = 1, \ldots, P$, $q = 1, \ldots, Q$, $r = 1, \ldots, R$. The 3MPCA model can be written in matrix notation as

$$\mathbf{X} = \mathbf{A}\ \mathbf{G}(\mathbf{C}' \otimes \mathbf{B}') + \mathbf{E}, \tag{2}$$

where $\mathbf{X}$, $\mathbf{G}$ and $\mathbf{E}$ denote the three-way arrays $\underline{\mathbf{X}}$, $\underline{\mathbf{G}}$ and $\underline{\mathbf{E}}$, respectively, written as two-way matrices of order ($I \times JK$), ($P \times QR$) and ($I \times JK$), respectively, and $\otimes$ denotes the Kronecker product; specifically, $\mathbf{X}$ is the supermatrix containing the K frontal planes of $\underline{\mathbf{X}}$ next to each other, and $\mathbf{G}$ and $\mathbf{E}$ are obtained analogously from $\underline{\mathbf{G}}$ and $\underline{\mathbf{E}}$, respectively. The data array $\underline{\mathbf{X}}$ to be analysed can be the raw scores array, but usually some kind of preprocessing is applied. The type of preprocessing is an important but difficult issue which cannot be settled at once (Harshman & Lundy, 1984; Kroonenberg, 1985). What preprocessing method is suitable heavily depends on the type of data. In what follows we assume the data array to have been preprocessed adequately.

Tucker (1966) originally proposed fitting the 3MPCA model by performing three PCAs of derived supermatrices, each yielding one of the component matrices. For instance, matrix $\mathbf{A}$ is obtained by PCA applied to $\mathbf{X}$, that is, by minimizing

$$f(\mathbf{A}, \mathbf{F}) = ||\mathbf{X} - \mathbf{A}\mathbf{F}||^2, \tag{3}$$

subject to the constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}$ and where $||.||$ denotes the Eucidean norm. $\mathbf{A}$ is then obtained as the matrix containing the first P eigenvectors, normalized to unit length, of $\mathbf{X}\mathbf{X}'$. The component matrices $\mathbf{B}$ ($J \times Q$) and $\mathbf{C}$ ($K \times R$) are obtained by analogously performing a PCA of a supermatrix of order ($J \times IK$) and ($K \times IJ$), respectively. The component matrices are restricted to being columnwise orthonormal, as can be done without loss of generality. Finally, the core $\underline{\mathbf{G}}$ is obtained by minimization of $||\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C}' \otimes \mathbf{B}')||^2$ over $\mathbf{G}$, hence from

$$\mathbf{G} = \mathbf{A}'\mathbf{X}(\mathbf{C} \otimes \mathbf{B}). \tag{4}$$

Tucker's method of fitting the 3MPCA model has the disadvantage that the estimators have unclear properties. To avoid this problem, an alternating least squares algorithm for fitting the 3MPCA model has been offered by Kroonenberg and De Leeuw (1980). This algorithm,

TUCKALS3, aims to minimize

$$g(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{G}) = ||\mathbf{X} - \mathbf{A}\mathbf{G}(\mathbf{C}' \otimes \mathbf{B}')||^2, \tag{5}$$

where **A**, **B** and **C** are constrained to be columnwise orthonormal. This loss function gives the sum of squares of $\underline{\mathbf{X}}$ that are not explained by the model, and by subtracting this from the total sum of squares of $\underline{\mathbf{X}}$, we get the 'explained sum of squares', which is called the 'fit' here. The algorithm commences with rationally or randomly chosen initial matrices **A**, **B** and **C**, and **G** is obtained from **A**, **B** and **C** according to (4). Subsequently, **A**, **B** and **C**, and implicitly **G** as well, are updated such that they minimize (5) given the other parameters. This process continues until the algorithm converges to a local or global minimum. In practice, the algorithm is considered to have converged if the difference in fit between two subsequent updates is less than a predetermined value, the so-called convergence criterion. As a rational start for **A**, **B** and **C**, one often takes the **A**, **B** and **C** resulting from Tucker's method.

The numbers of components of the three component matrices have to be chosen before fitting the 3MPCA model to the data. We intend to use those numbers of components that cover the most important or most salient aspects of the data, thereby ignoring aspects of little importance (e.g., because they pertain to a small number of subjects or variables). A number of approaches appear to be used in practice for indicating the numbers of components in the 3MPCA model. Love and Tucker (1970) applied the scree-test (Cattell, 1966) to the eigenvalues obtained from the three separate PCAs of the three supermatrices in Tucker's (1966) method. Alternative approaches based on the separate PCAs can be used as well. Such procedures, however, do not take the relations between modes into account. Some authors looked for the best interpretable solution (Dai, 1985; Kroonenberg & De Leeuw, 1980). However, as a 3MPCA model leads to a huge number of possible solutions due to the rotational freedom of every solution of a 3MPCA, it may be difficult to choose between the several possibilities. Kroonenberg, Lammers, and Stoop (1985) took into account the numbers of components and explained sums of squares of several analyses, but their approach is not systematic. Kroonenberg and van der Voort (1987) compared solutions for several models for a particular data set, among them several 3MPCA models, by plotting the residual sums of squares (RSS) against the degrees of freedom (df) associated with these models. The degrees of freedom of a 3MPCA are defined by Weesie and Van Houwelingen (1983) as the number of observations in the data matrix ($= I \times J \times K$) minus the number of free parameters of the model (computed as $I \times P + J \times Q + K \times R + P \times Q \times R - P^2 - Q^2 - R^2$). A solution with a good RSS/df ratio, that is, a low RSS and large number of degrees of freedom, was chosen. However, their approach is not very systematic either, in that it lacks a fully operationalized strategy for choosing one's solution(s).

Here, we propose a new method that is systematic (like the PCA-based scree-test approach) and at the same time does take the relations between modes into account. Furthermore, we examine the performance of this new method compared to two other systematic approaches, both variants of the PCA-based approach. The performance of the three methods is examined by a simulation study. The use of the new method is illustrated by an empirical example.

As mentioned earlier, the TUCKALS3 algorithm may lead to a local rather than a global minimum. By using several differently initiated runs, one can reduce the chance of missing the global minimum. The main disadvantage of using multiple starts is the increase in computation time. The chance of finding a local minimum may, however, depend on how the algorithm is initiated. In the present paper, we investigate the sensitivity of the TUCKALS3

algorithm to hitting local minima. Specifically, by means of a simulation study we examine how often and in which situations rationally and randomly initiated runs of the TUCKALS3 algorithm lead to a local minimum. The same simulated data were used as in the simulation study to look at the effectiveness of the three methods in indicating the numbers of components.

2. Methods for choosing the numbers of components in the 3MPCA model

2.1. Methods based on separate PCAs

In the first type of approach, various methods can be used to indicate the numbers of components in each separate PCA. For instance, Tucker (1966) recommended using the 'scree-test' (Cattell, 1966) and discarding components associated with small eigenvalues. The latter advice is reminiscent of the Kaiser (1960) criterion, which is to retain only those components corresponding to eigenvalues greater than 1, or, more generally, greater than the mean eigenvalue. The scree-test (Cattell, 1966) is based on plotting the magnitude of the eigenvalues in descending order against their ordinal numbers. Often, the magnitude of successive eigenvalues drops sharply until a certain point and then levels off. After the identification of such a drop, only those components are retained that are associated with the eigenvalues in the range of sharp descent, that is, before the first one in the range where they start to level off.

As an alternative procedure for indicating the number of components in PCA, parallel analysis (PA) has been proposed by Horn (1965) as a sample-based adaptation of the population-based Kaiser criterion. Yet another procedure is based on Bartlett's (1950, 1951) chi-square test, where each eigenvalue, ordered from largest to smallest, is excluded sequentially and the null hypothesis that the remaining eigenvalues are equal is tested. If the null hypothesis is not rejected, the components related to the excluded eigenvalues are retained.

A third alternative approach is based on the minimum average partial rule (MAP) (Velicer, 1976; Zwick & Velicer, 1986), which employs a matrix of partial correlations between the variables with components partialed out. The components are partialed out sequentially, starting with the component explaining most of the variance, and each time the average of the squared partial correlations is computed. This average usually first decreases, but then, after having reached a minimum value, starts increasing again. According to the MAP, those components should be retained that were partialed out when the minimum average partial correlation value was attained.

Zwick and Velicer (1986) evaluated the performance of the scree-test, the Kaiser criterion, PA, Bartlett's test and MAP in two-way PCA by means of a Monte Carlo study. They found that PA and MAP were best across all situations. The scree-test, based on judgments of three judges, was generally accurate but appeared to be more sensitive to smaller sample sizes and less saturated components (i.e., components with small loadings) than PA and MAP.

Niesing (1997) used a judge-independent operationalization of the scree-test combined with the Kaiser criterion, which he called the quotient of differences in additional values (QDA) procedure. The QDA consists of finding the value of q that maximizes

$$a(q) = \frac{\lambda_q - \lambda_{q+1}}{\lambda_{q+1} - \lambda_{q+2}}, \tag{6}$$

subject to $\lambda_q >$ mean (λ), where λ_q is the eigenvalue associated with the qth component in a PCA. QDA includes the Kaiser criterion because only those values of q are considered for which $\lambda_q >$ mean(λ). QDA includes an operationalization of the scree-test for the following reason. The explained variance of the qth component corresponds with λ_q. If the $(q + 1)$th component explains much less variance than the qth component, the difference $\lambda_q - \lambda_{q+1}$ will be considerably larger than zero. If further decreases in λ are small, the value of λ_{q+2} will be close to the value of λ_{q+1}, and the difference $\lambda_{q+1} - \lambda_{q+2}$ will be close to zero. Therefore, the number of components q for which the maximum of $a(q)$ is reached corresponds to a value after which the eigenvalues level off, and hence QDA is an operationalization of the scree-test. Niesing (1997) made a comparison between QDA and PA for PCA (and some generalizations) and found that QDA performed significantly better than PA in indicating the number of components to retain in PCA.

We compare the performance of our new method, DIFFIT, described in section 2.2.1, with two methods based on performing separate PCAs of supermatrices in a simulation study. On the basis of the comparisons by Zwick and Velicer (1986) and Niesing (1997), we chose to use QDA and MAP to indicate the number of components in separate PCAs for each of the three modes.

2.2.1. DIFFIT: A systematic approach for choosing the numbers of components based on a series of 3MPCA fit values. A possible disadvantage of applying methods designed for two-way PCA to the three supermatrices is that they disregard the relations between the modes. In order to take these relations into account, it is necessary to consider the 3MPCA solutions of combinations of numbers of components. Kroonenberg and van der Voort (1987) considered the results of 3MPCA solutions for a large number of combinations of numbers of components, but, as mentioned, their approach is not very systematic. Here, we propose a new method, called DIFFIT where an optimal ratio of the DIFFerence in FIT (the explained sum of squares) of 3MPCA solutions and the associated numbers of components is determined to indicate the numbers of components in a systematic way. The method employs a diagnostic for evaluating the 'salience value' of every solution.

In DIFFIT, we first fit a large number of 3MPCA models, using different combinations of numbers of components P, Q and R. To reduce the possibility of missing the global minimum, it is advisable to use multiple starts for the 3MPCAs. Note that only for combinations with $PQ \geq R$ and $PR \geq Q$ and $QR \geq P$ does the model fit have to be computed, because a model where, for instance, $R > PQ$ gives the same fit as a model with $R = PQ$ (see Wansbeek & Verhees, 1989). Also, 3MPCA models with $P > \max(I, JK)$, $Q > \max(J, IK)$ and/or $R > \max(K, IJ)$ can be omitted, because these models do not fit better than those where P, Q, R equal $\max(I, JK)$, $\max(J, IK)$ and $\max(K, IJ)$, respectively. In practice, values of P, Q and R can be restricted to maximal values chosen *a priori*, and these values will be denoted by P_{max}, Q_{max}, and R_{max}, respectively. Next, we compute all fitted values for solutions with the same total number of components (i.e. $P + Q + R$), and to each total number of components (s) the solution with the highest fit value is assigned. Thus, we have fitted values for $s = 3, \ldots, S$ where S is the sum of the highest numbers of components used in our analyses. For the resulting best fits it can be shown that they must increase monotonically with s, unless one of the best solutions pertains to a local minimum. Next, the difference in fit of s and its predecessor $s - 1$, called dif_s, is computed, $s = 4, \ldots, S$; for $s = 3$, dif_s is equal to the fit of the $s = 3$ (hence $P = Q = R = 1$) model, implying that dif_3 is taken compared to a model with

zero components. These differences in fit play a similar role to that of the eigenvalues obtained in PCA, because the eigenvalues also indicate the increase in fit by addition of one component. Retaining s components is warranted if dif_s is relatively large and if every dif_{s+n} ($n = 1, \ldots, S - s$) is relatively small compared to the other values of s. We propose applying a kind of scree-test to the *dif* values. However, contrary to the eigenvalues in PCA, subsequent values of dif_s do not necessarily decrease. Therefore, rather than comparing subsequent values of dif_s, we compare only dif_s values that are at least as high as all their successors. Specifically, these sequentially maximal values of dif_s are denoted by $dif_{t(m)}$ for $m = 1, \ldots, M$, where M is the number of sequentially maximal values and $t(m)$ refers to the values of s for which dif_s is sequentially maximal, that is, $t(m)$ runs through those values for which $dif_s > dif_{s+n}$, $n = 1, \ldots, S - s$. These successive maxima ($dif_{t(m)}$) are compared with each other using the ratio

$$b_{t(m)} = \frac{dif_{t(m)}}{dif_{t(m+1)}}, \tag{7}$$

which is called the 'salience value' of the solution with $t(m)$ components. A relatively large value of $b_{t(m)}$ indicates that the inclusion of $t(m)$ components (instead of $t(m-1)$ components) increases the fit of the model considerably, whereas the inclusion of any component beyond the $t(m)$th component hardly increases the fit. This notion suggests that those numbers of components should be used which are related to the value of $t(m)$ with the highest salience value, or one of the highest salience values $b_{t(m)}$. However, inclusion of the numbers of components related to $t(m)$ is not sensible if the value of $dif_{t(m)}$ itself is small, because then the addition of the $t(m)$th component no longer increases the fit substantially. This is comparable to excluding components associated with small eigenvalues in PCA. Therefore, we require $dif_{t(m)}$ to exceed a particular threshold value, and discard all solutions for which $dif_{t(m)}$ does not reach the threshold. In PCA, according to Kaiser's criterion, components are only included when they contribute more than the average contribution of each component. Here we choose, analogously, the average of the contributions of every sensible addition of a component as a threshold value. The maximal total number of components is $s_{\max} = \max(I, JK) + \max(J, IK) + \max(K, IJ)$. Hence, in principle, we start with $s = 0$ and add components up to the $s_{\max}$th component. However, not all additions are sensible. For instance, for $s = 1$ and $s = 2$, no sensible models exist; also, for $s = 4$, no sensible models exist, because taking, for instance, $P = Q = 1$ and $R = 2$ leads to the same fit as $P = Q = R = 1$, since $R > PQ$. Therefore, for $s = 0, \ldots, s_{max}$, we have $s_{max} - 3$ sensible additions of components, which on average account for a proportion of $T = (||\mathbf{X}||^2/(s_{max} - 3))$ of the variance. Therefore, we propose to use that solution for which $b_{t(m)}$ is maximal given $dif_{t(m)} > T$.

An interesting special case of our procedure emerges when analysing a two-way array as if it is a three-way array with a single entry in the third mode. In that case $K = 1$, and $dif_{t(m)}$ can be shown to correspond to the $t(m)$th eigenvalue of $\mathbf{X}'\mathbf{X}$. Thus, our criterion is based on the same information as the scree-test for two-way data. The criterion resembles the QDA criterion, but takes ratios of $dif_{t(m)}$ values rather than of differences of such values to avoid problems due to the (in general) rather irregular behaviour of the *dif* values compared to eigenvalues.

DIFFIT can be summarized in six steps:

1. Determine the fit of all 3MPCA models with (P, Q, R) components for which $PQ \geq R$, $PR \geq Q$, $QR \geq P$, up to $P = P_{\max}$, $Q = Q_{\max}$, $R = R_{\max}$.

2. For each value of s determine the best fit among models for which $P + Q + R = s$ ($s = 3, 5, 6, \ldots, P_{\max} + Q_{\max} + R_{\max}$).
3. Determine $dif_{t(m)}$ for $m = 1, \ldots, M$, where $t(m)$ indicates the elements of the subset of values of s for which $dif_s > dif_{s+n}$ for $n = 1, \ldots, S - s$.
4. Compute $b_{t(m)} = dif_{t(m)}/dif_{t(m+1)}$.
5. Determine, among the values of $t(m)$ for which $dif_{t(m)} > ||\mathbf{X}||^2/(s_{\max} - 3)$, the value for which $b_{t(m)}$ is maximal, and denote this as s_c.
6. Choose the numbers of components associated with the best fit among all models for which $P + Q + R = s_c$.

It should be noted that DIFFIT may be too rigid for some data sets. On the one hand, more than one interesting solution may exist. On the other hand, the most interesting solution may be overlooked: a different solution could have a slightly lower salience value, or a solution with a relatively high salience could have a $dif_{t(m)}$ value that is slightly lower than T (the threshold value). These two reasons would cause the most interesting solution to drop out of the procedure. In practice, therefore, it is advisable to inspect the suboptimal solution, and to base one's final choice on issues like interpretability in addition to the results of DIFFIT.

2.2.2 The use of DIFFIT illustrated by an empirical example. In this section, an empirical example is presented to illustrate the use of DIFFIT in deciding on the numbers of components to retain in 3MPCA. A study by Jansen and Bus, 1984, investigates the process of learning to read.[1] Seven pupils were tested weekly (except during holidays) on 37 occasions on five different tests, which were intended to measure different aspects of reading ability. One participant was omitted from the analyses, since he accounted for a large part of the missing data. This data set has been analysed by 3MPCA and discussed by Kroonenberg (1983). Before performing DIFFIT on this data set, the data were preprocessed in the same way as discussed by Kroonenberg (1983). That is, the scores on the five variables were rescaled so that they ranged from 0 to 1. Subsequently, the scores were centred across the participant mode and the rest mode jointly. We refer to Kroonenberg (1983) for a discussion of the rationale behind these steps.

The use of DIFFIT on this data set will be illustrated following the six steps, as discussed in Section 2.2.1. Note that the maximum useful numbers of components for **A**, **B** and **C** were 6, 5 and 30, respectively. The 3MPCA models with (P, Q, R) components for which $PQ \geq R$, $PR \geq Q$, $QR \geq P$ were analysed, leading to a total of 271 analyses (step 1). The fit of a selection of these models and the accompanying numbers of components are presented in Table 1.

Then the best fit among models per sum of numbers of components s ($s = 3, 5, 6, \ldots, 41$) is determined (step 2), as well as the difference in fit between the models with sums of numbers of components s and $s - 1$, denoted by dif_s. A selection of these numbers of components, their sum (s), the accompanying fit and the difference in fit with the predecessor, is presented in Table 2.

Next, the dif_s values that are at least as high as all their successors, denoted by $dif_{t(m)}$, are compared by computing the salience value $b_{t(m)}$ (steps 3 and 4). Salience values associated with the various $dif_{t(m)}$ values are presented in the final column of Table 2.

[1] This data set can be downloaded from http://ruls01.fsw.LeidenUniv.nl/~kroonenb/

Table 1. The fit of a number of 3MPCA models of the 'learning to read data' with several combinations of numbers of components

P	Q	R	$s = P + Q + R$	Fit (%)
1	1	1	3	41.91
1	2	2	5	48.07
2	1	2	5	44.37
2	2	1	5	70.12
2	2	2	6	76.91
1	3	3	7	48.93
2	2	3	7	77.66
2	3	2	7	80.77
3	1	3	7	45.14
3	2	2	7	79.63
3	3	1	7	70.78
2	2	4	8	77.70
2	3	3	8	81.63
2	4	2	8	80.89
3	2	3	8	81.59
3	3	2	8	83.82
4	2	2	8	79.89
6	5	30	41	100

The salience value for which the accompanying dif_s is larger than $T = ||\mathbf{X}||^2/(s_{\max} - 3) = 50.99/(41 - 3) = 1.34$ (step 5), is maximal for $s = 5$. Therefore, the numbers of components to retain according to DIFFIT are (2,2,1) (step 6). The (2,2,1) solution appeared to have a nice interpretation. Additionally, it is verified that other solutions with $s = 5$ gave considerably smaller fit values. Furthermore, the salience value of 4.16 is so

Table 2. A selection of the best fits of the 3MPCA models of the 'learning to read data' by the sum of numbers of components, the accompanying $dif_{t(m)}$ value, and, if defined, the accompanying salience value

P	Q	R	$s = P + Q + R$	fit (%)	dif_s	$b_{t(m)}$
1	1	1	3	41.91	41.91	1.49
2	2	1	5	70.12	28.21	4.16
2	2	2	6	76.91	6.78	1.76
2	3	2	7	80.77	3.86	1.27
3	3	2	8	83.82	3.05	1.33
3	3	3	9	86.11	2.29	1.93
4	3	3	10	87.20	1.09	–
4	3	4	11	88.39	1.19	1.06
4	4	4	12	89.27	0.88	–
5	4	4	13	90.22	0.95	–
5	4	5	14	91.34	1.12	1.19
6	5	30	41	100	0	∞

much larger than the other salience values associated with $dif_{t(m)}$ values larger than 1.34, that other solutions were not considered.

It is interesting to note that Kroonenberg (1983) fitted a (2,2,2) 3MPCA model, but he interpreted the two components of matrices **A** and **B**, and only the first component of **C** and the first slab of the core array. The latter was motivated by noting that the second component of **C** (and thus the second slab of the core array) only accounted for a small amount of variance.

3. Design of the simulation experiment

A Monte Carlo experiment was performed to compare the three methods for indicating the numbers of components (separate PCAs followed by QDA, separate PCAs followed by MAP, and DIFFIT), and to study the sensitivity to local minima of the TUCKALS3 algorithm. In total 360 three-way data arrays $\underline{\mathbf{X}}$, written as two-way matrices **X** of order $(I \times JK)$, were generated by

$$\mathbf{X} = \mathbf{AG}(\mathbf{C}' \otimes \mathbf{B}') + \varepsilon\mathbf{N}, \tag{8}$$

where **A** is a random matrix the elements of which are sampled from the standard normal distribution, **B** and **C** are random orthonormal matrices, **G** is a core matrix, **N** is a random matrix sampled from the standard normal distribution, and ε is a coefficient for manipulating the error level. **A**, **B**, **C** and **G** together constitute the structural part $\mathbf{AG}(\mathbf{C}' \otimes \mathbf{B}')$ of the data. The component matrix **A** was assumed to represent component scores of individuals (which were assumed to be a random sample from a population with normally distributed components), whose number is usually larger than the number of conditions and variables. **B** and **C** were chosen to be orthonormal to avoid collinearity between the components within **B** and **C** due to random variation. This is no severe limitation because they can always be rotated to orthonormality without loss of generality, provided that the rotation of the component matrices is compensated for in the core matrix.

We chose two types of core matrix. The first type is chosen such that a component in a particular mode is related with few other components for the other modes, so as to mimic relatively simple underlying processes. In itself this is not a severe limitation because it can be shown that an arbitrary core can be rotated such that a considerable number of elements in the core **G** will become zero. Murakami, ten Berge, and Kiers (1998), for example, showed that if the condition $P = QR - 1$ is satisfied in a (P, Q, R) core matrix there is a rotation method which yields at least $QR(QR - 2) - (R - 2)$ zero elements. However, there is no guarantee that a simple core will accompany orthonormal component matrices. To cover a broader class of situations, a second type of core matrix was chosen to consist of random matrices sampled from a uniform distribution with elements ranging from -0.5 to 0.5. The latter core matrices may sometimes lead to structural parts of the data matrices which can be fitted almost as well by smaller numbers of components. Therefore, the numbers of *salient* components in such data are smaller than the actual numbers of components. To avoid such undesirable situations we only considered data for which, for the *structural part* of the data (based on a core size of (P, Q, R)), the best solution for $s = P + Q + R - 1$ components should be at most 98% of the total sum of squares. In other words, when fitting the 3MPCA models to the structural part of the data, the *dif* value for the actual numbers of components $(s = P + Q + R)$ should at least be 2%, implying that the step from $s - 1$ to s components

pertains to adding a 'salient' component. Also, to avoid situations where the structural part is to a very large extent explained by a solution with a core size of (1,1,1), we required that the proportion of variance accounted for by the (1,1,1) solution should not exceed 60% of the total sum of squares of the structural part of the data.

The error level was manipulated by means of the coefficient ε, which influences the variance of the distribution of the error part ($\mathbf{E} = \varepsilon\mathbf{N}$). A larger variance of the error distribution increases the total variance. We normalized $\mathbf{N}$ such that it had the same sum of squares as the structural part, so that $100\varepsilon^2/(1 + \varepsilon^2)$ gives the expected percentage error sum of squares in the data, which we here denote as the 'error level'.

To gain insight into the properties of the examined methods for indicating the numbers of components, and into the occurrence of local minima in different conditions, four variables were varied: the size of the data matrix (I, J, K), namely (50,10,5), (100,10,5) and (25,25,25); the error level, namely 14% ($\varepsilon = 0.4$), 39% ($\varepsilon = 0.8$) and 59% ($\varepsilon = 1.2$); the numbers of underlying true components (P, Q, R) denoted by core size, namely (2,2,2), (3,2,2), (3,3,3) and (4,3,2); and the type of core matrix, namely simple (and fixed) or random. The simple cores for the four different numbers of underlying components were chosen as follows:

$$\mathbf{G}_{(2,2,2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{G}_{(3,2,2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{G}_{(3,3,3)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{G}_{(4,3,2)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The design was crossed completely. In each cell 5 replications were taken. Hence, $3 \times 4 \times 3 \times 2 \times 5 = 360$ data arrays were generated. Each data set was analysed by carrying out five runs of the TUCKALS3 algorithm, four of which were initiated randomly and one rationally. As a rational start for $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, the solution obtained from Tucker's method was used. The best of the resulting five solutions was considered the globally optimal solution. All data sets were analysed with all combinations of (P, Q, R) ranging from (1,1,1) to (5,5,5); leaving out all unnecessary analyses, we end up with 74 analyses of each data set. Combinations of (P, Q, R) larger than (5,5,5) were not used because, given the numbers of underlying components, these combinations would almost certainly be irrelevant, and including all of them would result in a considerable increase in computation time. The convergence criterion was set at 10^{-6} times the current function value.

In the comparison of the different methods of indicating the numbers of components to retain, we counted how often each method in each condition recovered the underlying numbers of components. To see if differences can be distinguished from random fluctuations, we analysed these frequencies (between 0 and 5) by means of a repeated multivariate measures analysis of variance (RMANOVA), using only main effects and first-order interactions of method and the independent variables. We chose to use a conservative test by setting the significance level at 0.001.

In the investigation of TUCKALS3's sensitivity to local minima, we counted how often the rational start of the TUCKALS3 algorithm led to a local minimum in each condition; a local minimum was defined as having a loss function value higher than 1.0001 times the function value of the solution (out of five) with the lowest function value. Thus, the lowest function value was considered to be the global minimum. These frequencies were analysed by means of an RMANOVA (again using only main effects and first-order interactions of method and the independent variables) to test the effects of the different independent variables. Here also the significance level was chosen to be 0.001.

4. Results

4.1. Recovery of the numbers of components in 3MPCA

For each data set, the numbers of components (P, Q, R) indicated by MAP, QDA and DIFFIT were compared with the true numbers of components. The percentages of correctly selected triples (P, Q, R) are plotted for each method for each data size, for each error level, for each core size and for each core type in Fig. 1.

As can be seen from Fig. 1, large differences exist in performance between the methods and between the conditions, which, moreover, in the RMANOVA all turned out to be significant ($p < 0.001$). Generally, DIFFIT performed better than QDA, which performed

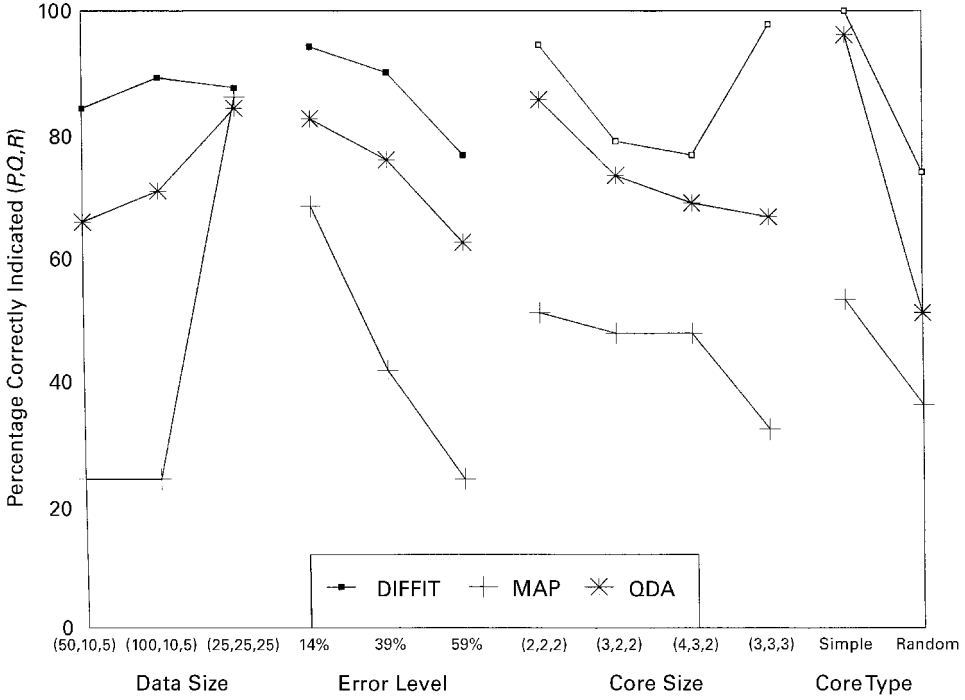

Figure 1. Percentage of correctly selected numbers of components for each method (DIFFIT, QDA and MAP) for each data size, error level, core size and core type.

better than MAP. Increasing data size (defined as IJK, the total number of data elements) led to more correctly selected numbers of components. The performance of the three methods deteriorated as the error level increased. Increasing core size (PQR, the number of elements in the core) led to fewer correctly chosen triples (P, Q, R) (except for DIFFIT, which performs better with core size (3,3,3) than with any other core size). The numbers of components of the core were chosen correctly more often if the data were constructed using a simple core than if using a random core. Interactions between method and conditions all turned out to be significant ($p < 0.001$), but their pattern differed, as will now be discussed. As data size (IJK) increased, the performance of the three methods became more similar: DIFFIT, MAP and QDA performed almost equally in the condition with the largest data set (25,25,25). The interaction between method and error level mainly pertained to the relatively more quickly deteriorating performance of the MAP method as the error level increased. The interaction between method and core size mainly pertained to the difference in performance of the three methods for core size (4,3,2) versus (3,3,3): MAP performed worse, QDA performed equally well and DIFFIT performed better for core size (3,3,3) than for core size (4,3,2). The interaction between method and core type mainly pertained to the relatively poor performance of QDA for data constructed using a random core. It is noteworthy that DIFFIT is the only method that performed reasonably well for data based on random cores.

It has been seen that the correct numbers of components were frequently not recovered. Then it is of interest to see if methods have a tendency to over- or underestimate the dimensionalities. The three methods generally underestimated the numbers of components (more than 99% of the cases, where each estimated size of one of the three modes by one of the three methods is considered a case).

4.2. The number of starts of TUCKALS3 leading to local minima

In our Monte Carlo experiment we analysed the 360 constructed data sets by 3MPCA with every combination of values P, Q, R which it was necessary to consider, ranging from (1,1,1) to (5,5,5), resulting in $360 \times 74 = 26\,640$ 3MPCA models. Every 3MPCA model was estimated by the TUCKALS3 algorithm, using one rational and four random starts. For each TUCKALS3 analysis we determined whether and how often the random start led to a local minimum. We used the same independent variables as before, except that, rather than inspecting the effect of core size itself, we investigated the effect of the difference between the estimated numbers of components and the correct numbers of components. We expressed the 'distance' between the correct numbers of components and the estimated numbers of components by the sum distance measure (SDM), which was defined as the sum of the absolute difference between the correct number and the estimated number of components for each of the three modes. For example, for a correct 'triple' (3,3,3) and an estimated 'triple' (2,5,5), the SDM is $1 + 2 + 2 = 5$. The percentages of local minima per start per TUCKALS3 analysis were analysed by means of an RMANOVA to test the effects of data size, error level, SDM and core type.

The percentages of rationally and randomly initiated TUCKALS3 analyses leading to a local minimum, by data size, error level, SDM and core type, are plotted in Fig. 2. Overall, a randomly initiated TUCKALS3 analysis led more often to a locally optimal solution than a rationally initiated analysis (16% and 13%, respectively). The percentage of rational and random starts leading to a local minimum increased with increasing data size and, albeit

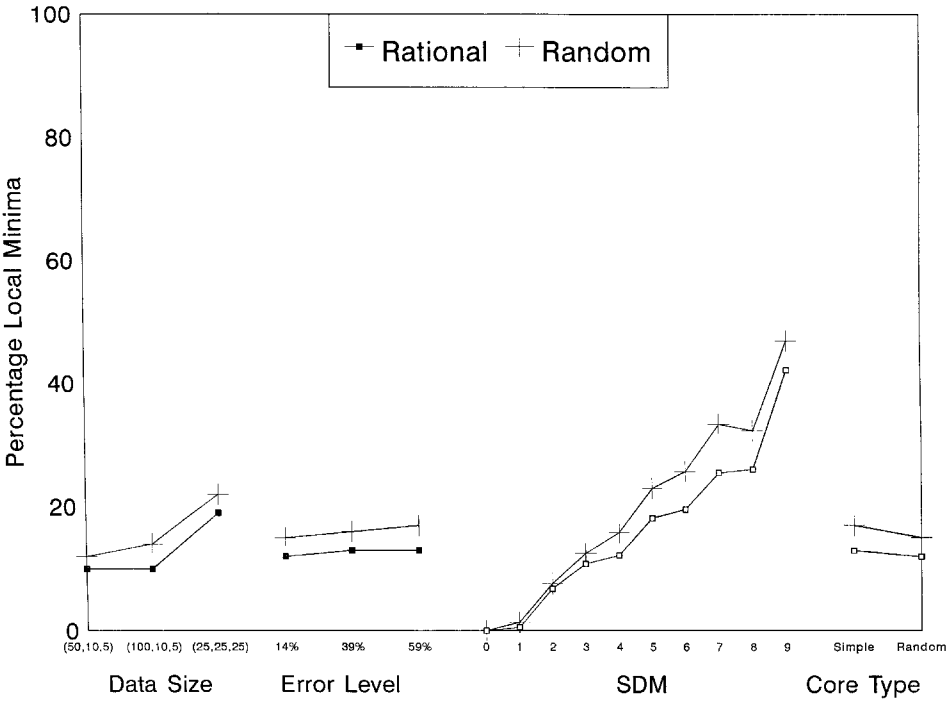

Figure 2. Percentage of rationally initiated runs (Rational) and of randomly initiated runs (Random) leading to a local minimum for each data size, error level, SDM and core type.

negligibly, with increasing error level. The sensitivity to local optima of both rationally and randomly initiated runs increased considerably with increasing SDM. It is particularly striking that if the correct number of components was used, *none* of the TUCKALS3 analyses led to a local minimum. Rationally initiated TUCKALS3 analyses of data constructed from a random core led less often to local minima than data constructed from a simple core. The reported effects are all significant, which is not surprising in view of the huge number of cases, but the effect sizes are rather small, as was indicated by the partial η^2 values. These values were 0.02, 0.03, 0.00, 0.09 and 0.00 for type of start, data size, error level, SDM and core type, respectively. We only deem the effect sizes of type of start, data size and SDM to be relevant.

5. Discussion

5.1. Choosing the numbers of components in 3MPCA

The first issue studied here concerned the choice of the numbers of components for 3MPCA of a given three-way array. Three methods, two based on two-way component analyses (MAP and QDA) and one specifically designed for 3MPCA (DIFFIT), were compared. The results of our simulation study indicate that DIFFIT is much better than MAP and QDA in selecting the correct numbers of components. DIFFIT performed better in all conditions of this experiment, and it does not seem to require large numbers of observations as much as the

other methods. In particular, MAP and QDA performed worse for data sizes (50,10,5) and (100,10,5), which implies that the ratio between number of observed entities and number of components is relatively low. Such an effect of the ratio of number of variables to number of factors has been found in other contexts as well. Marsh, Hau, Balla, and Grayson (1998) found that the higher this ratio the better the correct identification and estimation of factors in a factor analysis. Velicer and Fava (1987) reported similar results concerning factor analysis as well as component analysis. The present results suggest that the ratio between observed entities and numbers of components influences not only the success in recovering the structure of the data, but also the success of ordinarily used two-way methods in determining the numbers of components. On the other hand, DIFFIT does not seem to be affected very much by a poor ratio of number of entities to number of components.

The use of DIFFIT requires computation of 3MPCA solutions for all (useful) combinations of numbers of components. In this experiment, only solutions for numbers of components ranging from (1,1,1) to (5,5,5) were computed to decrease the computing time, excluding some solutions for $s \geq 10$. This does not constitute a serious limitation of the experiment. First of all, it is similar to what happens in practice, where bounds are put on the numbers of components; for the simulated data the bounds were fairly wide. Secondly, the cases where $s \geq 10$ usually, as was seen for the cases with $s \geq 10$ that were computed, correspond to solutions for which the *dif* value is low and often does not exceed the threshold value T.

In our simulation experiment, the error part of the data matrices was sampled from the standard normal distribution. The results of a pilot study indicated that sampling the error part from a uniform distribution hardly altered the achievements of the three methods.

The methods for indicating the numbers of components considered here are based on the idea that a component should only be retained if it explains a substantial amount of variance. The structural part of the more complex data sets (constructed with random core) was chosen such that the (1,1,1) model explained less than 60% and the last component explained more than 2% of the total sum of squares. We investigated whether the performance of the methods was actually related to the fit of the structural part of the (1,1,1) model and to the contribution of the last component. The performance of the methods appeared to be hardly related to the fit of the (1,1,1) model to the structural part of the data, but is clearly related to the contribution of the last component in the analysis of the structural part. For example, the numbers of components for the 29 data sets for which this contribution was between 2% and 4% are indicated correctly by DIFFIT in 41% of cases, which is considerably worse than the performance of DIFFIT for all data sets (87% correct). In fact, this accounts for 17 of the 47 cases where DIFFIT indicated the wrong numbers of components. A similar result was found for MAP and QDA. Thus, if we were also to leave out the above, still rather extreme, situations the comparison would lead to a similar conclusion, but the performance was better, and in fact, for DIFFIT would lead to 92% correctly indicated numbers of components.

As DIFFIT performed considerably better than methods based on two of the best two-way techniques, the use of DIFFIT for indicating the numbers of components appears to be the best choice. The fact that DIFFIT requires considerably more computation time than QDA and MAP does not make the method unfeasible: one full 3MPCA (i.e., with one rational start and four random starts) took approximately 10 seconds in our study (using a Pentium 133 MHz PC). The computation time increased with increasing data size: the mean computation times were 4, 8 and 19 seconds for data sizes (50, 10, 5), (100, 10, 5) and (25, 25, 25), respectively. Computation times may become prohibitive for very

large data sets. For such data sets, however, one may resort to compression-based algorithms (Bro & Andersson, 1998) that are very efficient and fit almost as well as the original algorithm.

The methods for indicating the numbers of components discussed here aim to find components explaining a substantial amount of variance. This kind of method deliberately ignores weak components, even though they can pertain to replicable information in the data. If one does want to trace such weak components, one may resort to cross-validation techniques to establish the replicability of components (Louwerse, Smilde, & Kiers, 1999). We would like to emphasize that the final decision about the numbers of components to use has to take into account substantive information and interpretability of the results. DIFFIT can be used as a first step in choosing the numbers of components. In particular, the DIFFIT approach can be used to find those solutions for which the salience value is maximal or close to maximal. A choice between these solutions can then be made on other grounds.

5.2. The number of starts of TUCKALS3 leading to local minima

The second topic studied here dealt with the question of whether, how often and in which situations the rational and random starts lead to a local minimum in estimating the 3MPCA model by the TUCKALS3 algorithm. In interpreting the results of the experiment, it is assumed that at least one of the five analyses per data set reached the global minimum, although the possibility that the solution with the lowest function value in fact ended in a local minimum cannot entirely be excluded. However, the more of the five analyses that ended in the same 'global' minimum, the higher the chance that this minimum *is* the global minimum. The results of the simulation experiment indicate that an important factor in the occurrence of local minima is the distance between the correct triple of numbers of components and the numbers of components as employed in a 3MPCA. If the latter distance was zero, all five analyses reached the same minimum in 99.9% of the cases. Hence, if the numbers of components are correctly chosen, the probability of the rationally initiated TUCKALS3 algorithm hitting a local minimum appears to be very small. Also, in general, rationally initiated TUCKALS3 analyses lead less often to a local minimum than randomly initiated runs, and therefore we recommend using the rational start in any case. Of course, the probability of missing the global minimum decreases if multiple starts are used. However, the results of our simulation experiment indicate that, if the numbers of components used do not differ much from the 'ideal' numbers of components, using only one (rational) start is not very likely to lead to serous problems.

Acknowledgements

This research has been made possible by funding from the Netherlands organization for scientific research (NWO) to the first author. The authors are obliged to Jos ten Berge, two anonymous reviewers and the editor for useful comments on an earlier version of this paper. We thank A. G. Bus for putting her data set at our disposal.

References

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, *3*, 77–85.

Bartlett, M. S. (1951). A further note on tests of significance in factor analysis. *British Journal of Psychology*, *4*, 1–2.

Bro, R., & Andersson, C. A. (1998). Improving the speed of multi-way algorithms. Part II: Compression. *Chemometrics and Intelligent Laboratory Systems, 42,* 105–113.

Cattell, R. B. (1966). The meaning and strategic use of factor analysis. In R. B. Cattell (Ed.). *Handbook of multivariate experimental psychology* (pp. 174–243). Chicago: Rand McNally.

Dai, K. (1985). Application of a three-mode factor analysis to brand images of whisky. *Reports of Statistical Application Research, 32(1),* 11–22.

Harshman, R. A., & Lundy, M. E. (1984). Data preprocessing and the extended PARAFAC model. In H. G. Law, C. W. Snyder Jr, J. A Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 216–284). New York: Praeger.

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

Jansen, M. G. H., & Bus, A. G. (1984). Individual growth patterns in early reading performance. *Kwantitatieve Methoden, 14,* 97–109.

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

Kroonenberg, P. M. (1983). *Three mode principal component analysis. Theory and applications.* Leider: DSWO Press.

Kroonenberg, P. M., & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika, 45,* 69–97.

Kroonenberg, P. M., Lammers, C. J., & Stoop, I. (1985). Three-mode principal component analysis of multivariate longitudinal organizational data. *Sociological Methods and Research, 14(2),* 99–136.

Kroonenberg, P. M., & van der Voort, T. H. A. (1987). Multiplicatieve decompositie van interacties bij oordelen over de werkelijkheidswaarde van televisiefilms [Multiplicative decomposition of interactions for judgments of realism of television films]. *Kwantitatieve Methoden, 23,* 117–144.

Louwerse, D. J., Smilde, A. K., & Kiers, H. A. L. (1999). Cross-validation of multiway component models. *Journal of Chemometrics, 13,* 491–510.

Love, W. D., & Tucker, L. R. (1970). *A three-mode factor analysis of serial learning.* Office of Naval Research Report.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Murakami, T., ten Berge, J. M. F., & Kiers, H. A. L. (1998). A case of extreme simplicity of the core matrix in three-mode principal component analysis. *Psychometrika, 63,* 255–261.

Niesing, J. (1997). *Simultaneous component and factor analysis methods for two or more groups: a comparative study.* Leiden: DSWO Press.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31,* 279–311.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321–327.

Velicer, W. F. & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image and factor analysis. *Multivariate Behavioral Research, 22,* 193–209.

Wansbeek, T., & Verhees, J. (1989). Models for multidimensional matrices in econometrics and psychometrics. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 543–552). Amsterdam: North Holland.

Weesie, J., & van Houwelingen, H. (1983). *GEPCAM users' manual (first draft).* Utrecht: Institute of Mathematical Statistics, State University of Utrecht.

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