

How to perform multiblock component analysis in practice

Kim De Roover · Eva Ceulemans ·
Marieke E. Timmerman

Published online: 12 July 2011
© Psychonomic Society, Inc. 2011

Abstract To explore structural differences and similarities in multivariate multiblock data (e.g., a number of variables have been measured for different groups of subjects, where the data for each group constitute a different data block), researchers have a variety of multiblock component analysis and factor analysis strategies at their disposal. In this article, we focus on three types of multiblock component methods—namely, principal component analysis on each data block separately, simultaneous component analysis, and the recently proposed clusterwise simultaneous component analysis, which is a generic and flexible approach that has no counterpart in the factor analysis tradition. We describe the steps to take when applying those methods in practice. Whereas plenty of software is available for fitting factor analysis solutions, up to now no easy-to-use software has existed for fitting these multiblock component analysis methods. Therefore, this article presents the MultiBlock Component Analysis program, which also includes procedures for missing data imputation and model selection.

Keywords Multigroup data · Multilevel data · Simultaneous component analysis · Principal component analysis · Clusterwise simultaneous component analysis

K. De Roover (✉) · E. Ceulemans
Department of Educational Sciences,
Katholieke Universiteit Leuven,
Andreas Vesaliusstraat 2,
B-3000 Leuven, Belgium
e-mail: Kim.DeRoover@ped.kuleuven.be

M. E. Timmerman
University of Groningen,
Groningen, Netherlands

Introduction

In the behavioral sciences, researchers often gather multivariate multiblock data—that is, multiple data blocks, each of which contains the scores of a different set of observations on the same set of variables. For an example, one can think of multivariate data from different groups of subjects (e.g., inhabitants from different countries). In that case, the groups (e.g., countries) constitute the separate data blocks. Another example is data from multiple subjects that have scored the same variables on multiple measurement occasions (also called multioccasion–multisubject data; see Kroonenberg, 2008). In such data, the data blocks correspond to the different subjects.

Both the observations in the data blocks and the data blocks themselves can be either fixed or random. For instance, in the case of multioccasion–multisubject data, the data blocks are considered fixed when the researcher is interested in the specific subjects in the study and random when one aims at generalizing the conclusions to a larger population of subjects. In the latter case, a representative sample of subjects is needed to justify the generalization. When the observations are random and the data blocks fixed, multiblock data are referred to as *multigroup* data (Jöreskog, 1971) in the literature. When both observations and data blocks are random, the data are called *multilevel* (e.g., Maas & Hox, 2005; Muthén, 1994; Snijders & Bosker, 1999).

Researchers are often interested in the underlying structure of such multiblock data. For instance, given multioccasion–multisubject scores on a number of emotions, one can wonder whether certain emotions covary across the measurement occasions of separate subjects or fluctuate independently of one another, and whether and

how this structure is similar across subjects (i.e., across the data blocks). For capturing the structure in multiblock data, a number of component analysis, as well as factor analysis, methods are available. Component analysis and factor analysis differ strongly with respect to their theoretical underpinnings, but both of them model the variation on the variables by a smaller number of constructed variables—called *components* and *factors*, respectively—which are based on the covariance structure of the observed variables. Which component or factor analysis method is the most appropriate depends on the research question at hand. For well-defined, confirmatory questions, factor analysis is usually most appropriate. For exploratory analysis of data that may have an intricate structure, as is often the case for multiblock data, component analysis is generally most appropriate.

To test specific hypotheses about the underlying structure, structural equation modeling (SEM; Haavelmo, 1943; Kline, 2004) is commonly used. SEM is applied, for example, to test whether the items of a questionnaire measure the theoretical constructs under study (Floyd & Widaman, 1995; Keller et al., 1998; Novy et al., 1994). Moreover, multigroup SEM (Jöreskog, 1971; Kline, 2004; Sörbom, 1974) allows testing different levels of factorial invariance among the data blocks (e.g., Lee & Lam, 1988), going from weak invariance (i.e., same factor loadings for all data blocks) to strict invariance (i.e., intercepts, factor loadings, and unique variances equal across data blocks).

When there are no a priori hypotheses about the underlying structure, one may resort to component analysis or exploratory factor analysis (EFA). We will first discuss a family of component methods that explicitly focus on capturing structural differences between the data blocks and then, briefly, the family of factor analysis methods. Note that many other multiblock component methods exist that focus, for example, on redundancy (Escofier & Pagès, 1998) or on modeling block structured covariance matrices (Flury & Neuenschwander, 1995; Klingenberg, Neuenschwander, & Flury, 1996).

If one expects the structure of each of the data blocks to be different, standard principal component analysis (PCA; Jolliffe, 2002; Pearson, 1901) can be performed on each data block. In case one thinks that the structure will not differ across the data blocks, simultaneous component analysis (SCA; Kiers, 1990; Kiers & ten Berge, 1994b; Timmerman & Kiers, 2003; Van Deun, Smilde, van der Werf, Kiers, & Van Mechelen, 2009) can be applied, which reduces the data for all the blocks at once to find one common component structure for all the blocks. Finally, if one presumes that subgroups of the data blocks exist that share the same structure, one may conduct clusterwise simultaneous component analysis (clusterwise SCA-ECP,

where ECP stands for equal cross-product constraints on the component scores of the data blocks; De Roover, Ceulemans, Timmerman, Vansteelandt, et al., *in press*; Timmerman & Kiers, 2003). This method simultaneously searches for the best clustering of the data blocks and for the best fitting SCA-ECP model within each cluster. This flexible and generic approach encompasses separate PCA and SCA-ECP as special cases.

For the separate PCA and SCA-ECP approaches, similar factor-analytic approaches exist, which are specific instances of exploratory structural equation modeling (Asparouhov & Muthén, 2009; Dolan, Oort, Stoel, & Wicherts, 2009; Lawley & Maxwell, 1962). While component and factor analyses differ strongly with respect to their theoretical backgrounds, they often give comparable solutions in practice (Velicer & Jackson, 1990a, b). However, no factor-analytic counterpart exists for the clusterwise SCA-ECP method.

While plenty of software is available for the factor-analytic approaches (e.g., LISREL, Dolan, Bechger, & Molenaar, 1999; Jöreskog & Sörbom, 1999; Mplus, Muthén, & Muthén, 2007; and Mx, Neale, Boker, Xie, & Maes, 2003), no easy-to-use software program exists for applying the multiblock component analysis methods described above. Thus, although the component methods are potentially very useful for substantive researchers (e.g., De Leersnyder & Mesquita, 2010; McCrae & Costa, 1997; Pastorelli, Barbaranelli, Cermak, Rozsa, & Caprara, 1997), it might be difficult for researchers to apply them. In this article, we describe software for fitting separate PCAs, SCA-ECP, and clusterwise SCA-ECP models. This Multi-Block Component Analysis (MBCA) software (Fig. 1) can be downloaded from <http://ppw.kuleuven.be/okp/software/MBCA/>. The program is based on MATLAB code, but it can also be used by researchers who do not have MATLAB at their disposal. Specifically, two versions of the software can be downloaded: one for use within the MATLAB environment and a *stand-alone* application that can be run on any Windows computer. The program includes a model selection procedure and can handle missing data.

The remainder of the article is organized in three sections. In Section **Multiblock component analysis**, we first discuss multiblock data, how to preprocess them and how to deal with missing data. Subsequently, we discuss clusterwise SCA-ECP as a generic modeling approach that comprises separate PCAs and SCA-ECP as special cases. Finally, we describe the different data analysis steps: checking data requirements, running the analysis, and model selection. The clusterwise SCA-ECP approach is illustrated by means of an empirical example. Section **Multiblock component analysis program** describes the handling of the MBCA software. Section **Conclusion** adds a general conclusion to the article.

Multiblock component analysis

Data structure, preprocessing, and missing values

In this section, we first describe the data structure that is required by the multiblock component methods under study. Second, the preprocessing of the data is discussed. Third, the problem of missing values is reviewed shortly, since an important feature of the MBCA software is that it can handle missing data.

Data structure

Clusterwise SCA-ECP, as well as SCA-ECP and separate PCAs, is applicable to all kinds of multivariate multiblock data—that is, data that consist of I data blocks \mathbf{X}_i ($N_i \times J$) that contain scores of N_i observations on J variables, where the number of observations N_i ($i = 1, \dots, I$) may differ between data blocks. These I data blocks can be concatenated into an N (observations) \times J (variables) data matrix \mathbf{X} , where $N = \sum_{i=1}^I N_i$. More specific requirements (e.g., minimal number of observations in each data block) will be discussed in the [Checking data requirements](#) section.

As an example, consider the following empirical data set from emotion research, which will be used throughout the article. Emotional granularity refers to the degree to which a subject differentiates between negative and positive emotions (Barrett, 1998); that is, subjects who score high on emotional granularity describe their emotions in a more fine-grained way than subjects scoring low. To study emotional granularity, 42 subjects were asked to rate on a 7-point scale the extent to which 22 target persons (e.g., mother, father, partner, . . .) elicited 16 negative emotions, where the selected target persons obviously differ across subjects. Thus, one may conceive these data as consisting of 42 data blocks \mathbf{X}_i , 1 for each subject, where each data block holds the ratings of the 16 negative emotions for the 22 target persons selected by subject i . Note that, in this case, the number of observations N_i is the same for all data blocks, but this is not necessary for the application of any of the three component methods considered. The data blocks $\mathbf{X}_1, \dots, \mathbf{X}_{42}$ can be concatenated below each other, resulting in a 924×16 data matrix \mathbf{X} .

Preprocessing

Before applying any of the multiblock component methods, one may consider whether or not the data should be preprocessed. Since we focus on differences and similarities in within-block correlational structures, we disregard between-block differences in variable means and in variances. Note that variants of the PCA and SCA methods exist in which the differences in means (Timmerman, 2006)

and variances (De Roover, Ceulemans, Timmerman, & Onghena, 2011; Timmerman & Kiers, 2003) are explicitly modeled. To eliminate the differences in variable means and variances, the data are centered and standardized per data block. This type of preprocessing, which is implemented in the MBCA software, is commonly denoted as *autoscaling* (Bro & Smilde, 2003). The standardization also results in a removal of arbitrary differences between the variables in measurement scale.

Missing values

In practice, data points may be missing. For instance, in our emotion data, 4% of the data points are missing, because some subjects neglected to rate certain emotions for some of their target persons. To judge the generalizability of the results obtained, one has to consider the method for dealing with the missing data in the analysis and the mechanism(s) that plausibly caused the missing data. To start with the latter, Rubin distinguished between “missing completely at random,” (MCAR) “missing at random,” (MAR) and “not missing at random” (NMAR) (Little & Rubin, 2002; Rubin, 1976). MCAR means that the missing data are related neither to observed nor to unobserved data. When data are MAR, the missing data are dependent on variables in the data set but are unrelated to unobserved variables. NMAR refers to missing data that depend on the values of unobserved variables.

To deal with missing data in the analysis, we advocate the use of imputation. Imputation is much more favorable than the simplest alternative—namely, to discard all observations that have at least one missing value. The latter may result in large losses of information (Kim & Curry, 1977; Stumpf, 1978) and requires the missing data to be MCAR. In contrast, imputation requires the missing data to be MAR, implying that it is more widely applicable. The procedure to perform missing data imputation in multiblock component analysis is described in the Missing Data Imputation section. and is included in the MBCA software.

The clusterwise SCA-ECP model

A clusterwise SCA-ECP model for a multiblock data matrix \mathbf{X} consists of three ingredients: a $I \times K$ binary partition matrix \mathbf{P} , which represents how the I data blocks are grouped into K mutually exclusive clusters; $K J \times Q$ cluster loading matrices \mathbf{B}^k , which indicate how the J variables are reduced to Q components for all the data blocks that belong to cluster k ; and a $N_i \times Q$ component score matrix \mathbf{F}_i for each data block. Figures 1 and 2 presents the partition matrix \mathbf{P} and the cluster loading matrices \mathbf{B}^k , and Table 1 presents the component score matrix \mathbf{F}_2 (of subject 2) of a clusterwise SCA-ECP model with three clusters and two components for our emotion data. The partition matrix \mathbf{P}

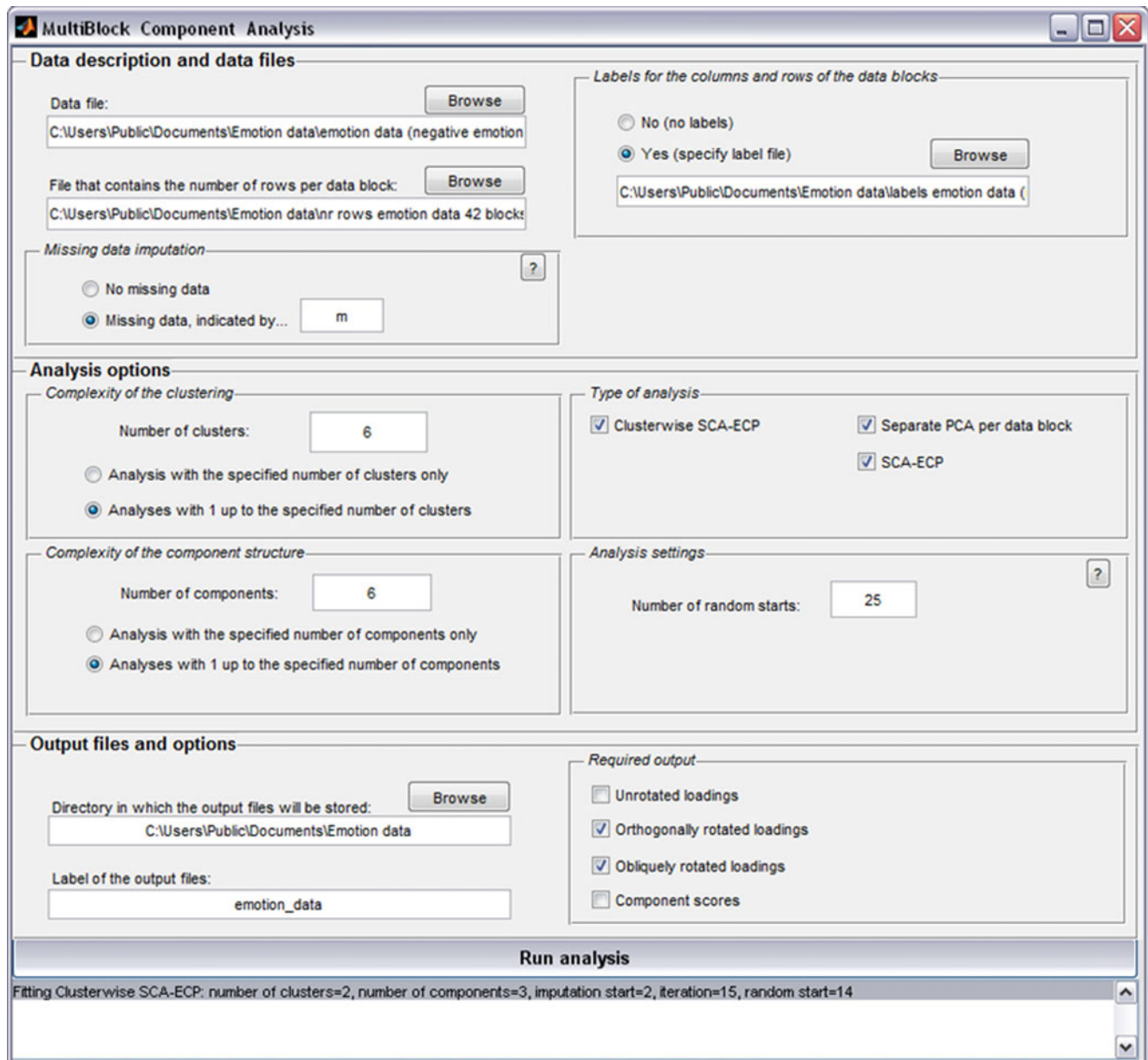


Fig. 1 Interface of the MultiBlock component analysis software

shows that 15 subjects are assigned to the first cluster (i.e., 15 subjects have a one in the first column and a zero in the other columns), while the second and third cluster contain 14 and 13 subjects, respectively.

The cluster loading matrices \mathbf{B}^k in Fig. 2 display the component structure for each of the subject clusters. Because we analyzed autoscaled data and have orthogonal components, the loadings can be interpreted as correlations between the variables and components. Each component of a cluster loading matrix can be interpreted by considering the common content of the variables that load highly positive or negative on that component (e.g., loadings with an absolute value greater than .50). Specifically, for cluster

1, the first component can be labeled *negative affect*, since virtually all negative emotions have a high positive loading on this component. The second component of this cluster is named *jealousy* due to the high loading of *jealous*. For cluster 2, the first component is termed *cold dislike*, since it consists of the negative emotions that are experienced when feeling a dislike for someone, without feeling sad about this. The second component is made up out of negative feelings that arise when feeling sad; therefore, it is named *sadness*. The first component of cluster 3 has some similarities to the first component of cluster 2, but with the important difference that additional emotions load high on this component—that is, *uneasy*, *miserable*, and *sad*.

Fig. 2 Output file for the clusterwise SCA-ECP analysis of the emotion data, showing the partition matrix and the orthogonally rotated cluster loading matrices for the model with two clusters and two components. The components in cluster 1 can be labeled “negative affect” and “jealousy,” while the components in cluster 2 can be interpreted as “cold dislike” and “sadness,” and the ones in cluster 2 as “hot dislike” and “low self-esteem,” respectively

emotion_data_Clusterwise_SCA-ECP_orthogonal.txt - Notepad

File Edit Format View Help

Analysis with 3 clusters and 2 components

Partition matrix

subject	1	2	3
subject 1	1	0	0
subject 2	1	0	0
subject 3	0	0	1
subject 4	0	0	1
subject 5	0	1	0
subject 6	1	0	0
subject 7	0	1	0
subject 8	1	0	0
subject 9	0	0	1
subject 10	0	1	0
subject 11	0	0	1
subject 12	1	0	0
subject 13	0	0	1
subject 14	1	0	0
subject 15	0	1	0
subject 16	1	0	0
subject 17	0	0	1
subject 18	1	0	0
subject 19	0	1	0
subject 20	1	0	0
subject 21	0	1	0
subject 22	0	1	0
subject 23	0	1	0
subject 24	0	1	0
subject 25	0	0	1
subject 26	0	0	1
subject 27	0	1	0
subject 28	0	1	0
subject 29	0	1	0
subject 30	0	0	1
subject 31	1	0	0
subject 32	0	0	1
subject 33	0	1	0
subject 34	1	0	0
subject 35	1	0	0
subject 36	0	0	1
subject 37	1	0	0
subject 38	1	0	0
subject 39	1	0	0
subject 40	0	1	0
subject 41	0	0	1
subject 42	0	0	1

Component loadings

	Cluster 1		Cluster 2		Cluster 3	
bored	0.7788	-0.1270	0.7505	-0.1092	0.6886	-0.1069
uneasy	0.8334	0.0025	0.4987	0.1921	0.5970	0.2025
miserable	0.6354	0.2841	0.1566	0.6217	0.5765	0.3130
angry	0.7772	0.3170	0.7745	0.2073	0.7516	0.0919
confused	0.5441	0.3317	0.1633	0.7520	0.5493	0.2728
dislike	0.7802	0.0851	0.8821	-0.0137	0.7640	0.2231
inferior	0.3299	0.5100	-0.0755	0.6153	0.1898	0.6614
sad	0.6330	0.4938	0.0963	0.7910	0.6679	0.4885
frustrated	0.7360	0.3959	0.7551	0.2454	0.7211	0.3973
jealous	-0.2317	0.7566	-0.2447	0.4180	-0.0077	0.6131
fearful	0.6051	0.2813	0.2189	0.6445	0.1879	0.6178
nervous	0.7553	0.1418	0.1702	0.6703	0.1678	0.7657
uncomfortable	0.8559	0.0422	0.6321	0.2958	0.3836	0.6038
disgust	0.8369	0.0928	0.9001	-0.0819	0.7513	0.2555
upset	0.8570	0.2163	0.5506	0.3861	0.6541	0.3124
hatred	0.8182	0.2093	0.8988	-0.0508	0.7942	0.0863

Thus, it seems that for this cluster, dislike is more emotionally charged; thus, it is labeled *hot dislike*. The second component of cluster 3 can be interpreted as a *low self-esteem* component, because it consists of negative emotions that stem from having a low feeling of self-worth. We can conclude that the clusters differ strongly from one another in the nature of the dimensions that underlie the negative emotions. On the basis of these results, we can hypothesize that the subjects in cluster 1 are the ones with the least granular emotional experience against the target persons, since most negative emotions strongly co-occur for these subjects. The subjects in clusters 2 and 3 seem to display a higher granularity in their emotional ratings, since they differentiate between feelings of dislike and feelings of sadness (cluster 2) or low self-esteem (cluster 3) toward the target persons. To evaluate whether the structural differences between the three subject clusters may, indeed, be interpreted as differences in

emotional granularity, we related the cluster membership to the average intraclass correlation coefficients (ICCs; Shrout & Fleiss, 1979; Tugade, Fredrickson, & Barrett, 2004) measuring absolute agreement, which were calculated across the negative emotions for each subject. The three clusters differ significantly, $F(2) = 5.12$, $p = .01$, with mean ICCs of .91 ($SD = .07$), .82 ($SD = .10$), and .88 ($SD = .05$) for clusters 1–3, respectively. Since higher ICC values indicate a lower granularity, cluster 1 contains the least granular subjects, while cluster 2 contains the most granular subjects.

In Table 1, the component score matrix for subject 2 is presented. Since subject 2 belongs to cluster 1 (see the partition matrix in Fig. 2), we can derive how this subject feels about each of the 22 target persons in terms of *negative affect* and *jealousy*. For instance, it can be read that target person 15 (disliked person 3) elicits the most negative affect in subject 2 (i.e., a score of 1.93 on the first component) and that this subject has the strongest

Table 1 Component scores for subject 2 out of the emotion data, given a clusterwise SCA-ECP model with two clusters and two components. Note that subject 2 is assigned to the first cluster

Target Person	Negative Affect	Jealousy
1. Mother	-0.28	0.24
2. Father	-0.57	0.11
3. Partner	-0.47	2.15
4. Liked relative 1	-0.72	0.72
5. Liked relative 2	-0.61	-1.42
6. Disliked relative 1	0.06	-1.02
7. Disliked relative 2	1.86	-0.06
8. Disliked relative 3	1.79	-0.66
9. Disliked relative 4	1.03	-0.88
10. Best friend 1	-0.72	-0.02
11. Best friend 2	-0.75	-0.92
12. Best friend 3	-0.37	0.23
13. Disliked person 1	1.40	-0.31
14. Disliked person 2	1.43	2.73
15. Disliked person 3	1.93	-0.81
16. Acquaintance 1	-0.95	-0.75
17. Acquaintance 2	-0.62	0.62
18. Acquaintance 3	-0.47	0.33
19. Mentor 1	-1.05	0.48
20. Mentor 2	-0.48	-0.17
21. Mentor 3	-0.81	-1.33
22. Myself	-0.54	0.59

feelings of jealousy (i.e., a score of 2.73 on the second component) toward target person 14 (disliked person 2).

To reconstruct the observed scores in each data block \mathbf{X}_i , the information in the three types of matrices is combined as follows:

$$\mathbf{X}_i = \sum_{k=1}^K p_{ik} \mathbf{F}_i^k \mathbf{B}^{k'} + \mathbf{E}_i, \quad (1)$$

where p_{ik} denotes the entries of the partition matrix \mathbf{P} , \mathbf{F}_i^k is the component score matrix for data block i when assigned to cluster k , and \mathbf{E}_i ($N_i \times J$) denotes the matrix of residuals. Since data block i is assigned to one cluster only, the index k in \mathbf{F}_i^k is mostly omitted in the remainder of this article, for reasons of parsimony. For example, to reconstruct the data block for subject 2, we read in the partition matrix (Fig. 2) that this subject is assigned to the first cluster and, subsequently, multiply the component scores in Table 1 with the component loadings of cluster 1 in Fig. 2. It can be concluded that the separate PCA and SCA-ECP strategies for multiblock data are indeed special cases of the clusterwise SCA-ECP model. On the one hand, when K , the number of clusterwise SCA-ECP clusters, is equal to I , the number of data blocks, the model boils down to

separate PCAs with an equal number of components for each data block. On the other hand, when K equals one, all data blocks belong to the same cluster, and the clusterwise SCA-ECP model reduces to a regular SCA-ECP model.

In clusterwise SCA-ECP, the columns of each component score matrix \mathbf{F}_i are restricted to have a variance of one, and the correlations between the columns of \mathbf{F}_i (i.e., between the cluster-specific components) must be equal for all data blocks that are assigned to the same cluster. With respect to the latter restriction, note that the parameter estimates of an SCA-ECP solution have rotational freedom. Thus, to obtain components that are easier to interpret, the components of a clusterwise SCA-ECP solution can be freely rotated within each cluster without altering the fit of the solution, provided that the corresponding component scores are counterrotated. For instance, the cluster loading matrices in Fig. 2 were obtained by means of an orthogonal normalized varimax rotation (Kaiser, 1958). When an oblique rotation is applied, the cluster-specific components become correlated to some extent. In that case, the loadings should not be read as correlations, but they can be interpreted similarly as weights that indicate the extent to which each variable is influenced by the respective components.

Steps to take when performing multiblock component analysis

When applying one of the multiblock component methods in practice, three steps have to be taken: checking the data requirements, running the analysis, and selecting the model. In the following subsections, each of these steps will be discussed in more detail.

Checking data requirements

As a first step, one needs to check whether the different data blocks contain a sufficient number of observations, whether the data have been preprocessed adequately, and whether and which data are missing. For the different component models to be identified, the number of observations N_i in the data blocks should always be larger than the number of components Q to be fitted. Moreover, when the observations in the data blocks and/or the data blocks themselves are a random sample, this sample needs to be sufficiently large and representative; otherwise, the generalizability of the obtained results is questionable.

With respect to preprocessing, as discussed in the [Preprocessing](#) section, it is often advisable to autoscale all data blocks, which is done automatically by the MBCA program. However, autoscaling is not possible when a variable displays no variance within one or more data blocks, which may occur in empirical data. For instance, for

our emotion example, it is conceivable that some subjects rate a certain negative emotion to be absent for all target persons. In such cases, one of the following options can be considered. First, one may remove the variables that are invariant for one or more data blocks. Second, one may discard the data blocks for which one or more variables are invariant. When many variables or data blocks are omitted, this leads to a great loss of data, however. Therefore, a third option, which is also provided in the MBCA software, is to replace the invariant scores by zeros, implying that the variables in question have a mean of zero but also a variance of zero in some of the data blocks. This strategy has the disadvantage that the interpretation of the component loadings becomes less straightforward. Specifically, even the loadings on orthogonal components can no longer be interpreted as correlations between the variables and the respective components.

With respect to missing data, when the multiblock data contain missing values, it seems straightforward to autoscale each variable with respect to the nonmissing data entries only. This way the nonmissing values for each variable will have a mean of zero and a variance of one per data block, regardless of any assumed or imputed values for the missing data. It may also be wise to remove variables that are missing completely within certain data blocks (i.e., an entire column of a data block is missing), since such missingness patterns are rather likely to be NMAR and, hence, yield biased analysis results.

Running the analysis

The second step consists of performing the multiblock component analysis, with the appropriate number of components Q and number of clusters K in case of clusterwise SCA-ECP. Given (K and) Q , the aim of the analysis is to find a solution that minimizes the following loss function:

$$L = \sum_{i=1}^I \|\mathbf{X}_i - \widehat{\mathbf{X}}_i\|^2, \quad (2)$$

where $\widehat{\mathbf{X}}_i$ equals $\sum_{k=1}^K p_{ik} \mathbf{F}_i^k \mathbf{B}^{k'}$, $\mathbf{F}_i \mathbf{B}_i'$, and $\mathbf{F}_i \mathbf{B}'$ for clusterwise SCA-ECP, separate PCAs, and SCA-ECP, respectively. In case the data contain missing values, $N_i \times J$ binary weight matrices \mathbf{W}_i , containing zeros if the corresponding data entries are missing and ones if not, are included in the loss function:

$$L = \sum_{i=1}^I \|(\mathbf{X}_i - \widehat{\mathbf{X}}_i) * \mathbf{W}_i\|^2. \quad (3)$$

Note that $*$ denotes the Hadamard (i.e., elementwise) product. On the basis of the loss function value L , the

percentage of variance accounted for (VAF) can be computed as follows:

$$VAF(\%) = \frac{\|\mathbf{X}\|^2 - L}{\|\mathbf{X}\|^2} \times 100. \quad (4)$$

The algorithms for estimating the multiblock component models and for missing data imputation are described in the following subsections.

Algorithms In this section, we discuss the algorithms for performing separate PCAs, SCA-ECPs, and clusterwise SCA-ECPs. Each of these algorithms is based on a singular value decomposition. However, unlike the separate PCA algorithm, which boils down to the computation of a closed form solution, the SCA-ECP and clusterwise SCA-ECP algorithms are iterative procedures.

Separate PCAs for each of the data blocks are obtained on the basis of the singular value decomposition of data block \mathbf{X}_i into \mathbf{U}_i , \mathbf{S}_i , and \mathbf{V}_i , with $\mathbf{X}_i = \mathbf{U}_i \mathbf{S}_i \mathbf{V}_i'$ (Jolliffe, 2002). Least squares estimators of \mathbf{F}_i and \mathbf{B}_i are $\mathbf{F}_i = \sqrt{N_i} \mathbf{U}_{i(Q)}$ and $\mathbf{B}_i = \frac{1}{\sqrt{N_i}} \mathbf{V}_{i(Q)} \mathbf{S}_{i(Q)}$, respectively, where $\mathbf{U}_{i(Q)}$ and $\mathbf{V}_{i(Q)}$ are the first Q columns of \mathbf{U}_i and \mathbf{V}_i respectively, and $\mathbf{S}_{i(Q)}$ consists of the first Q rows and columns of \mathbf{S}_i .

To estimate the SCA-ECP solution, an alternating least squares (ALS) procedure is used (see Timmerman & Kiers, 2003, for more details) that consists of four steps:

1. *Rationally initialize loading matrix \mathbf{B}* : Initialize \mathbf{B} by performing a singular value decomposition of the total data matrix \mathbf{X} into \mathbf{U} , \mathbf{S} , and \mathbf{V} , with $\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}'$. A rational start of \mathbf{B} is then given by $\mathbf{B} = \mathbf{V}_{(Q)}$, where $\mathbf{V}_{(Q)}$ contains the first Q columns of \mathbf{V} .
2. *(Re)estimate the component score matrices \mathbf{F}_i* : For each data block, decompose $\mathbf{X}_i \mathbf{B}$ into \mathbf{U}_i , \mathbf{S}_i , and \mathbf{V}_i , with $\mathbf{X}_i \mathbf{B} = \mathbf{U}_i \mathbf{S}_i \mathbf{V}_i'$. A least squares estimate of the component scores \mathbf{F}_i for the i th data block is then given by $\mathbf{F}_i = \sqrt{N_i} \mathbf{U}_i \mathbf{V}_i'$ (ten Berge, 1993).
3. *Reestimate the loading matrix \mathbf{B}* : $\mathbf{B} = ((\mathbf{F}' \mathbf{F})^{-1} \mathbf{F}' \mathbf{X})'$, where \mathbf{F} is the vertical concatenation of the component scores of all data blocks.
4. Repeat steps 2 and 3 until the decrease of the loss function value L for the current iteration is smaller than the convergence criterion, which is $1e-6$ by default.

Clusterwise SCA-ECP solutions are also estimated by means of an ALS procedure (see De Roover, Ceulemans, Timmerman, Vansteelandt, et al., *in press*, for more details):

1. *Randomly initialize partition matrix \mathbf{P}* : Randomly assign the I data blocks to one of the K clusters, where each cluster has an equal probability of being assigned to. If one of the clusters is empty, repeat this procedure until all clusters contain at least one element.

2. *Estimate the SCA-ECP model for each cluster:* Estimate the \mathbf{F}_i and \mathbf{B}^k matrices for each cluster k by performing a rationally started SCA-ECP analysis, as described above, on the \mathbf{X}_i data blocks assigned to the k th cluster.
3. *Reestimate the partition matrix \mathbf{P} :* Each data block \mathbf{X}_i is tentatively assigned to each of the K clusters. On the basis of the loading matrix \mathbf{B}^k of the cluster k and the data block \mathbf{X}_i , a component score matrix for block i in cluster k is computed, and the fit of data block i in cluster k is evaluated. Eventually, the data block is assigned to the cluster for which the fit is maximal. When one of the K clusters is empty after this procedure, the data block with the worst fit in its current cluster is moved to the empty cluster.
4. Steps 2 and 3 are repeated until the decrease of the loss function value L for the current iteration is smaller than the convergence criterion, which is $1e-6$ by default.

Note that the clusterwise SCA-ECP algorithm may end in a local minimum. Therefore, it is advised to use a multistart procedure (e.g., 25 starts; see De Roover, Ceulemans, Timmerman, Vansteelandt, et al., [in press](#)) with different random initializations of the partition matrix \mathbf{P} .

Missing data imputation To perform missing data imputation while fitting multiblock component models, weighted least squares fitting (Kiers, 1997) is used to minimize the weighted loss function (Eq. 3). This iterative procedure, which assumes the missing values to be missing at random (MAR), consists of the following steps:

1. Set t , the iteration number, to one. Initialize the $N \times J$ missing values matrix \mathbf{M}^t by sampling its values from a standard normal distribution (random start) or by setting all entries to zero (zero start).
2. Compute the imputed data matrix $\tilde{\mathbf{X}}^t = \mathbf{W} * \mathbf{X} + \mathbf{W}^c * \mathbf{M}^t$, where \mathbf{W}^c is the binary complement of \mathbf{W} (i.e., with ones for the missing values and zeros for the nonmissing values).
3. Perform a multiblock component analysis on $\tilde{\mathbf{X}}^t$ (see the Algorithms section).
4. Set $t = t + 1$ and $\mathbf{M}^t = \hat{\mathbf{X}}^t$, where $\hat{\mathbf{X}}^t$ holds the reconstructed scores from step 3.
5. Steps 2–4 are repeated until the decrease of the loss function value L for the current iteration is smaller than the convergence criterion, which is set to $1e-6$ times 10% of the data size $N \times J$; the latter product is added to keep the computation time for larger data sets under control.

In the MBCA program, the described procedure is performed with five different starts (i.e., one zero start and four random starts) for the missing values matrix \mathbf{M}^t , and the best solution is retained. Note that the computation time

will be considerably longer when missing data imputation is performed.

A simulation study was performed to investigate how the clusterwise SCA-ECP algorithm with missing data imputation performs in terms of goodness of recovery. A detailed description of the simulation study is provided in Appendix 1. From the study, which included missing data generated under different mechanisms, it can be concluded that the clustering of the data blocks, as well as each of the cluster loading matrices, is recovered very well in all simulated conditions. The overall mean computation time in the simulation study amounts to 22 min and 25 s, which is about 260 times longer than the computation time of clusterwise SCA-ECP on the complete data sets.

Model selection

In the previous step, the number of components Q and number of clusters K were assumed to be known. This is often not the case, however. In component analysis, the resulting model selection problem is often solved by fitting component models with different numbers of components and then selecting the model with the best balance between complexity and fit. To this end, a generalization of the well-known scree test (Cattell, 1966) can be used, based on a plot of the VAF (Eq. 4) against the number of components. Using this plot, the “best” number of components is determined by searching for the number of components after which the increase in fit with additional components levels off. The decision may be based on a visual inspection of the scree plot, but a number of automated scree test procedures have been proposed as well (e.g., DIFFIT, Timmerman & Kiers, 2000; CHULL, Ceulemans & Kiers, 2006).

Building on the CHULL procedure, we propose to select the component solution for which the scree ratio

$$sr_{(Q)} = \frac{VAF_Q - VAF_{Q-1}}{VAF_{Q+1} - VAF_Q} \quad (5)$$

is maximal, where VAF_Q is the VAF of a solution with Q components. Note that the lowest and highest number of components considered will never be selected, since, for them, the scree ratio (Eq. 5) cannot be calculated. For selecting among separate PCA solutions or SCA-ECP solutions, this scree criterion can readily be applied. For clusterwise SCA-ECP, model selection is more intricate, however, because the number of clusters also needs to be determined (which is analogous to the problem of determining the number of mixture components in mixture models; e.g., McLachlan & Peel, 2000). As a way out, one may use a two-step procedure in which, first, the best number of clusters is determined and, second, the best

number of components. More specifically, the first step of this procedure starts by calculating the scree ratio $sr_{(K|Q)}$ for each value of K , given different values of Q :

$$sr_{(K|Q)} = \frac{VAF_K - VAF_{K-1}}{VAF_{K+1} - VAF_K}. \quad (6)$$

Subsequently, for each number of components Q , the best number of clusters K is the number of clusters for which the scree ratio is maximal. The overall best number of clusters K^{best} is determined as the K -value that has the highest average scree ratio across the different Q -values. The second step aims at selecting the best number of components. To this end, given K^{best} , the scree ratios are calculated for each number of components Q :

$$sr_{(Q|K^{\text{best}})} = \frac{VAF_Q - VAF_{Q-1}}{VAF_{Q+1} - VAF_Q}. \quad (7)$$

The best number of components Q^{best} is the number of components Q for which the scree ratio is maximal.

We applied this procedure for selecting an adequate clusterwise SCA-ECP solution for the emotion data out of solutions with one to six clusters and one to six components. Table 2 contains the scree ratios for determining the number of clusters and the number of components. Upon inspection of the $sr_{(K|Q)}$ ratios in Table 2 (above), we conclude that the best number of clusters differs over the solutions with one to six components. Therefore, we computed the average scree ratios across the different numbers of components, which equaled 1.88, 2.01, 1.08, and 1.32 for two to five clusters, respectively, and decided that we should retain three clusters. The $sr_{(Q|K^{\text{best}})}$ values in Table 2 (below) suggest that the best number of components Q^{best} is two. Hence, we selected the model with three clusters and two components, which was discussed in the [Data structure, preprocessing, and missing values](#) section.

To evaluate the model selection procedure, we performed a simulation study, of which details can be found in Appendix 2. The simulation study revealed that the model selection procedure works rather well, with a correctly

selected clusterwise SCA-ECP model in 91% of the simulated cases.

Multiblock component analysis program

The most up-to-date versions of the MBCA software can be downloaded from <http://ppw.kuleuven.be/okp/software/MBCA/>. When clicking on the .exe file (for the stand-alone version) or typing “MultiBlock_Component_Analysis” in the MATLAB command window (for the MATLAB version), the interface of the software program (see Fig. 1) opens, consisting of three panels: “data description and data files,” “analysis options,” and “output files and options.” In this section, first, the functions of each panel of the software interface are clarified. Next, performing the analysis and error handling are described. Finally, the format and content of the output files is discussed.

Data description and data files

In the “data description and data files” panel, the user first loads the data by clicking the appropriate “browse” button and selecting the data file. This file should be an ASCII (.txt) file, in which the data blocks are placed below each other, with the rows representing the observations and the columns representing the variables. The columns may be separated by a semicolon, one or more spaces, or horizontal tabs (see Fig. 3). Missing values should be indicated in the data file by “.”, “/,” “*” or the letter “m” (e.g., in Fig. 3, the missing data values are indicated by “m”). If some data are missing, the user should select the option “missing data, indicated by . . .” in the “missing data imputation” section of the panel and specify the symbol by which the missing values are indicated.

Next, the user selects a “number of rows” file (also an ASCII file) by using the corresponding “browse” button. The selected file should contain one column of integers, indicating how many observations the consecutive data blocks contain, where the order of the numbers corresponds to the order of the data blocks in the data file (see Fig. 3).

Table 2 Scree ratios for the numbers of clusters K given the numbers of components Q and averaged over the numbers of components (above), and for the numbers of components Q given three clusters (below), for the emotion data. The maximal scree ratio in each column is highlighted in boldface

	1 Comp	2 Comp	3 Comp	4 Comp	5 Comp	6 Comp	Average
2 clusters	1.97	2.54	1.85	1.52	1.53	1.85	1.88
3 clusters	1.64	1.11	3.03	2.82	2.04	1.42	2.01
4 clusters	1.38	0.93	0.71	0.98	1.02	1.48	1.08
5 clusters	0.90	2.09	0.83	0.88	2.41	0.82	1.32
3 Clusters							
2 components	1.80						
3 components	1.30						
4 components	1.43						
5 components	1.21						

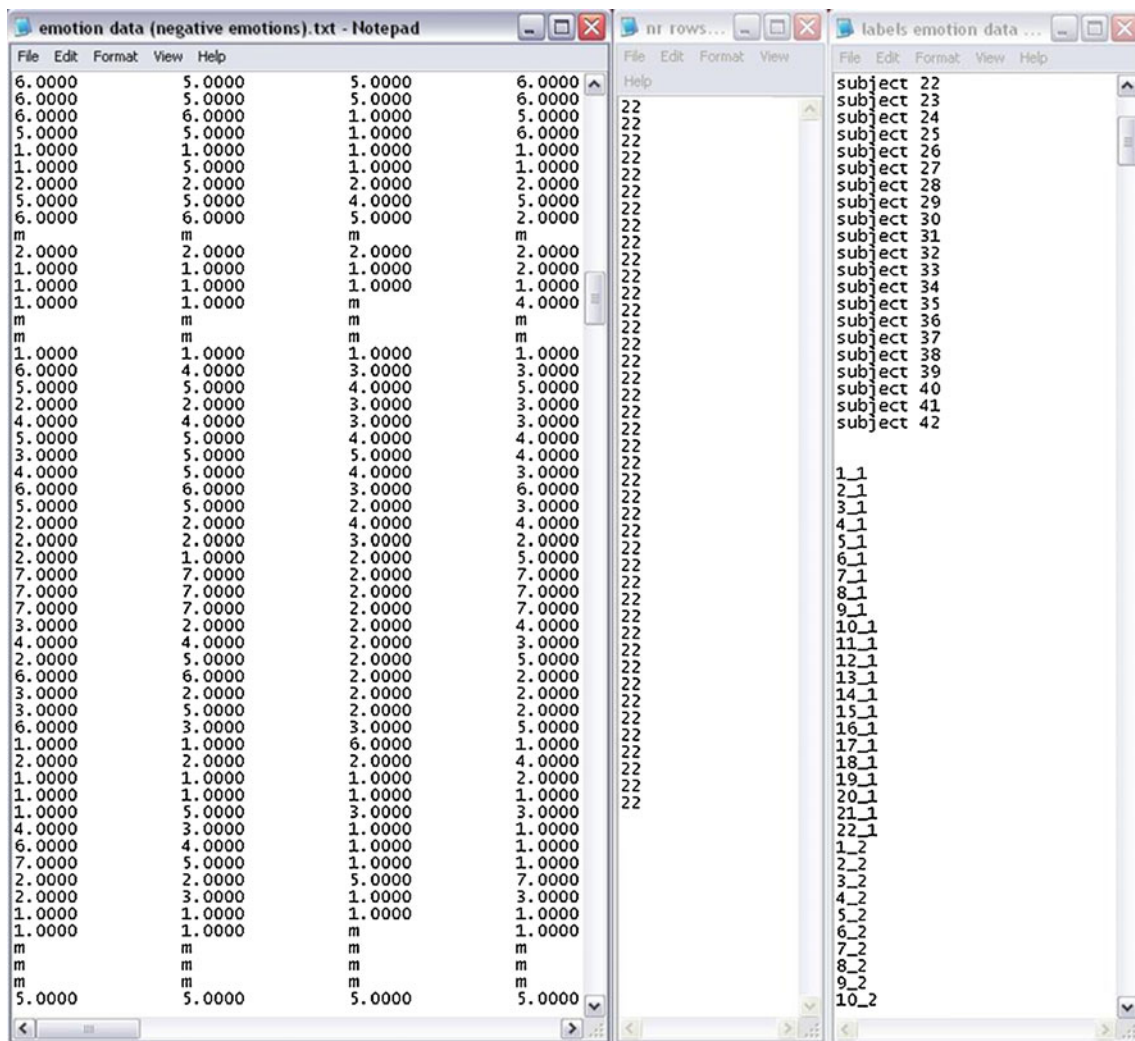


Fig. 3 Screenshot of (from left to right) a data file, a number of rows file, and a labels file. An "m" in the data file indicates a missing value

Finally, the user may choose to upload a file with meaningful labels for the data blocks, the observations within the data blocks, and the variables. The labels file should be an ASCII file containing three groups of labels, in the form of strings that are separated by empty lines, in the following order: block labels, object labels, and variable labels. Note that tabs are not allowed in the label strings. If the user does not load a labels file, the option "no (no labels)" in the right-hand part of the panel is selected. In that case, the program will use default labels in the output (e.g., "block1" for the first data block, "block1, obs1" for the first object of the first data block, and "column1" for the first variable).

Analysis options

In the "type of analysis" section of the "analysis options" panel, the user can choose which types of multiblock component analysis need to be performed, on the basis of the expected differences and/or similarities between the

underlying structure of the different data blocks (as was explained in the Introduction). The user selects at least one of the methods: clusterwise SCA-ECP, separate PCA per data block, and SCA-ECP.

In the case of clusterwise SCA-ECP analysis, the user specifies the number of clusters in the "complexity of the clustering" section. The maximum number of clusters is 10, unless the data contain fewer than 10 data blocks (in that case, the maximum number of clusters is the number of data blocks). In addition to that, the user chooses one of the following two options: "analysis with the specified number of clusters only" or "analyses with 1 up to the specified number of clusters." In the latter case, the software generates solutions with one up to the specified number of clusters and specifies which number of clusters should be retained according to the model selection procedure (see the [Model Selection](#) section).

In the "complexity of the component structure" section, the user specifies a number of components between 1 and 10. Just as in specifying the number of clusters (for

clusterwise SCA-ECP), the user can choose to perform the selected analyses with one up to the specified number of components or with the specified number of components only. In the former case, the model selection procedure (described in the [Model selection](#) section) will be applied to suggest what the best number of components is.

Finally, in the “analysis settings” section, the user can indicate how many random starts will be used, with a maximum of 1,000. The default setting is 25 random starts, on the basis of a simulation study by De Roover et al. ([in press](#)).

Output files and options

In the panel “output files and options,” the user indicates, by clicking the appropriate “browse” button, the directory in which the output files are to be stored. The user may also specify a meaningful label for the output files, to be able to differentiate among different sets of output files (for instance, for different data sets) and to avoid the output files to be overwritten next time the software program is used. The specified label is used as the first part of the name of each output file, while the last part of the file names refers to the content of the file and is added by the program. It is important to note that the label for the output files should not contain empty spaces.

In the “required output” section, the parameters to be printed in the output files can be selected. More specifically, the user indicates whether output files with unrotated, orthogonally rotated, and/or obliquely rotated loadings are needed and whether the component scores—counterrotated accordingly—are to be printed in those output files as well. Note that the output files often become very large when component scores are printed. For orthogonal rotation of the component matrices, the normalized varimax rotation (Kaiser, 1958) is used, while oblique rotation is performed according to the HKIC criterion (Harris & Kaiser, 1964; Kiers & ten Berge, 1994a).

Analysis

Performing the analyses

After specifying the necessary files and settings, as described in the previous sections, the user clicks the “run analysis” button to start the analysis. The program will start by reading and preprocessing the data; then the requested analyses are performed. During the analysis, the status of the analyses is displayed in the box at the bottom of the software interface, such that the user can monitor the progress. The status information consists of the type of analysis being performed at that time and the number of (clusters and) components being used (see Fig. 1). For clusterwise SCA-ECP analysis,

the random start number is included in the status. When analyses with missing data imputation are performed, the start number and iteration number of the imputation process are added to the status as well. When the analysis is done, a screen pops up to notify the user. After clicking the “OK” button, the user can consult the results in the output files stored in the selected output directory.

Error handling

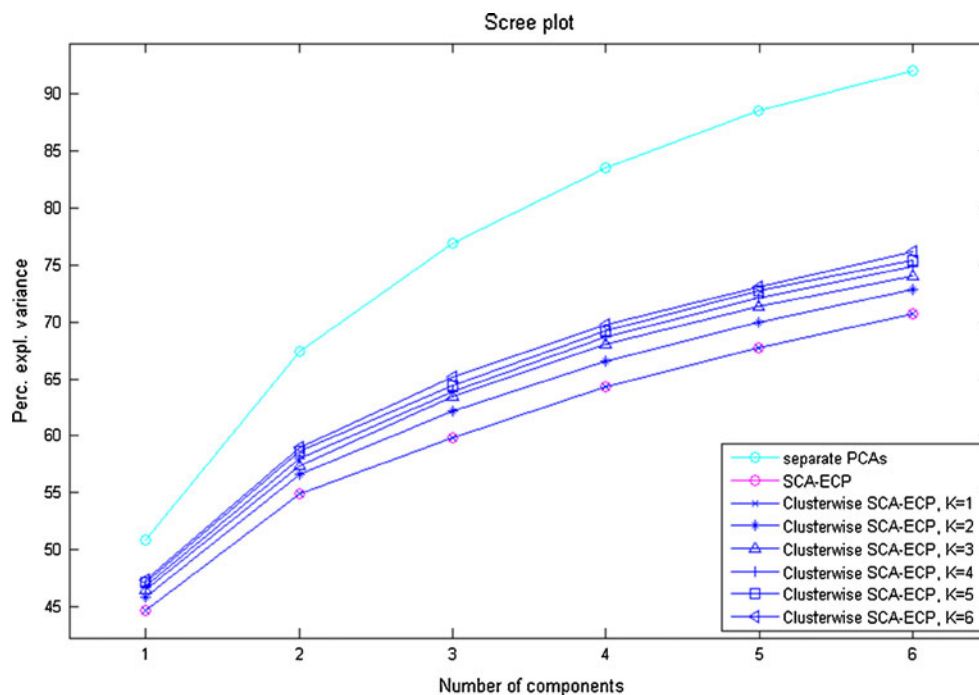
If the files or options are not correctly specified, one or more error screens will appear, with indications of the errors. After clicking “OK,” the analysis stops, and the content of the error messages is displayed in the box at the bottom of the interface. The user can then correct the files or settings and click “run analysis” again.

In some cases, a warning screen may appear. Specifically, a warning is given when missing data imputation is requested but no missing values are found, when missing data imputation is requested and the analyses are expected to take a very long time (i.e., when more than 10% of the data are missing and/or when more than 20 different analyses are requested, where each analysis refers to a particular K and Q value), or when some variables have a variance of zero for one or more data blocks (see the [Checking data requirements](#) section). In the latter case, a warning screen appears with the three options for dealing with invariant variables (as described in the [Checking data requirements](#) section). For the first two options, the number of data blocks or variables that would have to be removed for the data set at hand is stated between brackets. In addition to these three options, a fourth option is given that is a reference to a future upgrade of the software program containing a different variant of clusterwise SCA (i.e., clusterwise SCA-P; De Roover, Ceulemans, Timmerman, & Onghena, 2011). Also, a text file with information on which variables are invariant within which data blocks is created in the output directory and opened together with the warning screen. When the user chooses to continue the analysis, the third solution for invariant variables (i.e., replacing the invariant scores by zeros) is applied automatically by the software program. Otherwise, the user can click “no” to stop the analysis and remove data blocks and/or variables to solve the problem.

Output files

The MBCA program creates separate ASCII (.txt) output files for each combination of multiblock component method (separate PCAs, SCA-ECP, and/or clusterwise SCA-ECP) and rotation method (unrotated, orthogonal, and/or oblique; see Fig. 2 for an example). For each used number of (clusters and) components, these output files

Fig. 4 Percentage of explained variance for separate PCA, SCA-ECP and clusterwise SCA-ECP solutions, with the number of components varying from one to six and the number of clusters for clusterwise SCA-ECP varying from one to six for the emotion data



contain all obtained component loadings and, if requested, the component scores. For separate PCAs, the output is organized per data block. When the solutions are obliquely rotated, the component correlations are added to the output file in question. For separate PCAs, SCA-ECP, and clusterwise SCA-ECP, these correlations are respectively computed for each data block, across all data blocks and across all data blocks within a cluster. In the clusterwise SCA-ECP output files (e.g., Fig. 2), the partition matrices are printed as well.

In addition to the ASCII output files, the software program creates an output overview (.mht) file. For data with missing values, this file contains the percentage of missing values per data block and the total percentage of missing data. The file also displays the overall fit values for each of the performed analyses. When analyses are performed for at least four different numbers of clusters and/or components, the overview file shows the results of the model selection procedures for each component method. Specifically, the overview file suggests how many components and, if applicable, how many clusters should be retained. Sometimes, for clusterwise SCA-ECP, no suggestion can be made with respect to the number of clusters that should be used—for instance, because only two or three numbers of clusters are used. In that case, the best number of components is indicated for each number of clusters separately.

To further facilitate model selection, the output overview provides a scree plot (e.g., Fig. 4) in which the percentage of explained variance is shown as a function of the number of components for each number of clusters separately. When separate PCAs or SCA-ECP analyses are performed, an additional scree line is added to the scree plot. Moreover, all

the computed scree ratios are printed. Note that for clusterwise SCA-ECP analyses, a table of scree ratios is provided for each number of clusters, given the different numbers of components and averaged over the numbers of components (e.g., Table 2), as well as a similar table for the numbers of components given the different numbers of clusters. On the basis of these tables, the user can select additional solutions for further consideration. Of course, the interpretability of the different solutions should also be taken into account.

Finally, the output overview provides information on the fit of the different data blocks within all obtained solutions. This information can be consulted to detect data blocks that are aberrant (i.e., fitting poorly) within a certain model.

Conclusion

Behavioral research questions may concern the correlational structure of multivariate multiblock data. To explore this structure, a regular PCA or EFA is inappropriate, because this would mix up the between-block differences in means and in correlational structures. In this article, we gave an overview of more sophisticated factor analysis and component analysis techniques that have been proposed for investigating structural differences and similarities between blocks of data. We focused on multiblock component analysis, because this is a flexible approach that proved its usefulness in empirical practice. Moreover, for clusterwise SCA-ECP, which is the most general multiblock component model, no counterpart exists in factor analysis. An example from emotion research illustrated the value of this approach.

To facilitate the use of multiblock component analysis, we introduced the MBCA program and provided guidelines on how to perform a multiblock component analysis in practice.

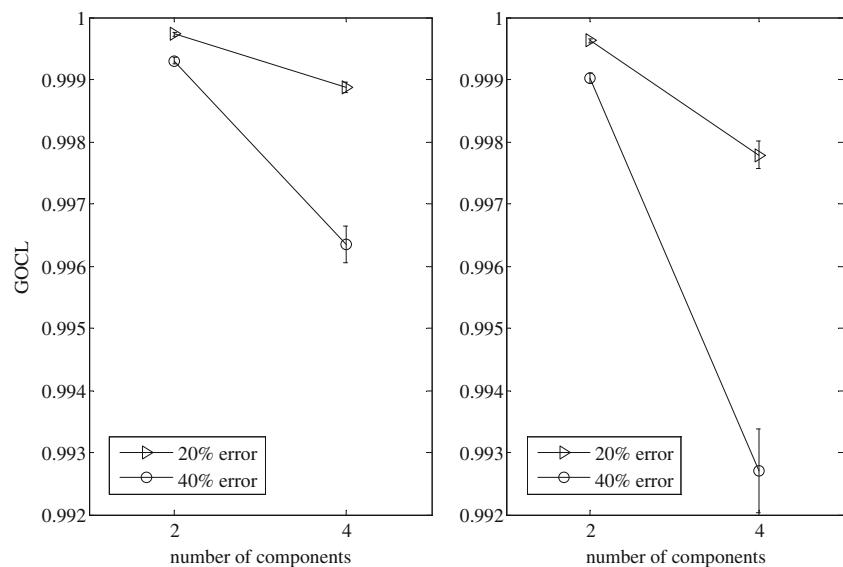
Author Notes The research reported in this article was partially supported by the fund for Scientific Research-Flanders (Belgium), Project No. G.0477.09 awarded to Eva Ceulemans, Marieke Timmerman, and Patrick Onghena, and by the Research Council of K. U. Leuven (GOA/2010/02).

Appendix 1 Simulation study to evaluate the performance of the missing data imputation procedure

To evaluate the missing data imputation procedure in terms of goodness of recovery, a simulation study was performed using the clusterwise SCA-ECP algorithm. The number of observations N_i within the data blocks was sampled uniformly between 80 and 120. Keeping the number of variables J fixed at 12 and the number of data blocks I at 40, six factors were manipulated and completely crossed:

1. the *missingness mechanism* at three levels: MCAR, MAR, NMAR (see the [Missing values](#) section);
2. the *percentage of missing values* at two levels: 10%, 25%;
3. the *number of clusters* K at two levels: 2, 4;
4. the *number of components* Q at two levels: 2, 4;
5. the *cluster size*, at three levels (see Milligan, Soon, & Sokol, 1983): equal (equal number of data blocks in each cluster), unequal with minority (10% of the data blocks in one cluster and the remaining data blocks distributed equally over the other clusters), unequal with majority (60% of the data blocks in one cluster and the remaining data blocks distributed equally over the other clusters);

Fig. 5 Mean GOCL and associated 95% confidence intervals as a function of the number of components and the amount of error variance ($e \times 100\%$) for 10% missing values (left panel) and for 25% missing values (right panel)



6. the *error level* e , which is the expected proportion of error variance in the data blocks \mathbf{X}_i , at two levels: .20, .40.

For each cell of the design, five data matrices \mathbf{X} were generated, consisting of I \mathbf{X}_i data blocks. These data blocks were constructed as follows:

$$\mathbf{X}_i = \mathbf{F}_i \mathbf{B}^{k'} + \mathbf{E}_i \quad (8)$$

where the entries of the component score matrices \mathbf{F}_i were randomly sampled from a multivariate normal distribution, of which the mean vector consists of zeros and of which the variance–covariance matrix was the identity matrix, and where the entries of the error matrices \mathbf{E}_i were randomly sampled from a standard normal distribution. To construct the partition matrix \mathbf{P} , the data blocks were randomly assigned to the clusters, subject to the restriction imposed by factor 5. The cluster loading matrices \mathbf{B}^k were obtained by sampling the loadings uniformly between -1 and 1 (see De Roover, Ceulemans, Timmerman, Vansteelandt, et al., [in press](#)). The congruence between the cluster loading matrices was relatively low, as indicated by Tucker congruence coefficients (Tucker, 1951): The congruence coefficients between the corresponding components of the cluster loading matrices amounted to .41 on average, when these matrices were orthogonally procrustes rotated to each other. Subsequently, the error matrices \mathbf{E}_i and the cluster loading matrices \mathbf{B}^k were rescaled—by multiplying these matrices with \sqrt{e} and $\sqrt{(1-e)}$ respectively—to obtain data that contained the desired expected proportion e of error variance (factor 6). Next, the resulting \mathbf{X}_i matrices were standardized columnwise and were vertically concatenated into the matrix \mathbf{X} .

Subsequently, within each cluster, a subset of the data entries (factor 2) was selected to be set missing. The

procedures to simulate MCAR, MAR, and NMAR (factor 1) were taken from Van Ginkel, Kroonenberg, and Kiers (2010). In order to obtain missing values that were MCAR, this subset was selected completely at random. To simulate missingness at random (MAR), we first determined within each cluster which variable had the highest average correlation with the rest of the variables (in what follows, we will refer to this variable as the *MAR variable*). Next, we set a subset of the values on the remaining variables as missing, where the probability that entry x_{nij} was set missing was based on a logistic transformation of the value of the same object n_i on the MAR variable. To obtain NMAR missingness, the probability that x_{nij} was set missing depended on a logistic transformation of x_{nij} itself.

In total, 3 (missingness mechanism) \times 2 (percentage of missing values) \times 2 (number of clusters) \times 2 (number of components) \times 3 (cluster size) \times 2 (error level) \times 5 (replicates) = 720 simulated data matrices were generated. Each data matrix was analyzed with the missing data imputation algorithm for clusterwise SCA-ECP analysis, using the correct values for the number of clusters K and components Q and 25 random starts.

To examine the goodness of recovery of the clustering of the data blocks, the *Adjusted Rand Index* (ARI; Hubert & Arabie, 1985) was computed between the true partition of the data blocks and the estimated partition. The ARI equaled 1 if the two partitions were identical and equaled 0 when the overlap between the two partitions was at chance level. With an overall mean ARI of 1.00 ($SD = .00$) the clusterwise SCA-ECP algorithm appeared to recover the clustering of the data blocks perfectly in all simulated conditions.

To evaluate the recovery of the cluster loading matrices, we obtained a goodness-of-cluster-loading-recovery statistic (GOCL) by computing congruence coefficients φ (Tucker, 1951) between the components of the true and estimated loading matrices and averaging across components and clusters (for more details, see De Roover, Ceulemans, Timmerman, Vansteelandt, et al., *in press*). The GOCL statistic takes values between 0 (*no recovery at all*) and 1 (*perfect recovery*). In the simulation study, the overall mean GOCL amounted to .9979 ($SD = .003$), implying an excellent recovery of the cluster loading matrices. An analysis of variance was performed with GOCL as the dependent variable and the six factors as independent variables. Only discussing effects that accounted for more than 5% of the variance in GOCL, the analysis revealed main effects of the number of components (intraclass correlation $\hat{\rho}_I = .27$) and the amount of error variance ($\hat{\rho}_I = .09$): The GOCL was lower for a higher number of components when more error variance was present in the data (Fig. 5). Also, interactions of the number of components with the amount of error variance ($\hat{\rho}_I = .23$), the percentage of missing values ($\hat{\rho}_I = .08$), and the

different combinations of percentage of missing values and amount of error variance ($\hat{\rho}_I = .07$), were found. These interactions imply that the effect of the number of components on GOCL is more outspoken when the data contain more error and/or when more data are missing (Fig. 5).

Appendix 2 Simulation study to evaluate the performance of the model selection procedure

To evaluate whether the proposed model selection procedure succeeds in selecting among clusterwise SCA-ECP solutions, the following seven factors were systematically varied in a complete factorial design, while keeping the number of variables J fixed at 12:

1. the *number of data blocks* I at two levels: 20, 40;
2. the *number of observations per data block* N_i at two levels: N_i sampled uniformly between 30 and 70, N_i sampled uniformly between 80 and 120;
3. the *number of clusters* K at two levels: 2, 4;
4. the *number of components* Q at two levels: 2, 4;
5. the *cluster size*, at three levels: see factor 5 in Appendix 1;
6. the *error level* e , which is the expected proportion of error variance in the data blocks \mathbf{X}_i , at two levels: .20, .40.
7. the *congruence of the cluster loading matrices* \mathbf{B}^k at three levels: low congruence, medium congruence, and high congruence, where low, medium, and high imply that the Tucker congruence coefficients (Tucker, 1951) between the corresponding components of the cluster loading matrices amount to .41, .72, and .93, on average, when these matrices are orthogonally procrustes rotated to each other.

For each cell of the design, five data matrices \mathbf{X} were generated, using the data construction procedure described in Appendix 1. The cluster loading matrices \mathbf{B}^k were generated according to the procedure described by De Roover et al. (*in press*), where the low and high congruence are simulated using randomly sampled loadings and the medium congruence by using simple structure loadings. The resulting 2 (number of data blocks) \times 2 (number of observations per data block) \times 2 (number of clusters) \times 2 (number of components) \times 3 (cluster size) \times 2 (error level) \times 3 (congruence of cluster loading matrices) \times 5 (replicates) = 1,440 simulated data matrices were analyzed with the clusterwise SCA-ECP algorithm, with the number of clusters K and components Q varying from one to six and using 25 random starts per analysis. Subsequently, the model selection procedure described in the [Model selection](#) section was applied on the obtained clusterwise SCA-ECP solutions.

The results of this simulation study indicate that the model selection procedure selected the correct clusterwise SCA-ECP model (i.e., correct K and Q) for 1,310 out of the 1,440 data sets (91%). With respect to the remaining data sets, in 6.6%, 2.2%, and 0.2% of the cases, only K , only Q , and both K and Q , respectively, were selected incorrectly. An analysis of variance was performed with the relative frequency of correct model selection within the cells of the design as the dependent variable and the seven factors as independent variables. The largest intraclass correlation was found for the main effect of the congruence of the cluster loading matrices ($\hat{\rho}_I = .08$): Specifically, the relative frequencies of correct model selection were .97, 1.00, and .76 for the low, medium, and high congruence of the cluster loading matrices (Fig. 6). The latter may seem to be counterintuitive in that it implies that the frequency of correct model selection does not decrease with an increasing congruence of the cluster loading matrices. However, this result can be explained by the data construction procedure, where the low and high congruence level loading matrices consisted of random numbers, while for the medium congruence level, the loadings had simple structure (for more details, see De Roover, Ceulemans, Timmerman, Vansteelandt, et al., [in press](#)). In the simple structure case, each component accounts for about the same proportion of variance, while in the random loadings case, the proportion of explained variance may differ strongly across the components. Consequently, in the latter case, it will be more difficult to distinguish components that are explaining less variance, from the error. In addition to that, most incorrect selections of the number of clusters occur in the conditions with highly congruent cluster loading matrices.

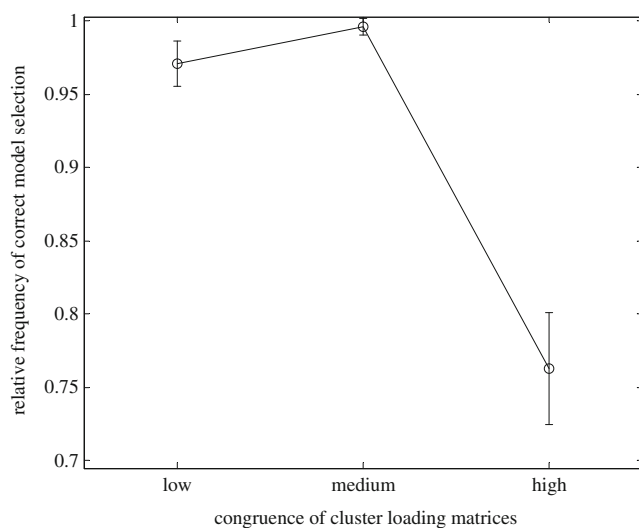


Fig. 6 Mean relative frequencies of correct model selection and associated 95% confidence intervals as a function of the congruence of the cluster loading matrices

References

- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438.
- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*, 579–599.
- Bro, R., & Smilde, A. K. (2003). Centering and scaling in component analysis. *Journal of Chemometrics*, *17*, 16–33.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.
- Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*, 133–150.
- De Leersnyder, J., & Mesquita, B. (2010). *Where do my emotions belong? A study of immigrants' emotional acculturation*. Manuscript submitted for publication.
- De Roover, K., Ceulemans, E., Timmerman, M. E., & Onghena, P. (2011). *A clusterwise simultaneous component method for capturing within-cluster differences in component variances and correlations*. Manuscript submitted for publication.
- De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (in press). Clusterwise simultaneous component analysis for the analysis of structural differences in multivariate multiblock data. *Psychological Methods*.
- Dolan, C., Bechger, T., & Molenaar, P. (1999). Using structural equation modeling to fit models incorporating principal components. *Structural Equation Modeling*, *6*, 233–261.
- Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling*, *16*, 295–314.
- Escoufier, B., & Pagès, J. (1998). *Analyses factorielles simples et multiples* (3rd ed.). Paris: Dunod.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286–299.
- Flury, B. D., & Neuenschwander, B. E. (1995). Principal component models for patterned covariance matrices with applications to canonical correlation analysis of several sets of variables. In W. J. Krzanowski (Ed.), *Recent advances in descriptive multivariate Analysis* (pp. 90–112). Oxford: Oxford University Press.
- Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, *11*, 1–2.
- Harris, C. W., & Kaiser, H. F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, *29*, 347–362.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.
- Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*, 409–426.
- Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8.30*. Chicago: Scientific Software.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.
- Keller, S. D., Ware, J. E., Bentler, P. M., Aaronson, N. K., Alonso, J., Apolone, G., et al. (1998). Use of structural equation modeling to test the construct validity of the SF-36 health survey in ten countries: Results from the IQOLA project. *Journal of Clinical Epidemiology*, *51*, 1179–1188.
- Kiers, H. A. L. (1990). *SCA: A program for simultaneous components analysis of variables measured in two or more populations*. Groningen, The Netherlands: iec ProGAMMA.

- Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, *62*, 251–266.
- Kiers, H. A. L., & ten Berge, J. M. F. (1994a). The Harris–Kaiser independent cluster rotation as a method for rotation to simple component weights. *Psychometrika*, *59*, 81–90.
- Kiers, H. A. L., & ten Berge, J. M. F. (1994b). Hierarchical relations between methods for simultaneous components analysis and a technique for rotation to a simple simultaneous structure. *British Journal of Mathematical and Statistical Psychology*, *47*, 109–126.
- Kim, J. O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, *6*, 215–241.
- Kline, R. B. (2004). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Klingenberg, C. P., Neuenschwander, B. E., & Flury, B. D. (1996). Ontogeny and individual variation: Analysis of patterned covariance matrices with common principal components. *Systematic Biology*, *45*, 135–150.
- Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken, NJ: Wiley.
- Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, *12*, 209–229.
- Lee, L. M. P., & Lam, Y. R. (1988). Confirmatory factor analyses of the Wechsler Intelligence Scale for children–revised and the Hong Kong–Wechsler Intelligence Scale for children. *Educational and Psychological Measurement*, *48*, 895–903.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley Interscience.
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92.
- McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*, 509–516.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.
- Milligan, G. W., Soon, S. C., & Sokol, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *5*, 40–47.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398.
- Muthén, B. O., & Muthén, L. K. (2007). *Mplus user's guide* (5th ed.). Los Angeles: Muthén & Muthén.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th ed.). Richmond, VA: Medical College of Virginia, Department of Psychiatry.
- Novy, D. M., Frankiewicz, R. G., Francis, D. J., Liberman, D., Overall, J. E., & Vincent, K. R. (1994). An investigation of the structural validity of Loewinger's model and measure of ego development. *Journal of Personality*, *62*, 86–118.
- Pastorelli, C., Barbaranelli, C., Cermak, I., Rozsa, S., & Caprara, G. V. (1997). Measuring emotional instability, prosocial behavior and aggression in pre-adolescents: A cross-national study. *Personality and Individual Differences*, *23*, 691–703.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modelling*. London: Sage.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.
- Stumpf, F. A. (1978). A note on handling missing data. *Journal of Management*, *4*, 65–73.
- ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis*. Leiden: DSWO Press.
- Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, *59*, 301–320.
- Timmerman, M. E., & Kiers, H. A. L. (2000). Three-mode principal component analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology*, *53*, 1–16.
- Timmerman, M. E., & Kiers, H. A. L. (2003). Four simultaneous component models of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*, *68*, 105–122.
- Tucker, L. R. (1951). *A method for synthesis of factor analysis studies (Personnel Research Section Rep. No. 984)*. Washington, DC: Department of the Army.
- Tugade, M. M., Fredrickson, B. L., & Barrett, L. F. (2004). Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of Personality*, *72*, 1161–1190.
- Van Deun, K., Smilde, A. K., van der Werf, M. J., Kiers, H. A. L., & Van Mechelen, I. (2009). A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, *10*, 246.
- Van Ginkel, J. R., Kroonenberg, P. M., & Kiers, H. A. L. (2010). *Comparison of five methods for handling missing data in principal component analysis*. Unpublished Manuscript.
- Velicer, W. F., & Jackson, D. N. (1990a). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*, *25*, 97–114.
- Velicer, W. F., & Jackson, D. N. (1990b). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1–28.