Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 6

11-2014

Retained-Components Factor Transformation: Factor Loadings and Factor Score Predictors in the Column Space of Retained Components

André Beauducel
University of Bonn, Bonn, Germany, beauducel@uni-bonn.de

Frank Spohn
University of Hamburg, Hamburg, Germany, frank.spohn@uni-hamburg.de

Recommended Citation
Beauducel, André and Spohn, Frank (2014) "Retained-Components Factor Transformation: Factor Loadings and Factor Score Predictors in the Column Space of Retained Components," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 2 , Article 6.
DOI: 10.22237/jmasm/1414814700

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 106-130.

ISSN 1538 − 9472

Retained-Components Factor Transformation: Factor Loadings and Factor Score Predictors in the Column Space of Retained Components

André Beauducel
University of Bonn
Bonn, Germany

Frank Spohn
University of Hamburg
Hamburg, Germany

Factor loadings optimally account for the non-diagonal elements of the covariance matrix of observed variables. Principal component analysis leads to components accounting for a maximum of the variance of the observed variables. Retained-components factor transformation is proposed in order to combine the advantages of factor analysis and principal component analysis.

Keywords: Factor analysis, principal component analysis, exploratory factor analysis.

Introduction

Common factor analysis (FA) is regularly used in order to identify latent constructs accounting for the covariance of observed variables whereas principal components analysis (PCA) is primarily used in order to explain as much of the variance as possible with a minimum of components (Conway & Huffcutt, 2003; Fabrigar, Wegener, MacCallum, Strahan, 1999; Preacher & MacCallum, 2003). There is a broad literature referring to similarities and differences between FA and PCA (Bentler & De Leeuw, 2011; Ogasawara, 2003; Harris, 2001; Velicer & Jackson, 1990; Unkel & Trendafilov, 2010). It is also known that both methods can produce identical or extremely similar results under specific conditions (Schneeweiss, 1997; Schneeweiss & Mathes, 1995) and that PCA is often used as a substitute for FA (Sato, 1990). Nevertheless, an important difference between FA and PCA is that communalities or unique error variances of the variables are not estimated in PCA whereas they are

Dr. Beauducel is Professor for Psychological Methods, Psychological Assessment, and Evaluation. Email Dr. Beauducel at beauducel@uni-bonn.de. Email Mr. Spohn at frank.spohn@uni-hamburg.de.

estimated in FA (Harman, 1976; Tabachnick & Fidell, 2007). The estimation of communalities avoids an inflation of loadings in FA whereas an inflation of loadings regularly occurs in PCA (Widaman, 1993; Snook & Gorsuch, 1989).

Another difference between PCA and FA is related to the scores resulting from these methods: The component scores are clearly determined in PCA whereas the factor scores are indeterminate in FA (Guttman, 1955; Lovie & Lovie, 1995; Grice, 2001). Moreover, the component scores account for a maximum of the variance of the observed variables so that they represent an optimal data reduction. The principal components represent best summarizers for the observed variables (ten Berge & Kiers, 1997), which might be relevant for psychological assessment. By contrast, in FA different factor score predictors with different advantages and disadvantages have been proposed (Beauducel & Rabe, 2009; Krijnen, 2006; ten Berge, Krijnen, Wansbeek, & Shapiro, 1999), however, there is no factor score predictor that is an optimal summarizer of the observed variables. In consequence, a method that combines an optimal estimation of the loading size (without inflation) with scores that represent an optimal data reduction is not available. Researchers have to decide: If they want to have an optimal representation of a latent construct and the corresponding loadings, they should opt for FA, if they want to get optimal summarizers of the observed variables, they should use PCA. In the present paper we start from the idea that a researcher wants to get both: An optimal (not inflated) loading matrix representing the common variance of latent constructs adequately and optimal summarizers of the observed variables. A method that combines the estimation of loading magnitude of FA with the optimal data reduction of PCA could be the projection of the factor loadings on the column space of the loadings of the components retained in PCA. The focus on the components retained for rotation and interpretation is necessary because typically the number of components retained in PCA is considerably smaller than the number of observed variables. Accordingly, the transformation resulting in factor loadings in the column space of the retained PCA loadings is called ‘retained-components factor transformation’ (RFT). First, some definitions are given, then RFT is introduced and RFT-factor scores summarizing the observed variables like principal component scores are presented. Finally, some properties of RFT are described and illustrated by means of a small simulation study and by means of an empirical example.

Methodology

The principal component model

According to Hotelling (1933) it is possible to decompose **x**, the random vector of observations of order *p* by means of a linear combination of components $\boldsymbol{\zeta}$ with component loadings **L**. The observations ***x*** and the components $\boldsymbol{\zeta}$ are assumed to have an expectation zero ($E[\boldsymbol{x}] = 0$, $E[\boldsymbol{\zeta}] = 0$)

$$\mathbf{x} = \mathbf{L}\zeta \tag{1}$$

This decomposition by means of components represents a population model. The principal component representation implies that **L´L** is diagonal with elements ordered from large to small (ten Berge & Knol, 1985). Moreover, it is assumed that $\mathbf{L} \neq \mathbf{0}$. When the principal component model is applied to sample data, it will be reasonable to distinguish between a random vector $\boldsymbol{\zeta}_\mathbf{r}$ of order *q* representing the intended and substantial variance and therefore the components retained for interpretation and a random vector $\boldsymbol{\zeta}_\mathbf{n}$ of order *p* - *q* representing the unintended or trivial variances and therefore the components not retained for interpretation (Hotelling, 1933). Accordingly, it is necessary to distinguish between **M**, the *p* x *q* loading matrix of the retained components and **N** the *p* x (*p* - *q*) loading matrix of the components not retained with **L = [M | N]**. This yields

$$\mathbf{x} = \mathbf{M}\zeta_r + \mathbf{N}\zeta_n \tag{2}$$

with **M´N = 0** and **N´M = 0** following from **L´L** being diagonal. The population covariance matrix of the observed variables can be decomposed as follows:

$$\Sigma = \mathbf{LL}' = \mathbf{MM}' + \mathbf{NN}' \tag{3}$$

The common factor model

The common factor model that is assumed to hold in the population is given by

$$\mathbf{x} = \Gamma\xi = [\Lambda \mid \Psi]\xi = \Lambda\xi_c + \Psi\xi_u \tag{4}$$

where **x** is the random vector of observations of order *p*, ξ is the random vector of factors consisting of *q* common factors $\boldsymbol{\xi_c}$ and *p* orthogonal unique factors $\boldsymbol{\xi_u}$. $\boldsymbol{\Lambda}$ is the factor pattern matrix of order *p* by *q*, $\boldsymbol{\Psi}$ is the *p* x *p* diagonal unique loading matrix. It is assumed that $\boldsymbol{\Psi}$ contains only positive values and that $\boldsymbol{\Lambda} \neq \mathbf{0}$. The factors ξ are assumed to have an expectation zero (E$[\xi] = \mathbf{0}$) and the standard deviation of ξ is one. Moreover, the expectation of the covariance of $\boldsymbol{\xi_c}$ with $\boldsymbol{\xi_u}$ is zero. The covariance matrix $\boldsymbol{\Sigma}$ can be decomposed into

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}^2 \tag{5}$$

where $\boldsymbol{\Phi}$ represents the *q* by *q* factor correlation matrix.

Results

Retained-components factor transformation

In order to transform the retained component matrix **M** to be as similar as possible to the factor loading matrix $\boldsymbol{\Lambda}$ the following transformation was used:

$$\mathbf{MT} = \Lambda \tag{6}$$

with the transformation matrix **T** and the factor loading matrix $\boldsymbol{\Lambda}$ as a target matrix, much like in procrustes rotation (Hurley & Cattell, 1962). Solving Equation 6 for **T** yields

$$\mathbf{T} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\Lambda \tag{7}$$

Entering **T** into Equation 6 yields $\boldsymbol{\Lambda}^*$, because the transformed component loadings will in most cases be similar, but not identical to the target matrix $\boldsymbol{\Lambda}$. Accordingly, $\boldsymbol{\Lambda}^*$ contains the loadings resulting from retained-components factor transformation (RFT),

$$\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\Lambda = \Lambda^* \tag{8}$$

Equation 2 and Equation 4 both explain the variance of the observed variables **x**, so that they can be equated. This yields

$$\mathbf{M}\zeta_r + \mathbf{N}\zeta_n = \Lambda\xi_c + \Psi\xi_u \tag{9}$$

Premultiplication of Equation 9 with **M(M´M)$^{-1}$M´** yields

$$\begin{aligned}&\mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{M}\zeta_r + \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\mathbf{N}\zeta_n \\ &= \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\Lambda\xi_c + \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\Psi\xi_u\end{aligned} \tag{10}$$

Since **M´N = 0** and according to Equation 8 it is possible to write

$$\mathbf{M}\zeta_r = \Lambda^*\xi_c + \Psi^*\xi_u \tag{11}$$

with **Ψ* = M(M´M)$^{-1}$M´Ψ**. Equation 11 gives a factorial representation of the retained components **ζ$_r$**. Thus, each retained component is decomposed into a projection from the common factors and from the unique factors. According to Equation 2 it is possible to write

$$\mathbf{x} = \Lambda^*\xi_c + \Psi^*\xi_u + \mathbf{N}\zeta_n \tag{12}$$

Thus, the RFT has two error terms: One term representing the unique error of the factorial decomposition of the retained-components and the other error term represents the residual PCA components (i.e., those components that are not retained for interpretation). The covariance matrix of observed variables can be computed from RFT by means of

$$\begin{aligned}\Sigma &= \left(\Lambda^*\xi_c + \Psi\xi_u + \mathbf{N}\zeta_n\right)\left(\Lambda^*\xi_c + \Psi\xi_u + \mathbf{N}\zeta_n\right)' \\ &= \Lambda^*\Phi\Lambda^{*\prime} + \Lambda^*\xi_c\zeta_n{}'\mathbf{N}' + \Psi^{*s} + \Psi^*\xi_u\zeta_n{}'\mathbf{N}' + \mathbf{N}\mathbf{N}'\end{aligned} \tag{13}$$

Postmultiplication of Equation 9 with **ξ$_c$**´, subsequent premultiplication with **(N´N)$^{-1}$N´** and transposing yields

$$\xi_c\zeta_n{}' = \Phi\Lambda'\mathbf{N}\left(\mathbf{N}'\mathbf{N}\right)^{-1} \tag{14}$$

Postmultiplication of Equation 9 with $\boldsymbol{\xi}_{\mathbf{u}}$, subsequent premultiplication with $\left(\mathbf{N}'\mathbf{N}\right)^{-1}\mathbf{N}'$ and transposing yields

$$\xi_u \zeta_n{}' = \Psi \mathbf{N}\left(\mathbf{N}'\mathbf{N}\right)^{-1} \tag{15}$$

Entering Equation 14 and 15 into Equation 13 and some transformation yields

$$\begin{aligned}\Sigma &= \Lambda^* \Phi \Lambda^{*\prime} + \Psi^{*2} + \mathbf{NN}' + \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'(\Lambda\Phi\Lambda' + \Psi^2)\mathbf{N}\left(\mathbf{N}'\mathbf{N}\right)^{-1}\mathbf{N}' \\ &= \Lambda^* \Phi \Lambda^{*\prime} + \Psi^{*2} + \mathbf{NN}' + \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'(\mathbf{MM}' + \mathbf{NN}')\mathbf{N}\left(\mathbf{N}'\mathbf{N}\right)^{-1}\mathbf{N}' \\ &= \Lambda^* \Phi \Lambda^{*\prime} + \Psi^{*2} + \mathbf{NN}' + \mathbf{MM}'\mathbf{N}\left(\mathbf{N}'\mathbf{N}\right)^{-1}\mathbf{N}' \\ &= \Lambda^* \Phi \Lambda^{*\prime} + \Psi^{*2} + \mathbf{NN}'\end{aligned} \tag{16}$$

since $\mathbf{M}'\mathbf{N} = \mathbf{0}$. Thus, the residual covariances that are represented by the loadings of the irrelevant components **N** have no covariance with $\boldsymbol{\Lambda}^*$ of RFT, since $\mathbf{N}'\mathbf{M} = \mathbf{0}$ implies $\mathbf{N}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\boldsymbol{\Lambda} = \mathbf{0}$. In contrast, the residual covariances represented by **N** might be related to the FA-loadings, that is, $\mathbf{N}'\boldsymbol{\Lambda}$ is not necessarily zero. One would therefore expect that advantages of RFT over conventional FA in terms of stability of parameters occur when the PCA residuals in $\mathbf{NN}'$ primarily represent covariances due to sampling error. Moreover, RFT should help to avoid the overestimation of loadings as it occurs with PCA, because the PCA loadings are transformed in order to be as similar as possible to the factor loadings. Accordingly, a simulation study was performed in order to explore the quality of the sample RFT-loadings as estimators of population factor loadings.

Properties of the Retained-components factor score predictors

A main reason for proposing RFT was that it allows for factor score predictors that are optimal summarizers of the observed variables. This property holds for Harman´s ideal-variable factor score predictor. The weights for Harman's (1976) ideal-variable factor score predictor based on RFT are given by

$$\mathbf{B_H} = \Lambda^*\left(\Lambda^{*\prime}\Lambda^*\right)^{-1\prime} \tag{17}$$

Moreover, according to Equation 2, the weights of the retained components are given by

$$\mathbf{B_r} = \mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1} \tag{18}$$

The relation between the component scores and Harman's ideal variables scores for RFT can be expressed in terms of correlations. Therefore, the correlations between weighted composites of the observed variables can be computed. Entering $\mathbf{B_H}$ and $\mathbf{B_r}$ into the formula for the correlations between weighted composites (Harris, 2001), yields

$$\mathbf{B_H}'\Sigma\mathbf{B_r}\left(\mathbf{B_H}'\Sigma\mathbf{B_H}\mathbf{B_r}'\Sigma\mathbf{B_r}\right)^{-0.5} = \mathbf{C_{Hr}} \tag{19}$$

where the main diagonal of $\mathbf{C_{Hr}}$ contains the correlation matrix between Harman's factor score predictor based on RFT and the retained principal components. Entering Equation 17 and 18 into Equation 19 and some transformation yields

$$\mathbf{G}^{-1}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\Sigma\mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\left(\mathbf{G}^{-1}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\Sigma\mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\right)^{-1} = \mathbf{G}^{-1}\mathbf{G} = \mathbf{I} \tag{20}$$

with $\mathbf{G= (M'M)^{-1}M'\Lambda}$. Thus, the correlations between the component scores and Harman's ideal variables scores are all perfect for the RFT-solution. This implies that Harman's ideal variables scores of the RFT-solution are optimal summarizers of the observed variables as are the principal components.

Since Thurstone's (1935) least squares regression score predictor is often used and recommended (Krijnen, 2006), the relationship between the regression score predictor based on RFT and the principal component scores was explored. Since the principal component scores have the interesting property of being the optimal summarizers of the observed variables, they should be regarded as a criterion and the RFT regression factor scores as predictors. Thus, *q* multiple regressions and corresponding multiple correlations can be calculated for the *q* retained components. If the multiple correlations between the regression score predictors and the principal component scores as criterion is one, this indicates that the scores represent the same overall individual differences, even though they might be distributed differently on the factors and components. The weights for Thurstone's regression factor score predictor based on RFT are

$$\mathbf{B_T} = \Sigma^{-1}\Lambda^{*}\Phi = \Sigma^{-1}\mathbf{M}\left(\mathbf{M'M}\right)^{-1}\mathbf{M'}\Lambda\Phi \tag{21}$$

The corresponding regression weights for the prediction of the retained principal components from Thurstone’e regression factor score predictor are

$$\mathbf{B} = (\Phi\Lambda^{*\prime}\Sigma^{-1}\Lambda^{*}\Phi)^{-1}\Phi\Lambda^{*\prime}\ \mathbf{M}\left(\mathbf{M'M}\right)^{-1} \tag{22}$$

The multiple correlation is calculated as

$$\begin{aligned}\mathbf{R}^2 &= \mathbf{B'}\Phi\Lambda^{*\prime}\ \mathbf{M}\left(\mathbf{M'M}\right)^{-1}\left(\mathbf{M'M}\right)^{-1}\mathbf{M'}\Sigma\mathbf{M}\left(\mathbf{M'M}\right)^{-1} \\ &= \left(\mathbf{M'M}\right)^{-1}\mathbf{M'}\Lambda^{*}\Phi(\Phi\Lambda^{*\prime}\Sigma^{-1}\Lambda^{*}\Phi)^{-1}\Phi\Lambda^{*\prime}\ \mathbf{M}\left(\mathbf{M'M}\right)^{-1}\end{aligned} \tag{23}$$

Some transformation yields

$$\mathbf{R}^2 = (\mathbf{M'}\Sigma^{-1}\mathbf{M})^{-1} \tag{24}$$

A singular value decomposition of $\mathbf{\Sigma}$ yields $\mathbf{\Sigma = SDS'}$, with $\mathbf{D}$ containing a diagonal matrix of eigenvalues in descending order. Accordingly it is possible to write

$$\mathbf{LL'} = \mathbf{SD}^{1/2}\mathbf{D}^{1/2}\mathbf{S'} = \mathbf{S} \tag{25}$$

From $\mathbf{\Sigma^{-1} = (SD^{-1}S')'}$ we get $\mathbf{L\Sigma^{-1}L' = SD^{1/2}D^{-1}D^{1/2}S' = I}_{p\ x\ p}$, which implies

$$\mathbf{M}\sum^{-1}\mathbf{M'} = \mathbf{SD}^{1/2}\mathbf{D}^{-1}\mathbf{D}^{1/2}\mathbf{S'} = \mathbf{I_{q\,x\,q}} \tag{26}$$

and, accordingly, $\mathbf{(M'\Sigma^{-1}M)^{-1} = I}_{q\ x\ q}$, which implies that all multiple correlations with the RFT regression factor scores as predictors and each principal component as criterion are one.

Simulation Study

The expectation that the RFT-loadings are more stable than the FA-loadings when the residual covariances represent sampling error was investigated by means of a small simulation study based on orthogonal and oblique three-factor models. For the Models 1 to 4, the population FA-loadings were identical to the population RFT-loadings. Schneeweiss and Mathes (1995) have shown that factor loadings can be perfectly transformed into the retained component loadings when all unique factor loadings are equal. Whenever the retained component loadings can be perfectly transformed into the factor loadings, it follows from Equations 6 and 7 that the RFT-loadings will be identical to the factor loadings ($\mathbf{\Lambda}^* = \mathbf{\Lambda}$), because Equation 7 yields the transformation matrix for the transformation of the retained component loadings into factor loadings. The condition of equal uniqueness of all variables holds for population Models 1 and 2. Moreover, population Models 1 to 4 represent a perfect simple structure (independent clusters) where all non-salient loadings are zero and the salient loadings on each factor are identical even when there are different salient loadings on different factors for Models 3 and 4 (see Table 1). This implies that multiplication with a scalar will allow to transform each vector of factor loadings into the corresponding component loadings. Again, a perfect transformation of retained component loadings into factor loadings implies that the RFT-loadings and the factor loadings are identical.

Whereas Model 1 represents an orthogonal perfect simple structure with large salient loadings Model 2 represents an orthogonal perfect simple structure with moderate salient loadings. Model 3 represents an oblique perfect simple structure with large salient loadings and Model 4 represents an oblique perfect simple structure with moderate salient loadings (see Table 1).

Table 1: Population loadings for models with identical FA- and RFT-loadings

Model 1			Model 2		
.700	.000	.000	.500	.000	.000
.700	.000	.000	.500	.000	.000
.700	.000	.000	.500	.000	.000
.700	.000	.000	.500	.000	.000
.700	.000	.000	.500	.000	.000
.000	.700	.000	.000	.500	.000
.000	.700	.000	.000	.500	.000
.000	.700	.000	.000	.500	.000
.000	.700	.000	.000	.500	.000
.000	.700	.000	.000	.500	.000
.000	.000	.700	.000	.000	.500
.000	.000	.700	.000	.000	.500
.000	.000	.700	.000	.000	.500
.000	.000	.700	.000	.000	.500
.000	.000	.700	.000	.000	.500
Model 3			**Model 4**		
.714	.000	.000	.520	.000	.000
.714	.000	.000	.520	.000	.000
.714	.000	.000	.520	.000	.000
.714	.000	.000	.520	.000	.000
.714	.000	.000	.520	.000	.000
.000	.665	.000	.000	.472	.000
.000	.665	.000	.000	.472	.000
.000	.665	.000	.000	.472	.000
.000	.665	.000	.000	.472	.000
.000	.665	.000	.000	.472	.000
.000	.000	.616	.000	.000	.424
.000	.000	.616	.000	.000	.424
.000	.000	.616	.000	.000	.424
.000	.000	.616	.000	.000	.424
.000	.000	.616	.000	.000	.424
inter-factor correlations					
1.000			1.000	-.061	.363
-.032	1.000		-.061	1.000	-.475
.273	-.329	1.000	.363	-.475	1.000

Population FA-loadings and the corresponding RFT-loadings for models with unequal FA- and RFT-loadings ($\boldsymbol{\Lambda} \neq \boldsymbol{\Lambda}^*$) are given in Table 2 and 3. Models 5 and 6 are orthogonal, Model 5 has a simple structure with large salient loadings, and Model 6 has a simple structure with moderate salient loadings (see Table 2). Moreover, Model 7 has an oblique simple structure and high salient loadings whereas Model 8 represents an oblique simple structure with low to moderate salient loadings (see Table 3). The eight models with their corresponding population factor loading matrices presented in Tables 1, 2, and 3 were used in order to generate population correlation matrices according to Equation 5. It should be noted that even for those population models where the FA- and RFT-loadings were not equal, the means of the FA- and the RFT-loadings were generally similar (see Tables 2 and 3, bottom). The only exception was found for the first factor of Model 7, where the mean RFT-loading was a bit smaller than the mean factor loading. Overall, this demonstrates that the RFT-loadings are not inflated.

From each population 500 random normal samples with 50, 75, 150, 300, and 1000 cases were taken. Maximum likelihood factor analysis (MLFA), unweighted least squares factor analysis (ULFA), and PCA were performed for each sample correlation matrix. It should be noted that the relative size of the MLFA-loadings does not depend on the standard deviations of the observed variables, which means that MLFA is scale free (Lawley, 1940). On the other hand, PCA is not scale free so that the relative size of the PCA-loadings can be affected by different standard deviations of the observed variables. Since RFT is based on PCA, it is not recommended to calculate RFT for ML-factors when covariance matrices are analyzed. Therefore, the present simulation study was based on correlation matrices so that no effects of scaling on the loadings were expected. It was decided to include a correlation-based MLFA into the simulation study, because ML-estimation is rather common in the context of factor analysis.

Table 2: Popluation loading matrices of Model 5 and 6

	FA-loadings			RFT-loadings		
Model 5						
	.600	.000	.000	.638	.000	.000
	.650	.000	.000	.673	.000	.000
	.700	.000	.000	.705	.000	.000
	.750	.000	.000	.734	.000	.000
	.800	.000	.000	.759	.000	.000
	.000	.600	.000	.000	.638	.000
	.000	.650	.000	.000	.673	.000
	.000	.700	.000	.000	.705	.000
	.000	.750	.000	.000	.734	.000
	.000	.800	.000	.000	.759	.000
	.000	.000	.600	.000	.000	.638
	.000	.000	.650	.000	.000	.673
	.000	.000	.700	.000	.000	.705
	.000	.000	.750	.000	.000	.734
	.000	.000	.800	.000	.000	.759
M	**.233**	**.233**	**.233**	**.234**	**.234**	**.234**
Model 6						
	.400	.000	.000	.437	.000	.000
	.450	.000	.000	.474	.000	.000
	.500	.000	.000	.507	.000	.000
	.550	.000	.000	.535	.000	.000
	.600	.000	.000	.559	.000	.000
	.000	.400	.000	.000	.437	.000
	.000	.450	.000	.000	.474	.000
	.000	.500	.000	.000	.507	.000
	.000	.550	.000	.000	.535	.000
	.000	.600	.000	.000	.559	.000
	.000	.000	.400	.000	.000	.437
	.000	.000	.450	.000	.000	.474
	.000	.000	.500	.000	.000	.507
	.000	.000	.550	.000	.000	.535
	.000	.000	.600	.000	.000	.559
M	**.167**	**.167**	**.167**	**.167**	**.167**	**.167**

Note. "M" denotes the column mean.

Table 3: Popluation loading matrices of Model 7 and 8

	FA-loadings			RFT-loadings		
			Model 7			
	.673	.000	.000	.654	.001	.005
	.694	.000	.000	.666	.000	.002
	.714	.000	.000	.677	.000	.000
	.734	.000	.000	.688	.001	.002
	.755	.000	.000	.698	.001	.005
	.000	.624	.000	.001	.636	.006
	.000	.644	.000	.000	.648	.003
	.000	.665	.000	.000	.659	.000
	.000	.686	.000	.001	.671	.003
	.000	.706	.000	.001	.681	.006
	.000	.000	.573	.005	.006	.586
	.000	.000	.594	.003	.003	.597
	.000	.000	.616	.000	.000	.609
	.000	.000	.638	.003	.004	.620
	.000	.000	.659	.006	.007	.629
M	**.238**	**.222**	**.205**	**.227**	**.221**	**.205**
			inter-factor correlations			
	1.000			1.000		
	-.031	1.000		-.035	1.000	
	.271	-.327	1.000	.265	-.340	1.000
			Model 8			
	.478	.000	.000	.496	.002	.007
	.498	.000	.000	.508	.001	.003
	.520	.000	.000	.521	.000	.000
	.540	.000	.000	.531	.001	.004
	.561	.000	.000	.541	.002	.007
	.000	.428	.000	.002	.446	.010
	.000	.450	.000	.001	.458	.005
	.000	.471	.000	.000	.468	.001
	.000	.492	.000	.002	.478	.006
	.000	.514	.000	.003	.487	.011
	.000	.000	.376	.009	.012	.397
	.000	.000	.399	.004	.006	.407
	.000	.000	.423	.001	.001	.417
	.000	.000	.446	.006	.008	.425
	.000	.000	.470	.011	.015	.433
M	**.173**	**.157**	**.141**	**.176**	**.159**	**.142**
			inter-factor correlations			
	1.000			1.000		
	-.060	1.000		-.067	1.000	
	.359	-.468	1.000	.340	-.505	1.000

Note. "M" denotes the column mean.

Varimax-rotation was performed for the orthogonal models (Model 1, 2, 5, and 6) and Promax-rotation (Kappa=4) was performed for the oblique models (Model 3, 4, 7, and 8). Then, according to Equation 8 the RFT-loadings were computed from the unrotated sample PCA retained component loadings and the factor loadings. Varimax-rotation of the RFT-loadings was performed for the orthogonal models and Promax-rotation (Kappa=4) was performed for the oblique models.

Although the RFT constitutes a new model comprising aspects both from PCA and FA, researchers might want to use the RFT especially as a substitute for FA. Therefore, the root mean square (RMS) difference between the sample FA-loadings and the corresponding population FA-loadings was compared with the RMS difference between the sample RFT-loadings and the corresponding population FA-loadings (Figures 1 and 2). The RMS difference represents the overall difference between sample and population FA-loadings, but it does not indicate whether an over- or underestimation occurs. Therefore, the mean-difference between the mean sample loadings and the population FA-loadings was also calculated (see Table 4). The mean-difference is negative when the sample RFT-, FA-, or PCA-loadings underestimate the population FA-loadings and it is positive when the sample loadings overestimate the population FA-loadings.

Figure 1 contains the RMS differences between the sample MLFA-loadings, sample ULFA-loadings, sample ML-RFT-loadings, sample UL-RFT-loadings, sample PCA-loadings and the corresponding population FA-loadings for Models 1 to 4. RMS differences were equal or smaller for RFT-loadings based on ML-estimation than for MLFA-loadings. Moreover, RMS differences were equal or smaller for RFT-loadings based on UL- estimation than for ULFA-loadings. Thus, when the population FA-loadings and the population RFT-loadings are equal, the precision of the sample RFT-loadings as estimates of the population FA-loadings is at least as high as the precision of the FA-loadings. The mean-differences between sample MLFA-loadings, sample ULFA-loadings, sample RFT-loadings and the corresponding population FA-loadings were extremely small for Models 1 to 3 (see Table 4). They were a bit larger for Model 4, where a slight tendency for an underestimation of loadings was found for all methods. PCA-loadings have, in general, the largest RMS and, thus, the lowest precision as estimates of the FA-loadings, especially for sample sizes of 150 cases and above (see Figure 1) and the mean-differences were of a relevant size (see Table 4), indicating the known tendency of PCA-loadings to overestimate the population FA-loadings.

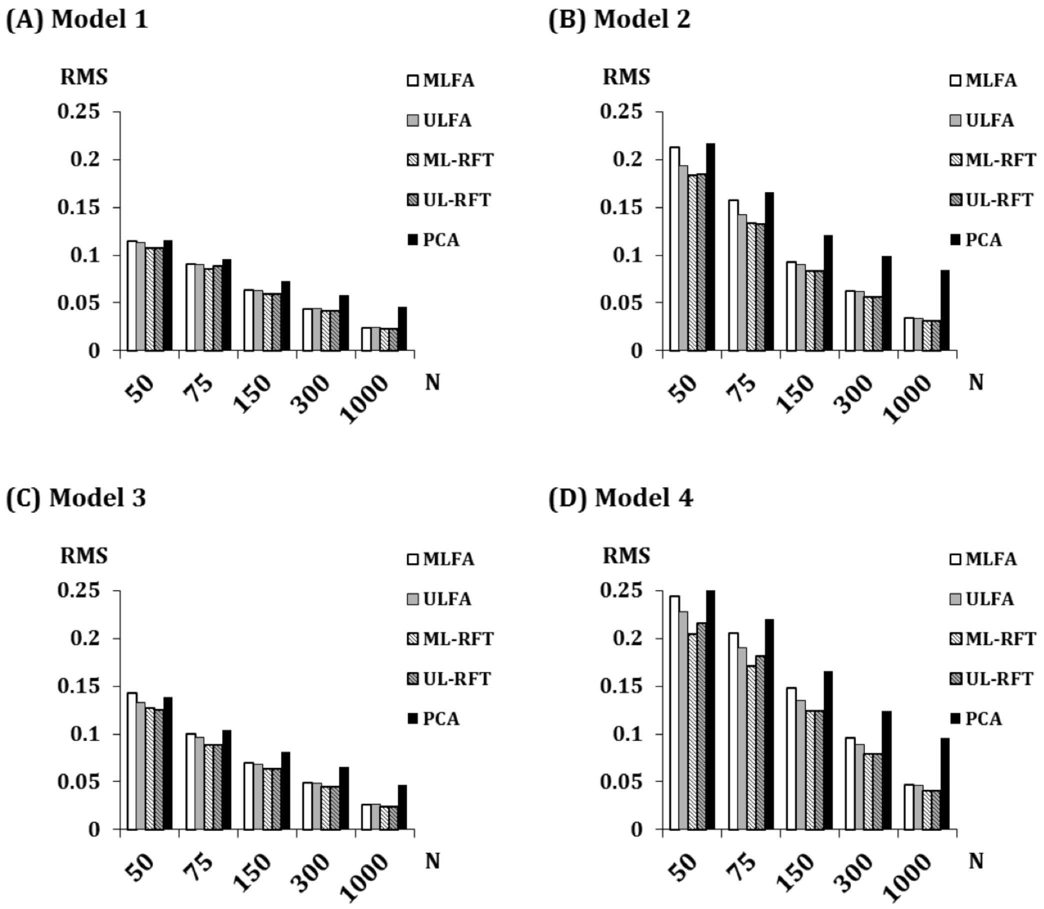

Figure 1. Root mean squared difference (RMS) between the sample MLFA-, ULFA-, ML-RFT-, UL-RFT-, PCA- loadings and the corresponding population factor loadings for Models 1 to 4.

Table 4. Mean-difference between sample loading estimates and population FA-loadings

Model	N	MLFA	ML-RFT	ULFA	UL-RFT	PCA
1	50	-.003	-.002	-.002	-.001	.019
	75	-.002	-.001	-.001	-.001	.021
	150	-.001	.000	-.001	.000	.022
	300	.000	.000	.000	.000	.023
	1000	.000	.000	.000	.000	.023
2	50	-.002	-.008	.000	.001	.032
	75	-.002	-.001	.001	.001	.036
	150	.000	.001	.001	.001	.041
	300	.000	.001	.000	.001	.043
	1000	.000	.000	.000	.000	.044
3	50	-.005	-.002	-.003	-.001	.021
	75	-.003	-.001	-.002	-.001	.023
	150	-.001	.000	-.001	.000	.025
	300	.000	.000	.000	.000	.026
	1000	.000	.000	.000	.000	.026
4	50	-.020	-.045	-.016	-.014	.013
	75	-.017	-.024	-.015	-.012	.020
	150	-.012	-.007	-.006	-.006	.032
	300	-.006	-.003	-.002	-.002	.041
	1000	-.002	-.002	-.001	-.001	.046
5	50	-.002	-.002	-.002	-.001	.019
	75	-.002	-.001	-.001	-.001	.020
	150	-.001	-.001	-.001	.000	.021
	300	.000	.000	.000	.000	.022
	1000	.000	.001	.000	.001	.023
6	50	-.001	-.018	.001	.002	.033
	75	.000	.001	.001	.002	.036
	150	.000	.001	.000	.002	.040
	300	.000	.001	.000	.001	.042
	1000	.000	.001	.000	.001	.043
7	50	-.021	-.009	-.010	-.008	.015
	75	-.012	-.009	-.010	-.008	.015
	150	-.009	-.007	-.008	-.007	.018
	300	-.006	-.006	-.006	-.006	.021
	1000	-.006	-.006	-.006	-.006	.021
8	50	-.017	-.044	-.015	-.014	.013
	75	-.018	-.017	-.016	-.015	.017
	150	-.013	-.009	-.011	-.008	.031
	300	-.007	-.004	-.006	-.003	.040
	1000	-.003	-.002	-.004	-.002	.046
M		**-.005**	**-.006**	**-.004**	**-.003**	**.028**

Note. "M" denotes the column mean.

The RMS differences between population FA-loadings and corresponding sample FA-loadings, sample RFT-loadings, and sample PCA-loadings were presented for Models 5 to 8 in Figure 2. Both for ML- and UL-estimation, the RMS differences were smaller for the RFT-loadings than for the FA-loadings. Although the population RFT-loadings were different from the population FA-loadings for Models 5 to 8, the sample RFT-loadings were at least as precise estimators of the population FA-loadings as the sample FA-loadings. The mean-differences between sample and population loadings were extremely small for MLFA-, ULFA-, ML-RFT-, and UL-RFT-loadings in Models 5 and 6. They tend to be a bit more negative for Model 7 and especially for Model 8 for samples comprising 50 and 75 cases (see Table 4). The overall mean-difference between UL-based RFT-loadings and population factor loadings was slightly smaller than the overall mean-difference for any other method (see Table 4, bottom). Again, the PCA-loadings had the lowest precision as estimates of the population FA-loadings both in terms of RMS (Figure 2) and in terms of the mean-differences, which indicate the overestimation of population FA-loadings by means of PCA (Table 4).

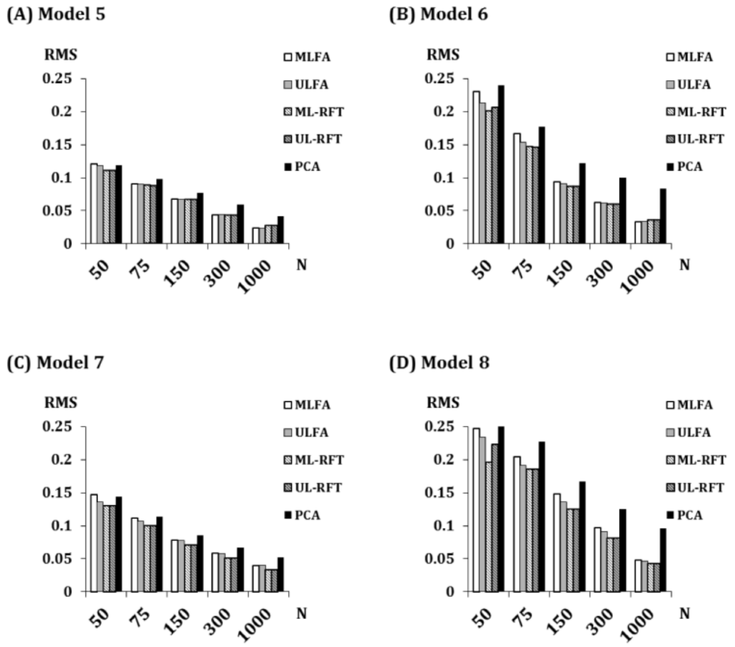

Figure 2. Root mean squared difference (RMS) between the sample MLFA-, ULFA-, ML-RFT-, UL-RFT-, PCA- loadings and the corresponding population factor loadings for Models 5 to 8.

The RMS differences between the inter-correlations of the population factors and the sample inter-correlations for MLFA, ULFA, the corresponding RFT, and the sample principal components were presented for the oblique population models (Model 3, 4, 7, and 8, see Figure 3). For Models 4, 7, and 8 and sample sizes below 150 cases the RMS differences were smaller for MLFA than for the RFT based on ML-estimation. Especially, when based on 50 cases, the RMS was large for the ML-based RFT for Models 4 and 8. However, this effect did not occur for the UL-based RFT. In contrast, when sample size was at least 150 cases the RMS was smaller for the ML-based RFT than for MLFA. For UL-based RFT the RMS tends to be equal or smaller than for ULFA. Overall, the mean-differences between sample inter-correlations and population factor inter-correlations indicate that the correlations tend to be underestimated (see Table 5). The effect of underestimation was most pronounced for PCA. Moreover, the underestimation of inter-factor correlations was less pronounced for RFT-solutions than for the FA-solutions with all methods (see Table 5).

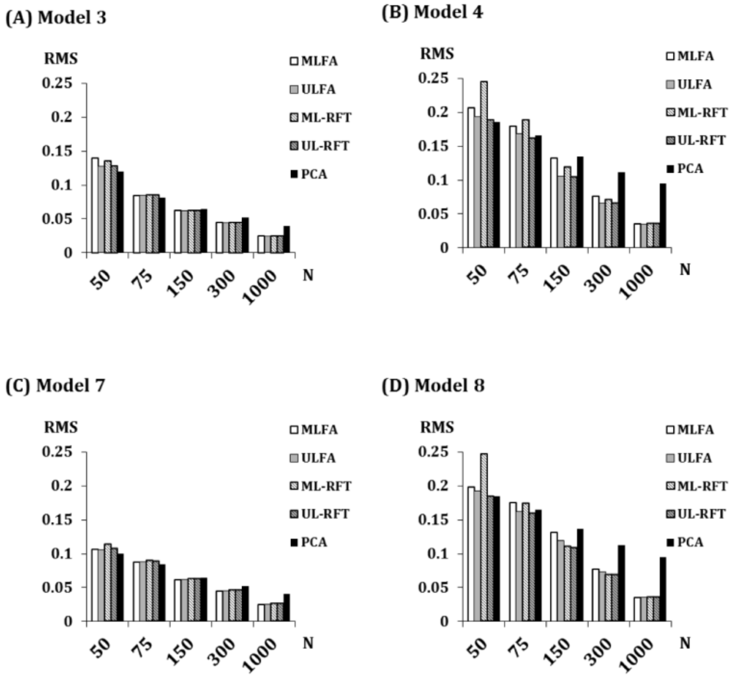

Figure 3. Root mean squared difference (RMS) between the inter-correlations for sample MLFA-, ULFA-, ML-RFT-, UL-RFT-, PCA and the corresponding population factor inter-correlations for the oblique models (Model 3, 4, 7, and 8).

Table 5. Mean-difference between sample factor and component inter-correlations and population inter-factor correlations for the oblique models (Model 3, 4, 7, and 8)

Model	N	MLFA	ML-RFT	ULFA	UL-RFT	PCA
	50	-0.054	-0.039	-0.045	-0.041	-0.073
	75	-0.023	-0.017	-0.022	-0.019	-0.057
3	150	-0.014	-0.011	-0.013	-0.011	-0.052
	300	-0.007	-0.006	-0.007	-0.006	-0.049
	1000	-0.003	-0.003	-0.003	-0.003	-0.047
	50	-0.226	-0.167	-0.21	-0.194	-0.232
	75	-0.189	-0.112	-0.17	-0.151	-0.207
4	150	-0.124	-0.076	-0.105	-0.082	-0.174
	300	-0.06	-0.038	-0.053	-0.041	-0.15
	1000	-0.017	-0.014	-0.017	-0.014	-0.135
	50	-0.047	-0.034	-0.041	-0.038	-0.072
	75	-0.027	-0.019	-0.025	-0.021	-0.061
7	150	-0.013	-0.01	-0.012	-0.01	-0.053
	300	-0.007	-0.005	-0.007	-0.005	-0.05
	1000	-0.002	-0.001	-0.002	-0.001	-0.047
	50	-0.211	-0.151	-0.207	-0.189	-0.228
	75	-0.183	-0.138	-0.166	-0.15	-0.208
8	150	-0.121	-0.079	-0.108	-0.087	-0.177
	300	-0.061	-0.04	-0.056	-0.042	-0.151
	1000	-0.018	-0.012	-0.017	-0.013	-0.135
M		**-0.07**	**-0.049**	**-0.064**	**-0.056**	**-0.118**

Note. "M" denotes the column mean.

Empirical Study

Since the simulation study focused on the loadings and factor inter-correlations, the empirical example presented in the following focused on the robustness of factor score predictors. A sample of 497 German participants (353 females; 71 %; age: M = 33.1; SD = 12.6) was recruited by means of newspaper advertising and through advertising in university courses. The participants indicated written informed consent and filled in 20 items (10 extraversion items, 10 neuroticism items) of the German Version of the Eysenck Personality Inventory (EPI; Eggert, 1983). Since there are more females in the sample, the data do not represent a balanced sample of the population. Nevertheless, asymmetries of demographic parameters are not rare in empirical research, so that it seemed reasonable to demonstrate RFT by means of this sample.

Two factors were extracted by means of ULFA, because two factors (Extraversion and Neuroticism) were expected to occur. The Promax-rotated ULFA-solution (Kappa=4), the corresponding Promax-rotated RFT-solution, and the Promax-rotated PCA-solution are presented in Table 6. The Neuroticism-factor is rather clear whereas the Extraversion-factor is rather weak, because four items do not load as expected. Overall, the ULFA loading pattern and the corresponding RFT loading pattern were very similar, although some of the largest ULFA loadings were a bit smaller in the RFT-solution. Moreover, inspection of Table 6 reveals the well-known overestimation of loadings that occurs with PCA.

Table 6. Pattern-loadings of Promax-solution of ULFA, UL-based RFT, and PCA for 20 items of the EPI

	ULFA		**UL-RFT**		**PCA**	
item	**N**	**E**	**N**	**E**	**N**	**E**
e01	-.02	**.48**	-.03	**.47**	-.04	**.58**
e03	**-.46**	.18	**-.46**	.18	**-.52**	.22
e05	.17	.28	.18	**.32**	.20	**.40**
e08	.16	**.37**	.16	**.40**	.17	**.50**
e10	**.38**	.18	.39	.19	**.44**	.24
e13	.00	**.51**	-.01	**.50**	-.01	**.62**
e15	-.16	**.44**	-.17	**.43**	-.20	**.53**
e17	-.05	**.57**	-.06	**.52**	-.08	**.65**
e20	.00	.03	.00	.03	.00	.04
e22	.15	.19	.16	.22	.18	.27
n02	**.47**	.11	**.47**	.11	**.53**	.14
n04	**.36**	.07	**.38**	.07	**.43**	.09
n07	**.63**	.15	**.59**	.14	**.67**	.18
n09	**.49**	-.07	**.49**	-.07	**.55**	-.09
n11	**.33**	-.16	**.35**	-.18	**.40**	-.22
n14	**.53**	-.03	**.53**	-.03	**.60**	-.04
n16	**.49**	-.08	**.49**	-.08	**.56**	-.09
n19	.24	.13	.25	.14	.29	.18
n21	**.38**	.07	**.40**	.07	**.45**	.09
n23	**.57**	-.06	**.55**	-.06	**.63**	-.07
Inter-correlations						
	-.05		-.05		-.03	
First 10 eigenvalues of unrotated PCA:						
3.34, 2.20, 1.40, 1.25, 1.09, 1.02, .96, .89, .86, .79						

In some occasions researchers want to get scores for each participant, so that Thurstone's regression score predictor was computed for the ULFA-solution and for the RFT-solution (see Equation 21). Since gender was distributed rather unequally, the robustness of factor score predictors might be questioned. In order to investigate the robustness of the score predictors, 150 random splits of the total sample into two subsamples (249 vs. 248 participants) were performed. The weights for the computation of score predictors were calculated for ULFA, for UL-based RFT, and for PCA component scores in each sub-sample. Then, the weights were applied to compute the score predictors and component score in the total sample so that the root mean squared (RMS) correlation between the scores based on the two sub-samples was computed as an indicator of the robustness of the score predictors. The RMS correlation was .94 with a standard deviation of .06 for ULFA, .97 with a standard deviation of .04 for the UL-RFT score predictors, and .95 with a standard deviation of .07 for PCA.

Conclusion

A transformation of the retained principal component loadings to be as similar as possible to the factor loading matrix was proposed. This transformation was called 'retained-components factor transformation' (RFT). It was shown that Harman's ideal variables factor score predictor based on RFT has perfect correlations with the principal components. It can therefore be concluded that Harman's factor score predictor based on RFT is an optimal summarizer of the observed variables. Moreover, Thurstone's regression score predictor based on RFT was shown to have a perfect multiple correlation with the principal components, indicating that the RFT based regression score predictor summarizes the same overall individual differences as the principal components, even when the variances are distributed differently on the RFT factors and principal components. Thus, the RFT based regression score predictor is also an optimal summarizer of the observed variables.

In a simulation study based on orthogonal and oblique simple structure the means of the population loadings were very similar for FA and RFT. This demonstrates that the RFT-loadings are not inflated as has been found for PCA-loadings when compared to FA-loadings (Widaman, 1993; Snook & Gorsuch, 1989). Moreover, the RMS difference between the population factor loadings and the sample loadings was overall equal or smaller for RFT-loadings than for FA-loadings and PCA-loadings. This implies that RFT-loadings can be used as estimates of population factor loadings. Moreover, the mean-difference between the RFT-loadings and the population factor loadings was smallest for the RFT based on UL-estimates. This indicates that the UL-based RFT-loadings might be slightly more precise than other estimates of the factor loadings. Moreover, the underestimation of inter-factor correlations was less pronounced for RFT-solutions than for the FA-solutions.

The empirical example was based on 20 items of the EPI. The simple structure of the two-factor solutions was not perfect and the sample had an unbalanced gender distribution. Thus, the sample contains imperfect data as they occur in empirical research. The Promax loading pattern of the ULFA-solution and the Promax loading pattern of the UL-based RFT-solution were very similar and would probably lead to the same interpretation of the factors whereas the PCA-loadings were again inflated. Nevertheless, many of the largest loadings in the ULFA-solution were smaller in the RFT-solution. The total sample was divided into two-subsamples and the weights for Thurstone's regression score predictor were computed in the subsamples. These weights were then applied to the total-sample in order to compute score predictors. The RMS of the correlation between the score predictors based on sub-sample weights was a bit smaller for ULFA than for the UL-based RFT. This indicates that score predictors that are based on UL-RFT could be a valuable alternative to conventional scores.

To summarize, RFT could be regarded as interesting in several applied settings because the simple structure models investigated in the present simulation study and in the empirical study are relevant for many areas of research. It was found that RFT allows for a model without inflated loadings, which can be used as estimates of population factor loadings. The underestimation of inter-factor correlations was less pronounced when based on RFT than for FA. Moreover, the RFT model implies score predictors that are optimal summarizers of the observed variables and the regression score predictor based on UL-RFT was more robust than the ULFA-based regression score predictor. In this sense, RFT combines the advantages of PCA (score predictors that are optimal summarizers of observed variables) with the advantages of FA (RFT-loadings are not inflated).

It should be noted that the computation of the RFT-loadings can be based on any initial factor model when the analyses are based on the inter-correlations of the observed variables (maximum likelihood, unweighted least-squares, principal axis factoring, etc.). Although RFT was also calculated for MLFA, it should be noted that this is only possible when the analyses are based on the inter-correlations of the observed variables. When covariances are used instead of correlations, MLFA will lead to a scale-free solution whereas PCA will depend on scaling, so that the RFT might be biased. Accordingly, when RFT is based on covariances ULFA or principal axis factoring would be an appropriate method. Moreover, the stability of the results of the simulation study indicates that the UL-based RFT-loadings should be preferred over ML-based RFT-loadings in small samples. A small R script that can be used in order to compute the RFT-loadings from an initial loading matrix is available from the authors (http://beauducel.de/research.html).

References

Beauducel, A. & Rabe, S. (2009). Model-related factor score predictors for confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 62*, 489-506.

Bentler, P. M. & De Leeuw, J. (2011). Factor analysis via components analysis. *Psychometrika*, *76*, 461-470.

Conway, J. M. & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*, 147-168.

Eggert, D. (1983). *Eysenck-Persönlichkeits-Inventar (EPI)* [Eysenck-Personality Inventory]. Göttingen: Hogrefe.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Grice, J. W. (2001). Computing and evaluation factor scores. *Psychological Methods*, *6*, 430-450.

Guttman, L. (1955). The determinacy of factor score matrices with applications for five other problems of common factor theory. *British Journal of Statistical Psychology*, *8*, 65-82.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: The University of Chicago Press.

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed.). Hillsdale, NJ: Erlbaum.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology*, *24*: 417-441.

Hurley, J. R. & Cattell, R. B. (1962). The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, *7*, 258-262.

Krijnen, W. P. (2006). Some results on mean square error for factor score prediction. *Psychometrika*, *71*, 395-409.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, *60*, 64-82.

Lovie, P. & Lovie, A. D. (1995). The cold equations: Spearman and Wilson on factor indeterminacy. *British Journal of Mathematical and Statistical Psychology*, *48*, 237-253.

Ogasawara, H. (2003). Oblique factors and components with independent clusters. *Psychometrika*, *68*, 299-321.

Preacher, K. J. & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13-43.

Sato, M. (1990). Some remarks on principal component analysis as a substitute for factor analysis in monofactor cases. *Journal of the Japanese Statistical Society*, *20*, 23-31.

Schneeweiss, H. (1997). Factors and principal components in the near spherical case. *Multivarite Behavioral Research*, *32*, 375-401.

Schneeweiss, H. & Mathes, H. (1995). Factor analysis and principal components. *Journal of Multivariate Analysis*, *55*, 105-124.

Snook, S. C. & Gorsuch, R. L. (1989). Component analysis versus common factor analysis: A monte carlo study. *Psychological Bulletin*, *106*, 148-154.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th Ed.). Boston, MA: Pearson Education.

Ten Berge, J. M. F. & Kiers, H. A. L. (1997). Are all varieties of PCA the same? A reply to Cadima & Jolliffe. *British Journal of Mathematical and Statistical Psychology*, *50*, 367-368.

Ten Berge, J. M. F. & Knol, D. L. (1985). Scale construction on the basis of components analysis: A comparison of three strategies. *Multivariate Behavioral Research*, *20*, 45-55.

Ten Berge, J. M. F., Krijnen, W. P., Wansbeek, T., & Shapiro, A. (1999). Some new results on correlation-preserving factor scores prediction methods. *Linear Algebra and its Applications*, *289*, 311–318.

Thurstone, L.L. (1935*). The Vectors of Mind*. Chicago: University of Chicago Press.

Unkel, S. & Trendafilov, N. T. (2010). A majorization algorithm for simultaneous parameter estimation in robust exploratory factor analysis. *Computational Statistics and Data Analysis*, *54*, 3348-3358.

Velicer, W. F. & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure (with comments and reply). *Multivariate Behavioral Research*, *25*, 1-114.

Widaman, K. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263-311.