A Brief Introduction to Factor Analysis

1. Introduction

Factor analysis attempts to represent a set of observed variables $X_1, X_2 \ldots X_n$ in terms of a number of 'common' factors plus a factor which is unique to each variable. The common factors (sometimes called latent variables) are hypothetical variables which explain why a number of variables are correlated with each other -- it is because they have one or more factors *in common*.

A concrete physical example may help. Say we measured the size of various parts of the body of a random sample of humans: for example, such things as height, leg, arm, finger, foot and toe lengths and head, chest, waist, arm and leg circumferences, the distance between eyes, etc. We'd expect that many of the measurements would be correlated, and we'd say that the explanation for these correlations is that there is a common underlying factor of body size. It is this kind of common factor that we are looking for with factor analysis, although in psychology the factors may be less tangible than body size.

To carry the body measurement example further, we probably wouldn't expect body size to explain all of the variability of the measurements: for example, there might be a lankiness factor, which would explain some of the variability of the circumference measures and limb lengths, and perhaps another factor for head size which would have some independence from body size (what factors emerge is very dependent on what variables are measured). Even with a number of common factors such as body size, lankiness and head size, we still wouldn't expect to account for all of the variability in the measures (or explain all of the correlations), so the factor analysis model includes a unique factor for each variable which accounts for the variability of that variable which is not due to any of the common factors.

Why carry out factor analyses? If we can summarise a multitude of measurements with a smaller number of factors without losing too much information, we have achieved some economy of description, which is one of the goals of scientific investigation. It is also possible that factor analysis will allow us to test theories involving variables which are hard to measure directly. Finally, at a more prosaic level, factor analysis can help us establish that sets of questionnaire items (observed variables) are in fact all measuring the same underlying factor (perhaps with varying reliability) and so can be combined to form a more reliable measure of that factor.

There are a number of different varieties of factor analysis: the discussion here is limited to principal axis factor analysis and factor solutions in which the common factors are uncorrelated with each other. It is also assumed that the observed variables are standardised (mean zero, standard deviation of one) and that the factor analysis is based on the correlation matrix of the observed variables.

2. The Factor Analysis Model

If the observed variables are $X_1, X_2 \ldots X_n$, the common factors are $F_1, F_2 \ldots F_m$ and the unique factors are $U_1, U_2 \ldots U_n$, the variables may be expressed as linear functions of the factors:

$$X_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_1U_1$$
$$X_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2m}F_m + a_2U_2$$
$$\ldots$$
$$X_n = a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nm}F_m + a_nU_n \qquad (1)$$

Each of these equations is a regression equation; factor analysis seeks to find the coefficients $a_{11}, a_{12} \ldots a_{nm}$ which best reproduce the observed variables from the factors. The coefficients $a_{11}, a_{12} \ldots a_{nm}$ are weights in the same way as regression coefficients (because the variables are standardised, the constant is zero, and so is not shown). For example, the coefficient a_{11} shows the effect on variable X_1 of a one-unit increase in F_1. In factor analysis, the coefficients are called loadings (a variable is said to 'load' on a factor) and, when the factors are uncorrelated, they also show the correlation between each variable and a given factor. In the model above, a_{11} is the loading for variable X_1 on F_1, a_{23} is the loading for variable X_2 on F_3, etc.

When the coefficients are correlations, i.e., when the factors are uncorrelated, the sum of the squares of the loadings for **variable X_1**, namely $a_{11}^2 + a_{12}^2 + \ldots + a_{13}^2$, shows the proportion of the variance of variable X_1 which is accounted for by the common factors. This is called the *communality*. The larger the communality for each variable, the more successful a factor analysis solution is.

By the same token, the sum of the squares of the coefficients for **a factor** -- for F_1 it would be $[a_{11}^2 + a_{21}^2 + \ldots + a_{n1}^2]$ -- shows the proportion of the variance of all the variables which is accounted for by that factor.

3. The Model for Individual Subjects

Equation (1) above, for variable 2, say, may be written explicitly for one subject *i* as

$$X_{2i} = a_{21}F_{1i} + a_{22}F_{2i} + a_{23}F_{3i} + \ldots + a_{2m}F_{mi} + a_2U_{2i} \qquad (2)$$

This form of the equation makes it clear that there is a value of each factor for each of the subjects in the sample; for example, F_{2i} represents subject *i*'s score on Factor 2. Factor scores are often used in analyses in order to reduce the number of variables which must be dealt with. However, the coefficients $a_{11}, a_{21}, \ldots a_{nm}$ are the same for all subjects, and it is these coefficients which are estimated in the factor analysis.

4. Extracting Factors and the Rotation of Factors

The mathematical process used to obtain a factor solution from a correlation matrix is such that each successive factor, each of which is uncorrelated with the other factors, accounts for as much of the variance of the observed variables as possible. (The amount of variance accounted for by each factor is shown by a quantity called the *eigenvalue*, which is equal to the sum of the squared loadings for a given factor, as will be discussed below). This often means that all the variables have substantial loadings on the first factor; i.e., that coefficients a_{11}, $a_{21, \ldots .} a_{nm}$ are all greater than some arbitrary value such as .3 or .4. While this initial solution is consistent with the aim of accounting for as much as possible of the total variance of the observed variables with as few factors as possible, the initial pattern is often adjusted so that each individual variable has substantial loadings on as few factors as possible (preferably only one). This adjustment is called *rotation to simple structure*, and seeks to provide a more interpretable outcome. As will be seen in the example which we'll work through later, rotation to simple structure can be seen graphically as the moving or rotation of the axes (using the term 'axis' in the same way as it is used in '*x*-axis' and '*y*-axis') which represent the factors.

5. Estimating Factor Scores

Given the equations (1) above, which show the variables $X_{1 \ldots} X_n$ in terms of the factors F_1 $\ldots F_m$, it should be possible to solve the equations for the factor scores, so as to obtain a score on each factor for each subject. In other words, equations of the form

$F_1 = b_{11}X_1 + b_{12}X_2 \ldots b_{1n}X_n$
$F_2 = b_{21}X_1 + b_{22}X_2 \ldots b_{2n}X_n$

….

$F_m = b_{m1}X_1 + b_{m2}X_2 \ldots b_{mn}X_n$ (3)

should be available; however, problems are caused by the unique factors, because when they are included with the common factors, there are more factors than variables, and no exact solution for the factors is available. [An aside: The indeterminacy of the factor scores is one of the reasons why some researchers prefer a variety of factor analysis called principal component analysis (PCA). The PCA model doesn't contain unique factors: all the variance of the observed variables is assumed to be attributable to the common factors, so that the communality for each variable is one. As a consequence, an exact solution for the factors is available in PCA, as in equations (3) above. The drawback of PCA is that its model is unrealistic; also, some studies have suggested that factor analysis is better than PCA at recovering known underlying factor structures from observed (Snook & Gorsuch, 1989; Widaman, 1993)].

Because of the lack of an exact solution for factor scores, various approximations have been offered, and three of these are available in SPSS. One of these approximations will be discussed in considering the example with real data. It is worth noting, however, that many researchers take matters into their own hands and use the coefficients $a_{11}, a_{12} \ldots a_{nm}$ from the factor solution as a basis for creating their own factor scores. Grice (2001), who describes

and evaluates various methods of calculating factor scores, calls this method a 'coarse' one, and compares it unfavourably with some 'refined' methods and another 'coarse' one based on factor score coefficients.

6. Calculating Correlations from Factors

It was mentioned above that an aim of factor analysis is to 'explain' correlations among observed variables in terms of a relatively small number of factors. One way of gauging the success of a factor solution is to attempt to reproduce the original correlation matrix by using the loadings on the common factors and seeing how large a discrepancy there is between the original and reproduced correlations -- the greater the discrepancy, the less successful the factor solution has been in preserving the information in the original correlation matrix. How are the correlations derived from the factor solution? When the factors are uncorrelated, the process is simple. The correlation between variables X_1 and X_2 is obtained by summing the products of the coefficients for the two variables across all common factors; for a three-factor solution, the quantity would be $(a_{11} \times a_{21}) + (a_{12} \times a_{22}) + (a_{13} \times a_{23})$. This processs will become clearer in the description of the hypothetical two-factor solution based on five observed variables in the next section.

7. A Hypothetical Solution

Variable	Loadings/Correlations		Communality		Reproduced correlations						
	Factor 1	Factor 2				X_1	X_2	X_3	X_4	X_5	
X_1	.7	.2	.53		X_1	.					
X_2	.8	.3	.73		X_2	.62	.				
X_3	.9	.4	.97		X_3	.71	.84	.			
X_4	.2	.6	.40		X_4	.26	.34	.42	.		
X_5	.3	.7	.58		X_5	.35	.45	.55	.48	.	
	Σx^2 2.07	Σx^2 1.14									

7.1 *The Coefficients*

According to the above solution,

$X_1 = .7F_1 + .2F_2$
$X_2 = .8F_1 + .3F_2$
…..
$X_5 = .3F_1 + .7F_2$.

As well as being weights, the coefficients above show that the correlation between X_1 and F_1 is .70, that between X_1 and F_2 is .20, and so on.

7.2 *Variance Accounted For*

The quantities at the bottom of each factor column are the sums of the squared loadings for that factor, and show how much of the total variance of the observed variables is accounted

for by that factor. For Factor 1, the quantity is $.7^2 + .8^2 + .9^2 + .2^2 + .3^2 = 2.07$. Because in the factor analyses discussed here the total amount of variance is equal to the number of observed variables (the variables are standardised, so each has a variance of one), the total variation here is five, so that Factor 1 accounts for (2.07/5) x 100 = 41.4% of the variance.

The quantities in the *communality* column show the proportion of the variance of each variable accounted for by the common factors. For X_1 this quantity is $.7^2 + .2^2 = .53$, for X_2 it is $.8^2 + .3^2 = .73$, and so on.

7.3 ***Reproducing the Correlations***

The correlation between variables X_1 and X_2 as derived from the factor solution is equal to (.7 x .8) + (.2 x .3) = .62, while the correlation between variables X_3 and X_5 is equal to (.9 x .3) + (.4 x .7) = .55. These values are shown in the right-hand side of the above table.

8. How Many Factors?

A factor solution with as many factors as variables would score highly on the amount of variance accounted for and the accurate reproduction of correlations, but would fail on economy of description, parsimony and explanatory power. The decision about the number of common factors to retain, or to use in rotation to simple structure, must steer between the extremes of losing too much information about the original variables on one hand, and being left with too many factors on the other. Various criteria have been suggested. The standard one (but not necessarily the best) is to keep all the factors which have eigenvalues greater than one in the original solution, and that is used in the example based on *workmot.sav*, which is referred to in Exercise 2 in *An Introduction to SPSS for Windows*. This example is described in greater detail in the next section.

9. An Example with Real Data

This example is based on variables *d6a* to *d6h* of *workmot.sav*. The items corresponding to the variables are given in Appendix 1 of *An Introduction to SPSS for Windows versions 9 and 10*. The root question is "How important are the following factors to getting ahead in your organisation?" Respondents rate the eight items, such as "Hard work and effort", "Good luck", "Natural ability" and "Who you know" on a six-point scale. The scale ranges from "Not important at all" (coded 1) to "Extremely important" (coded 6).

9.1 ***The correlation matrix***

Correlations

	D6A Hard work&effort	D6B Good luck	D6C Natural ability	D6D Who you know	D6E Good performance	D6F Office politics	D6G Adapatability	D6H Years of service
D6A Hard work&effort	1.000	-.320	.658	-.452	.821	-.302	.604	.178
D6B Good luck	-.320	1.000	-.062	.426	-.316	.355	-.078	.114
D6C Natural ability	.658	-.062	1.000	-.313	.709	-.220	.597	.314
D6D Who you know	-.452	.426	-.313	1.000	-.453	.597	-.243	.174
D6E Good performance	.821	-.316	.709	-.453	1.000	-.381	.679	.160
D6F Office politics	-.302	.355	-.220	.597	-.381	1.000	-.082	.200
D6G Adapatability	.604	-.078	.597	-.243	.679	-.082	1.000	.267
D6H Years of service	.178	.114	.314	.174	.160	.200	.267	1.000

Inspection of the correlation matrix suggests that certain kinds of items go together. For example, responses to "Hard work and effort", "Natural ability", "Good performance" and "Adaptability" are moderately correlated with each other and less so with other items.

9.2 *Carrying out the Factor Analysis*

Instructions for carrying out the factor analysis are given on pages 22 to 26 of *An Introduction to SPSS for Windows*. Additional commands are described in the following sub-sections.

9.3 *The Initial Factor Analysis Solution*

The first table shows the initial and final communalities for each factor. The final estimate of

Communalities

	Initial	Extraction
D6A Hard work&effort	.708	.740
D6B Good luck	.265	.262
D6C Natural ability	.581	.638
D6D Who you know	.477	.649
D6E Good performance	.782	.872
D6F Office politics	.421	.524
D6G Adapatability	.523	.577
D6H Years of service	.216	.275

Extraction Method: Principal Axis Factoring.

the communality which is given in the second column of the table, is arrived at by an iterative process. To start the ball rolling, an initial estimate is used. By default, this is the squared multiple correlation obtained when each variable is regressed on all the other variables. In other words, the amount of the variance of variable X_1 explained by all the other variables is taken as a reasonable first estimate of the amount of X_1's variance accounted for by the common factors.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.634	45.427	45.427	3.317	41.457	41.457
2	1.748	21.847	67.275	1.220	15.253	56.710
3	.766	9.569	76.844			
4	.615	7.687	84.530			
5	.415	5.188	89.719			
6	.364	4.553	94.271			
7	.306	3.824	98.096			
8	.152	1.904	100.000			

Extraction Method: Principal Axis Factoring.

The second table shows the eigenvalues and the amount of variance explained by each successive factor. The *Initial Eigenvalues* are for a principal components analysis, in which the communalities are one. The final communalities are estimated by iteration for the principal axis factor analysis, as mentioned earlier. As can be seen from the first table, they

are somewhat less than one, and the amount of variance accounted for is reduced, as can be seen in the second table in the section headed *Extraction Sums of Squared Loadings*. The rest of the factor analysis is based on two factors, because two factors have eigenvalues greater than one. There are other methods which can be used to decide on the number of factors, some of which may generally be more satisfactory than the rule used here (Fabriger *et al*, 1999). As an aside, it has been suggested that over-extraction (retaining more than the true number of factors) leads to less distorted results than under-extraction (retaining too few factors); Wood, Tataryn & Gorsuch, 1996.

Factor Matrix[a]

	Factor	
	1	2
D6A Hard work&effort	.857	.078
D6B Good luck	-.347	.376
D6C Natural ability	.749	.276
D6D Who you know	-.590	.548
D6E Good performance	.927	.110
D6F Office politics	-.465	.555
D6G Adapatability	.664	.368
D6H Years of service	.187	.490

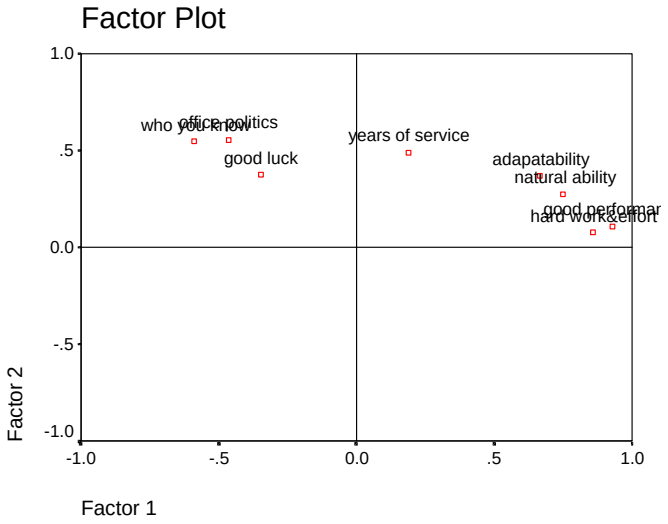
Extraction Method: Principal Axis Factoring.

a. 2 factors extracted. 9 iterations required.

The table headed *Factor Matrix* shows the coefficients $a_{11}, a_{12} \ldots a_{nm}$ for the factor analysis model (1); for example, the results show that variable *d6a* = .857 x F_1 + .078 x F_2, *d6b* = -.347 x F_1 + .376 x F_2, and so on.

As pointed out earlier, the sum of the squared loadings over factors for a given variable shows the communality for that variable, which is the proportion of the variance of the variable explained by the common factors; for example, the communality for *d6a* is $.857^2 + .078^2$, which is equal to .740, the second figure shown for that variable in the *Communalities* table above. The sum of the squared loadings for a given factor shows the variance accounted for by that factor. The figure for Factor 1, for example, is equal to $.857^2 + -.347^2 + .749^2 + \ldots + .187^2 = 3.32$, which is the figure given for the first factor in the second part of the *Total Variance Explained* table above. Finally, we may reproduce the correlations from the factor-solution coefficients. The correlation between *D6A* and *D6B*, for example, is equal to (.857 x -.347) + (.078 - .376) = -.268, a not unreasonable but not startlingly accurate estimate of the -.32 shown in the table of observed correlations.

9.4 *Plot of the Factor Loadings*



When there are two factors, the coefficients shown in the *Factor Matrix* table may be plotted on a two-dimensional graph, as above (choose *Unrotated factor solution* in the *Display* panel of the *Factor Analysis*➔*Extraction* display). This graph shows the phenomenon mentioned earlier, whereby most variables tend to have substantial loadings on the first factor.

9.5 *Rotated Factor Solution*

Rotated Factor Matrix[a]

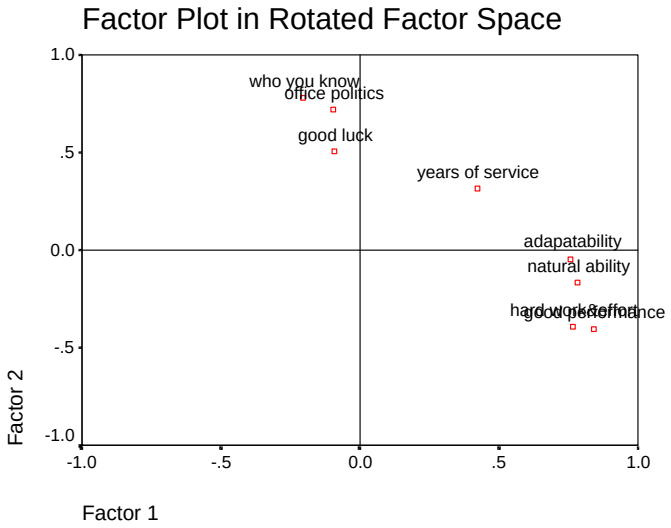
	Factor	
	1	2
D6A Hard work&effort	.765	-.394
D6B Good luck	-.092	.503
D6C Natural ability	.780	-.169
D6D Who you know	-.204	.779
D6E Good performance	.842	-.405
D6F Office politics	-.094	.718
D6G Adapatability	.758	-.046
D6H Years of service	.421	.313

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

The above table shows the factor loading/correlations for a rotated factor solution. Comparing the graphs for the rotated and unrotated solutions, it can be seen that the proximity of the points representing the variables to the axes (and the frame) have changed. This change was brought about by rotating the whole frame, including the axes, in a counter-clockwise direction. In this case the Varimax method was used; for each variable, this seeks to maximise the loading on one factor and to minimise the loadings on other factors. As discussed in *An Introduction to SPSS for Windows*, a clear pattern now emerges, with items having to do with performance and ability loading highest on Factor 1 and those to do with

who you know and good luck loading highest on the Factor 2. Years of service has lowish loadings on both factors.

9.6 ***Plot of the Rotated Factor Loadings***



The plot shows a clear pattern of the loadings, with all items identified with either Factor 1 or Factor 2, with the exception of *Years of Service*, which is poised between the two. This result illustrates the benefits of rotation.

9.7 ***Factor Scores***

In *An Introduction to SPSS for Windows*, a 'coarse' method (Grice, 2001) was used to create factor scores. We simply decided which variables were identified with each factor, and averaged the ratings on the selected items to creat a score for each subject on each of the two factors. *Years of service* was omitted because it was not clearly identified with one factor.

This method of creating factor scores is equivalent to using the equations

$$F_1 = b_{11}X_1 + b_{12}X_2 + b_{13}X_3 + b_{14}X_4 + b_{15}X_5$$
$$F_2 = b_{21}X_1 + b_{22}X_2 + b_{23}X_3 + b_{24}X_4 + b_{25}X_5$$

and assigning the value 1 to the *b*-coefficients for variables with loadings/correlations greater then .4 and zero to the *b*-coefficients for variables with loadings/correlations equal to or less than .4. Grice (2001) critically evaluates this method. One obvious problem is that factor scores created in this way may be highyl correlated even when the factors on which they are based are uncorrelated (orthogonal).

SPSS provides three more sophisticated methods for calculating factor scores. In the

Factor Score Coefficient Matrix

	Factor	
	1	2
D6A Hard work&effort	.184	-.091
D6B Good luck	.052	.120
D6C Natural ability	.230	.118
D6D Who you know	.094	.472
D6E Good performance	.473	-.186
D6F Office politics	.120	.312
D6G Adapatability	.197	.163
D6H Years of service	.118	.135

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

Factor Scores Method: Regression.

present example, the default *Regression* method was used. The first table above,

Factor Score Covariance Matrix

Factor	1	2
1	.882	-.090
2	-.090	.778

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

Factor Scores Method: Regression.

the *Factor Score Coefficient Matrix*, shows the coefficients used to calculate the factor scores; these coefficients are such that if the observed variables are in standardised form, the factor scores will also have a mean of zero and a standard deviation of one. Using the coefficients for Factor 1, the factor scores are equal to .184 x z*D6A* + .052 x z*D6B* + … + .118 x z*D6H*, where z*D6A*, etc are the standardised forms of the original observed variables. The second table above, *Factor Score Covariance Matrix* shows that although theoretically the factor scores should be entirely uncorrelated, the covariance is not zero, which is a consequence of the scores being estimated rather than calculated exactly.

10. Conclusion

This has been a very brief introduction to factor analysis, and only some of the many decisions which have to be made by users of the method have been mentioned. Fabrigar *et al* (1999) give a thorough review. Some the decisions they refer to are:

- The number of subjects (see also MacCallum, et. al., 1999)
- The type of factor analysis (e.g., principal axis factoring, principal components analysis, maximum likelihood)
- The method used to decide on the number factors (e.g., eigenvlaues greater than unity, scree plot, parallel analysis)
- The method and type of rotation

Grice (2001) adds another decision:

- The method used to calculate factor scores

These two articles are a good place to start when considering the use of factor analysis, or when designing a study in which factor analysis will be used.

Alan Taylor
Department of Psychology
5th June 2001
Modifications and additions 4th May 2002
Slight changes 27th May 2003
Additions 20th March 2004

11. References and Further Reading

Cureton, Edward. (1983). *Factor analysis: an applied approach.* Hillsdale, N.J.: Erlbaum. (QA278.5.C87)

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

Grice, J. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430-450. [Reserve]

Hair, J.F. (1998). *Multivariate data analysis*. New Jersey: Prentice Hall. (QA278.M85) (Chapter 3)

Harman, Harry. (1976). *Modern factor analysis* (3rd Ed., Revised). Chicago :University of Chicago Press. (QA278.5.H38/1976)

Kim, Jae-on & Charles W. Mueller. (1978). *Introduction to factor analysis : what it is and how to do it*. Beverly Hills, Calif. : Sage Publications. (HA29.Q35/VOL 13)

Kim, Jae-on & Charles W. Mueller. (1978). *Factor analysis : statistical methods and practical issues*. Beverly Hills, Calif. : Sage Publications. (HA29.Q35/VOL 14)

MacCallum, R.C. et.al. , (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.

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

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

Wood, J.M., Tataryn, D.J. & Gorsuch, R.L. (1996). The effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods*, 1, 354-365.