

Exploratory factor analysis (EFA)

EFA: An alternative medical example for the WAIS data set

For this example we use the same (real) WAIS data but we have fabricated a medical context for it, so the results for this example do not derive from real medical research. It is proposed to offer all adults between the ages of 40 and 75 years a free medical check-up accompanied by advice about maintaining good health. A Positive Health Inventory (PHI) is being developed in order to monitor the effectiveness of the free check-up programme. The PHI comprises 11 subtests, 6 of which are biomedical indicators (e.g., blood counts) of healthy function. We may suppose the subtests concern healthy function of lungs, muscular system, liver, skeletal system, kidneys, and heart. The remaining 5 subtests are functional measures of health. We may suppose these are scores on a step test, a stamina test, a stretch test, a blow test, and a urine flow test. We have a good idea of what the relationship among variables should be like in a healthy population, but it is acceptable to do an exploratory factor analysis and see what emerges. We will use CFA to analyse the same data later.

EFA: Missing values

Factor analysis is a large sample technique, and Tabachnik and Fidell (2007, p. 613) again give a rule of thumb; 'it is comforting to have at least 300 cases for a factor analysis'. Because the dataset will be large, it is almost certain that the problem of missing values will have to be addressed. The factor analysis **Options** dialog box allows you three choices for dealing with any missing values. The default is to exclude any cases with missing values (exclude cases listwise). The next option is to calculate each correlation using all data available for the two variables in that correlation (exclude cases pairwise). Finally, a missing observation can be replaced

with the mean of the available observations on that variable (replace with mean). These choices were discussed briefly in the section on Missing data in Chapter 3. We just remind you here that while the second option does make use of all available data, internal constraints on the correlation matrix may be breached if the correlations are not all calculated on the same set of cases. As always, it is worth investing a lot of effort to reduce missing data to an absolute minimum. Our example dataset for this chapter is rather small (only 128 cases) but at least there are no missing observations. The first few cases of the data are shown in Table 8.2. The full data set of 128 cases (med.factor.sav) is available on the book website.

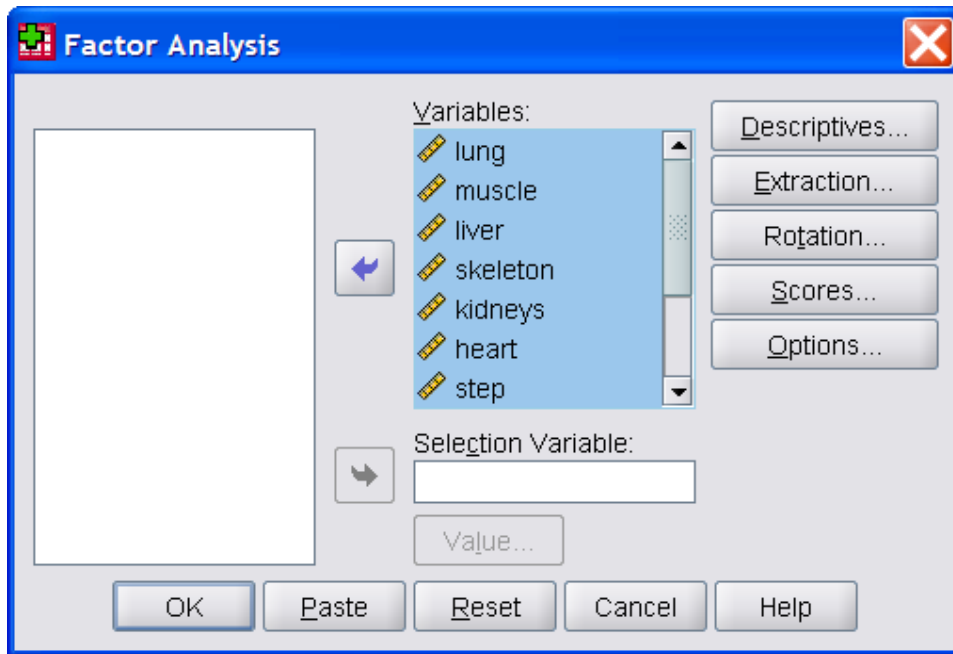
Table 8.2

The first few cases of the PHI subtest scores of a sample of 128 volunteers (the full dataset can be found as med.factor.sav on the website)

lung	muscle	liver	skeleton	kidneys	heart	step	stamina	stretch	blow	urine
20	16	52	10	24	23	19	20	23	29	67
24	16	52	7	27	16	16	15	31	33	59
19	21	57	18	22	23	16	19	42	40	61
24	21	62	12	31	25	17	17	36	36	77
29	18	62	14	26	27	15	20	33	29	88
18	19	51	15	29	23	19	20	50	37	54
19	27	61	12	19	24	17	11	38	21	72

EFA: requesting the factor analysis in SPSS

Enter the data into SPSS in 11 columns as in Table 8.2. Select **Analyze**, then **Data Reduction**, then **Factor**. Use the arrow to move all 11 subtests into the **Variables** box. Ignore the **Selection Variable** box, which you could use if you wanted to analyse a subset of cases. The main dialog box for factor analysis then looks like that in SPSS Dialog Box 8.1.

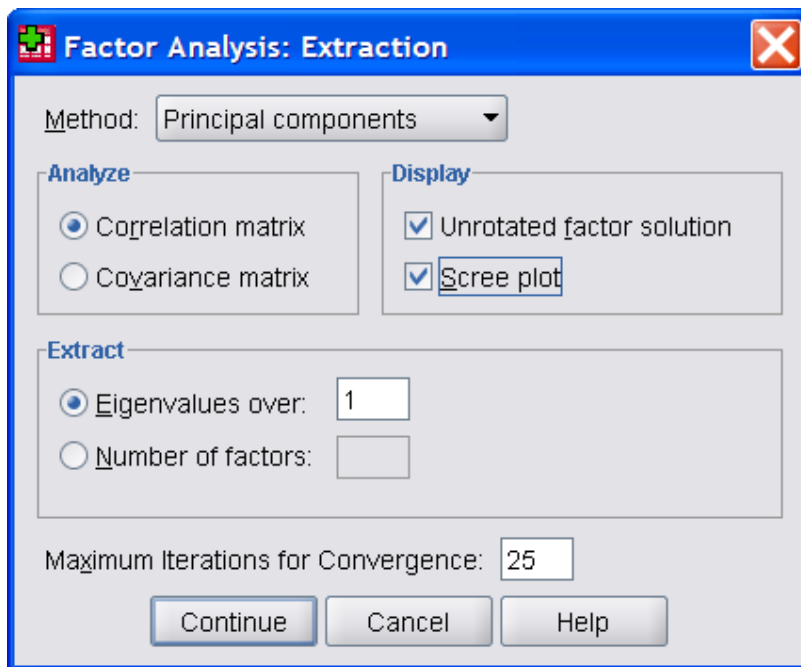


SPSS Dialog Box 8.1. The main dialog box for factor analysis

If you click on **Descriptives**, you will see that **Initial solution** in the Statistics group has a tick against it. This will give you initial eigenvalues, communalities (we'll explain what these are shortly) and the percentage of variance explained by each factor. You might like to look at the eigenvalues and proportions of variance, so leave the tick on. You could also request **Univariate descriptives**, but we don't need them so we won't bother. In the **Correlation Matrix** group are a number of diagnostic statistics. We will just select **KMO** and **Bartlett's test of sphericity** as examples. The KMO (Kaiser-Meyer-Olkin) index is a measure of sampling adequacy and the sphericity statistic tests whether the correlations among variables are too low for the factor model to be appropriate. The other diagnostics are clearly explained in Tabachnick and Fidell (2007) (see references at the end of the chapter and Further reading at the end of the book). The entries in this dialog box are straightforward and it is not reproduced here.

If you click on **Extraction** (see SPSS Dialog Box 8.2), you will see that the default for extracting factors is the **Principle components** method. Strictly speaking, the principle components method is not factor analysis, in that it uses all of the variance, including error variance, whereas factor analysis excludes the error variance by obtaining *estimates of communalities* of the variables (a communality is the proportion of variance accounted for by a factor in a factor analysis). However, principle components is the method most commonly used in the analysis of psychological data, and we will use it here. Also, as it is commonly referred to as a method of factor analysis in the psychology literature, we will follow that practice. So, we will accept the default, **Principle components** as our method of factor extraction. In the **Analyze** group, **Correlation matrix** has been selected and that is what we will use. The program will automatically generate a correlation matrix from the data. In the **Display** group, you will see that **Unrotated factor solution** is ticked. Add a tick against **Scree plot**. In the **Extract** group, the number '1' has been entered against **Eigenvalues over**. In principle components analysis, it is possible to extract as many factors as there are variables. That would not, however, involve any reduction of the data, and it is normal to set some criterion for the number of factors that are extracted. The sum of the squared loadings of the variables on a factor is known as the *eigenvalue* (or latent root) of the factor. Dividing the eigenvalue by the number of variables gives the proportion of variance explained by the factor. The higher the eigenvalue, the higher the proportion of variance explained by the factor, so it is possible to set a criterion eigenvalue for the acceptance of a factor as being important enough to consider. By convention, the usual criterion value is 1. The value of **Eigenvalue over** can be changed to allow more or fewer factors to be extracted, or you can decide on the number of factors you want to be extracted by clicking on the

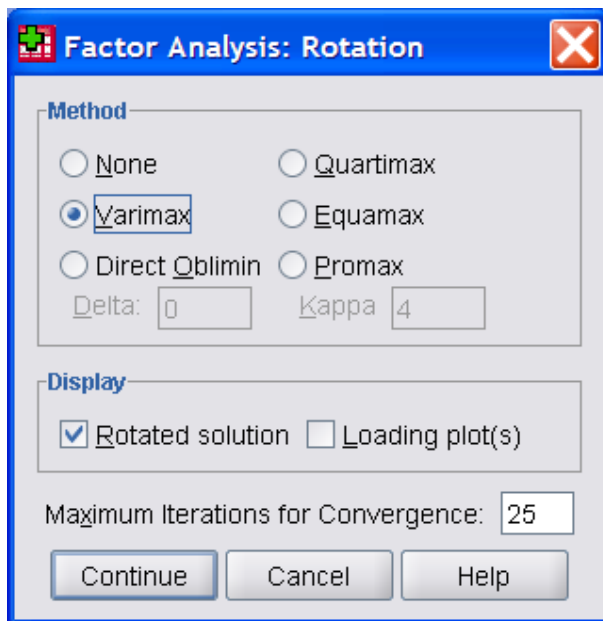
Number of factors button and entering the desired number in the box. At the bottom of the dialog box, the **Maximum Number of Iterations** is set at 25. This simply means that the analysis will stop after 25 iterations of the procedure to generate successive approximations to the best possible factor solution. This number can be increased, but we will leave it as it is. So the box now looks as in SPSS Dialog Box 8.2.



SPSS Dialog Box 8.2. Extracting factors from a correlation matrix.

Next, click **Rotation** to get SPSS Dialog Box 8.3. You will see that the default **Method** is **None**. We want a rotated solution as well as an unrotated one and the most usual method for an orthogonal rotation is **Varimax**. Click that button. If we had wanted an oblique solution, we probably would have clicked **Direct Oblimin** and retained the default value of **Delta** = zero. In the **Display** group, **Rotated solution** has been ticked, which is what we want. If you love complex graphics, you can tick **Loading plot(s)** to see a three-dimensional plot of the rotated factor loadings, but we won't do that. At the bottom, the **Maximum Iterations for Convergence** on the best

obtainable rotated solution is set at 25. We will accept that value. The box now appears as shown.



SPSS Dialog Box 8.3. Rotating the factors

If we wanted to make use of factor scores, we could click on **Scores** and **Save as variables**. That would result in a new column for each factor being produced in the data sheet. We do not wish to save the factor scores and we do not reproduce the dialog box here.

Click on **Options**. If you have any missing data, refer to our discussion at the start of the preceding section (Missing values) and make the best choice you can for dealing with the problem. In our example dataset we don't have any missing data, so we ignore the **Missing Values** box. In the **Coefficient Display Format** box, you can, if you wish, put ticks against **Sorted by size** and/or **Suppress absolute values less than**. The former will result in variables that load on the same factor being located together in successive rows of the rotated factor matrix, and the latter will eliminate loadings below some threshold from both of the factor matrices. The default for the

threshold value is **0.1**, but if you wanted to suppress low loadings in the output you would probably choose a value closer to 0.5 to get an uncluttered view of the factor structure. Of course you can choose a value in between (0.3 is often used); it is a trade off between losing some information and ease of interpretation of the results. The higher the value chosen the fewer results are displayed making the results easier to interpret but you run the risk of missing something interesting just below the threshold. We will not select either of these options on this occasion and we do not reproduce the dialog box here. Click on **Continue** to close any secondary dialog box that is open, then on **OK** in the main dialog box to run the analysis.

EFA: understanding the diagnostic output

The first table in the output (see SPSS Output 8.1) contains the diagnostics we requested. For the KMO index of sampling adequacy, values above 0.6 are required for good factor analysis. Our value of 0.69 is satisfactory. We need a significant value for Bartlett's test of sphericity and we have that. However, it is notoriously sensitive to sample size and is likely to be significant even when correlations are substantial. Consequently, this test is only recommended (e.g. by Tabachnick and Fidell – see Further reading) when there are fewer than about five cases per variable. We have about eleven cases per variable, so it is of no help to us here.

EFA: understanding the initial factor solution

The next table gives communalities (since this is principle components analysis, the *actual* proportion of variance accounted for by the factors), which you can ignore, so we do not reproduce that table here. Next comes the table with initial eigenvalues and proportions of variance explained by each factor. You will see that only the first three factors have eigenvalues greater than 1. As '1' was our criterion for retention of a

factor, that tells us that only the first three factors to be extracted are in the factor solution. Looking at the proportions of variance, we see that the bulk of the variance attributable to the retained factors was explained by the first (general) factor (30.7% out of 56.6%) in the initial solution, whereas the variance was more evenly distributed in the rotated solution (21.7%, 19.4% and 14.5%, respectively). This is always the case. Next comes the scree plot. What we are looking for here is an indication of how many factors should be retained. The eigenvalue criterion tells us that three factors should be retained, but the criterion of 'greater than one' is somewhat arbitrary, and the scree plot might suggest a different number. A sudden change in the slope is often a useful guide. In our plot, this discontinuity seems to occur at two factors rather than three, and we might therefore consider whether a two-factor solution might be preferable. In the end, though, we should generally make the decision on the basis of how many factors are theoretically interpretable. There is little point in retaining factors that do not make sense.

The next table presents the loadings of variables on the factors (or *components*, because we used the principal components method of extraction) for the initial (unrotated) solution. We just note that all of the variables have positive loadings on the first (general) factor, which suggests that a PHI score based on all eleven subtests might be reasonable. On the other hand, some of the loadings are quite modest, so we should not anticipate that the measure PHI would be highly reliable.

EFA: understanding the rotated factor solution

Next comes the rotated factor (or component) matrix. We see that the effect of the rotation is, as usual, to increase the size of high loadings and reduce the size of low loadings. The first two factors seem readily interpretable as biomedical, with high

loadings of lung, liver, kidney and heart function. There is also a relatively high loading (0.49) for the step test on the first factor, which was not anticipated. The second factor seems to be quite clearly identifiable as performance, with high loadings of stamina, stretch, blow and urine tests. The step test, which is supposed to belong with this group of subtests, also has a moderately high loading (0.48) on this factor. The fact that this test has very similar high loadings on two factors is an indication of a lack of *factor purity*. A solution is said to have factor purity when each test loads highly on just one factor, so that the factors are clearly defined by the groupings of tests that load on them. The situation is further complicated by the emergence of a third factor. This has high loadings on two subtests; the muscular and skeletal system tests. Both of these tests involve body strength and fitness, so perhaps the third factor represents something like muscular-skeletal strength, which might be distinct from both biomedical and performance. Remember that, even though we had some expectations about the factor solution, we are engaged in exploratory factor analysis, so our findings (and their interpretations) have the status of hypotheses, not model confirmation.

KMO and Bartlett's Test

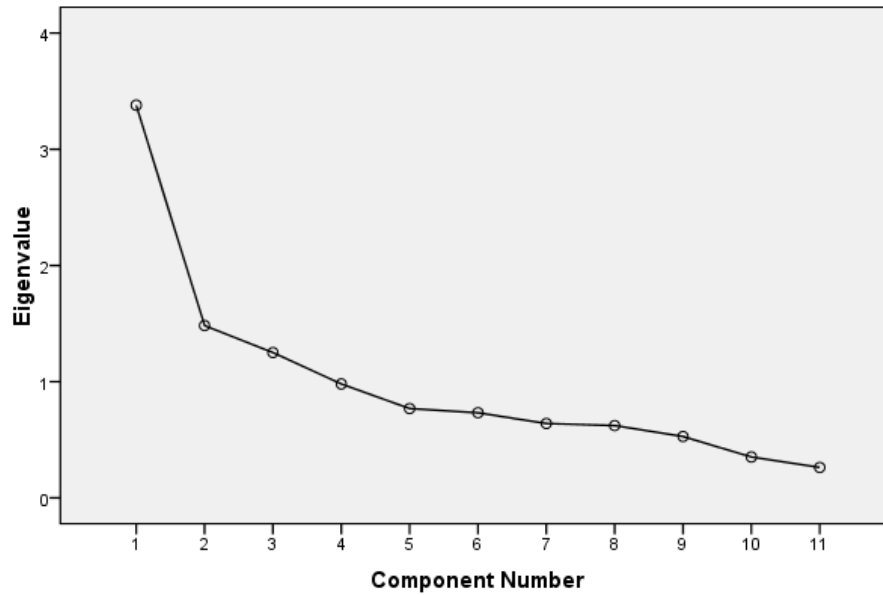
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.687
Bartlett's Test of Sphericity	Approx. Chi-Square	330.640
	df	55.000
	Sig.	.000

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.379	30.720	30.720	3.379	30.720	30.720	2.388	21.705	21.705
2	1.483	13.480	44.200	1.483	13.480	44.200	2.135	19.406	41.111
3	1.251	11.369	55.569	1.251	11.369	55.569	1.590	14.458	55.569
4	.980	8.913	64.482						
5	.769	6.989	71.471						
6	.733	6.664	78.136						
7	.640	5.822	83.957						
8	.622	5.656	89.614						
9	.528	4.803	94.417						
10	.352	3.199	97.616						
11	.262	2.384	100.000						

Extraction Method: Principal Component Analysis.

Scree Plot



Component Matrix^a

	Component		
	1	2	3
lung	.605	-.247	-.219
muscle	.320	-.471	.559
liver	.700	-.300	-.279
skeleton	.584	-.199	.564
kidneys	.610	-.059	-.479
heart	.469	-.435	-.231
step	.670	.132	-.106
stamina	.484	.350	.333
stretch	.638	.287	.285
blow	.587	.461	-.059
urine	.232	.665	-.074

Extraction Method: Principal Component Analysis.
a. 3 components extracted.

Rotated Component Matrix^a

	Component		
	1	2	3
lung	.659	.124	.160
muscle	.110	-.085	.785
liver	.783	.127	.170
skeleton	.185	.286	.763
kidneys	.731	.230	-.135
heart	.646	-.107	.185
step	.486	.480	.102
stamina	.016	.620	.289
stretch	.180	.652	.336
blow	.264	.699	-.041
urine	-.066	.649	-.277

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

SPSS Output 8.1. Selected parts of the factor analysis output

Before we leave our discussion of the rotated factor solution, we show in SPSS Output 8.2 a version of the rotated factor matrix that was obtained when the **Options** button in the main dialog box was clicked and ticks were put against **Sorted by size** and **Suppress absolute values less than** in the **Coefficient Display Format** box, and the default value of minimum size was changed from 0.1 to 0.5. You can see that these options make it easier to see the structure of the solution.

Rotated Component Matrix^a

	Component		
	1	2	3
liver	.783		
kidneys	.731		
lung	.659		
heart	.646		
step			
blow		.699	
stretch		.652	
urine		.649	
stamina		.620	
muscle			.785
skeleton			.763

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

SPSS Output 8.2. Rotated factor matrix, sorted and with loadings below 0.5 suppressed

The reliability of factor scales: internal consistency of scales

Many real correlation matrices may have no clear factor solution. When a factor solution is obtained, we need to consider how confident we can be about using it to measure factors. So far as the meaning of the factors is concerned, that is a matter of subjective theoretical judgement, about which there may be considerable disagreement. We can, however, resolve the statistical question of the *reliability* of factors. Even though we cannot be certain of the meaning of a factor, it is worth knowing whether a scale defined by factor loadings really is measuring a unitary construct. The usual index of the *internal consistency* of a scale (all of the items or

tests that load on the factor are tapping into the same construct) is Chronbach's *coefficient Alpha*. This may be thought of as the average of all of the possible split-half reliabilities of the items or tests that are taken to be indicators of the construct. Values of Alpha range from zero to one and generally values over 0.8 are required for the claim to be made that a scale has high internal consistency.

Requesting a reliability analysis in SPSS

According to our factor analysis, there are several scales that we might be interested in. First, there is the presumed general positive health scale, the first factor in the unrotated solution. We will begin with the assumption that all of the subtests contribute to this scale, though we may entertain some doubts about the contributions of the urine flow test (loading = 0.23) and the muscle test (loading = 0.32). To obtain Coefficient Alpha for this scale, click on **Analyze**, then **Scale**, then **Reliability Analysis**. Use the arrow to move all of the subtests into the **Items** box. You will see that **Alpha** is the default in the **Model** box and that is what we want. You can ignore **Scale label** unless you want to give a name to the scale. This dialog box is straightforward and is not shown here.

Click on **Statistics** and put a tick against **Scale if item deleted**. There are a lot of other options in this dialog box, but we don't need any others and we have not shown it here. Click **Continue** and **OK** for the analysis.

Understanding the reliability analysis for the full scale

The first thing in the output is a Case Processing Summary table that we do not reproduce here. That is followed by the Reliability Statistics table (see SPSS Output 8.3), which tells us that $\text{Alpha} = 0.64$, based on all 11 subtests. That is a bit low and

we wonder whether removal of one or more subtests might improve the internal consistency of those remaining. The final table, Item-Total Statistics (see SPSS Output 8.3) is able to help us with that question. Looking in the final column of the table, we see that if the urine flow test were deleted the value of Alpha would increase to 0.74. If you remove URINE from the variables in the Items box of the main dialog box and re-run the analysis, you will find that the new Alpha is indeed 0.74. If you also look at the new Item-Total Statistics box (which we do not reproduce here), you will discover that there are no further improvements to be made by removal of any other variables, so Alpha = 0.74 is the best we are going to get for the general positive health scale.

Reliability Statistics

Cronbach's Alpha	N of Items
.642	11

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
lung	306.50	789.055	.427	.602
muscle	307.64	861.193	.159	.641
liver	271.77	685.189	.452	.582
skeleton	313.59	834.243	.423	.615
kidneys	300.82	803.755	.454	.604
heart	303.28	837.054	.257	.628
step	310.29	832.412	.543	.610
stamina	310.59	850.338	.380	.622
stretch	287.65	675.316	.444	.584
blow	292.92	784.718	.499	.595
urine	259.55	673.320	.130	.743

SPSS Output 8.3. Main output from reliability analysis

Understanding the reliability analyses for the group scales

You can repeat the reliability analysis for the biomedical and performance groupings of tests (they are shown clearly in SPSS Output 8.2). We will just summarize the results here. For the four tests comprising the biomedical scale, Alpha = 0.69 and the reliability would not be improved by removal of any of the tests. For the four tests

comprising the performance scale, Alpha = 0.44, which can be increased to 0.54 by removal of the urine flow subtest. These values do not give us confidence in the internal consistency of the scales. However, perhaps the rather modest Alpha values for all three scales might have been expected given that the set of subtests was supposed to be doing two different things; providing an overall measure of positive health and measures of two (probably related) components of health. High reliability of the general positive health scale would mean that the components were not being well separated. Conversely, high reliabilities for the component scales would mean that the general measure might be less satisfactory.

An alternative index of reliability: coefficient Theta

There is another index of reliability (internal consistency) of the scale based on the first (general) factor to be extracted. This is coefficient Theta, which is closely related to coefficient Alpha. This index takes advantage of the fact that the first factor to be extracted accounts for the largest proportion of the variance. It is not provided in SPSS, but it is easily calculated using the formula:

$$\theta = \left(\frac{n}{n-1} \right) \left(1 - \frac{1}{\lambda} \right)$$

where n = number of items in the scale, and λ = the first (largest) eigenvalue. The largest eigenvalue for our unrotated solution was 3.38 (see SPSS Output 8.1). For our general positive health scale, that gives a reliability of Theta = $(11/10)(1-1/3.38) = 0.77$, which is approaching respectability.