SOME OBSERVATIONS ON THE SCREE TEST, AND ON COEFFICIENT ALPHA

Larry R. Nelson

Faculty of Education, Language Studies, and Social Work

Curtin University of Technology

Western Australia

Nelson, L.R. (2005). Some observations on the scree test, and on coefficient alpha. *Thai Journal of Educational Research and Measurement (ISSN 1685-6740): 3(1),* 1-17.

This paper re-visits two measurement topics which have featured prominently in our technical literature for the past five decades: the factor structure underpinning a set of test items, and the internal consistency of the score derived from a linear composite of the same items. The paper suggests possible enhancements to our use and understanding of the methods on which these topics are founded.

The scree test

The scree test for the number of factors dates back to Cattell (1966). Cattell reported that the process which led to the development of the scree test was one which had him extracting the principal components of correlation matrices, and then looking for "…various signs…" in the resultant eigenvalues (which he preferred to call latent roots). Writing in response to a request from Current Contents, which cited his work as a "Citation Classic" in 1983, Cattell wrote:

> To my delight, a very simple finding presented itself, namely, that if I plotted the principal components in their sizes, as a diminishing series, and then joined up the points all through the number of variables concerned, a relatively sharp break appeared where the number of factors ended and the 'detritus', presumably due to error factors, appeared. From the analogy of the steep descent of a mountain till one comes to the scree of rubble at the foot of it, I decided to call this the *scree* test.

Is the scree test still widely applied? Undoubtedly. In October 2004 I used the Google.com search engine, asking it to find Internet-based documents containing the phrase "scree test". Google returned several hundred hits.

I recently had the opportunity to review the work of a PhD student at the University of Alberta. The student was clearly aware of more recent articles on the number of factors question (for example, Nandakamur (1994),

was cited), but it was the venerable scree test that was applied to conclude that the two cognitive examinations used in the research were "essentially unidimensional" (Dawber, Rogers, and Carbonaro 2004; Dawber, 2004).

Even Tate (2004), in his thorough paper on contemporary procedures for answering the number of factors question, compared eigenvalue magnitudes as one of the processes used to suggest factorial structure. It is apparent that Cattell's suggestions regarding the study of eigenvalues, and the comparison of their relative magnitudes, are still actively applied.

A problem of the scree test relates to finding that "sharp break" referred to by Cattell. As Hayton et. al. stated (2004, p.193): "Although the scree test may work well with strong factors, it suffers from subjectivity and ambiguity, especially where there are either no clear breaks or two or more apparent breaks. Definite breaks are less likely with smaller sample sizes and when the ratio of variables to factors is low …."

I have spent a bit of time considering this problem. Table 1 represents part of my effort: the table presents scree test statistics for twelve selected data sets, A through L (F1 and F2 are the same data sets, and are further discussed below).

For each data set, Table 1 indicates the type of test used in the data set (cognitive or affective; in all cases, cognitive items were dichotomously scored, while affective items had polytomous scoring patterns); the number of items involved in the test, "Nits"; the number of test takers, "*N*"; the value of coefficient alpha for the test as a whole; and the first eigenvalues of the test's correlation matrix. All correlation matrices had 1s on their diagonal (that is, the correlation matrices were not reduced).

Extracting the eigenvalues of an unreduced correlation matrix is commonly associated with a principal components analysis, a procedure referred to in most texts as a "data reduction" method.

Under each eigenvalue is the percentage of total variance accounted for by the corresponding principal component. Since the correlation matrix has not been reduced by placing communality estimates on its diagonal, the sum of the eigenvalues will equal Nits, the number of test items. Each percentage is found by dividing the eigenvalue by Nits, and multiplying by 100.

Table 1: Scree test statistics for 12 data sets

Set	Type	Nits	*N*	alpha	eigen1	eigen2	eigen3	eigen4	eigen5	eigen6	eigen7
A	aff	100	*307*	0.94	23.95	4.66	3.31	2.90	2.42	2.31	1.99
				%	24.0%	4.7%	3.3%	2.9%	2.4%	2.3%	2.0%
				R^2	0.17	0.68	0.77	0.81	0.85	0.86	0.88
B	aff	64	*342*	0.97	26.17	5.69	3.65	1.85	1.56	1.42	1.32
				%	40.9%	8.9%	5.7%	2.9%	2.4%	2.2%	2.1%
				R^2	0.15	0.46	0.64	0.83	0.86	0.87	0.88
C	aff	20	*3000*	0.67	4.12	4.06	0.82	0.81	0.76	0.75	0.72
				%	20.6%	20.3%	4.1%	4.0%	3.8%	3.8%	3.6%
				R^2	0.36	0.25	0.97				
D	cog	40	*450*	0.87	7.18	2.34	1.45	1.31	1.26	1.18	1.17
				%	18.0%	5.8%	3.6%	3.3%	3.2%	3.0%	2.9%
				R^2	0.30	0.79	0.98				
E	cog	36	*1638*	0.86	6.31	1.53	1.30	1.15	1.08	1.07	1.01
				%	17.5%	4.3%	3.6%	3.2%	3.0%	3.0%	2.8%
				R^2	0.22	0.85	0.93	0.97	0.98		
F1	cog	30	*222*	0.79	4.83	1.83	1.54	1.39	1.36	1.26	1.15
				%	16.1%	6.1%	5.1%	4.6%	4.5%	4.2%	3.8%
				R^2	0.48	0.94	0.97	0.98			
F2	cog	17	*222*	0.81	4.47	1.60	1.28	1.09	0.98	0.91	0.86
				%	26.3%	9.4%	7.5%	6.4%	5.7%	5.3%	5.1%
				R^2	0.46	0.90	0.95	0.97			
G	cog	20	*140*	0.89	7.36	1.60	1.57	1.26	1.05	0.81	0.75
				%	36.8%	8.0%	7.9%	6.3%	5.2%	4.1%	3.8%
				R^2	0.35	0.84	0.84	0.89	0.93	0.98	
H	cog	70	*288*	0.90	9.63	3.31	2.22	1.95	1.91	1.73	1.65
				%	13.8%	4.7%	3.2%	2.8%	2.7%	2.5%	2.4%
				R^2	0.35	0.82	0.92	0.94	0.95	0.96	
I	cog	70	*288*	0.93	12.49	3.93	2.18	1.79	1.77	1.61	1.58
				%	17.8%	5.6%	3.1%	2.6%	2.5%	2.3%	2.3%
				R^2	0.26	0.72	0.92	0.94	0.95	0.96	
J	aff	4	*60*	0.39	1.93	1.31	0.65	0.11			
				%	48.1%	32.6%	16.3%	2.9%			
				R^2	1.00						
K	cog	9	*53*	0.34	1.78	1.51	1.22	1.08	0.92	0.82	0.71
				%	19.8%	16.7%	13.6%	12.0%	10.2%	9.1%	7.8%
				R^2	0.97	0.98					
L	cog	60	*649*	0.86	7.21	1.86	1.58	1.50	1.48	1.42	1.37
				%	12.0%	3.1%	2.6%	2.5%	2.5%	2.4%	2.3%
				R^2	0.32	0.95	0.98				

Thus far the Table 1 statistics discussed are familiar ones. To them I have here introduced "R^2", being, in this case, the proportion of eigenvalue variance which may by explained by fitting a straight line to the scree plot. R^2 is (thus) an indication of the degree to which the eigenvalues, when scree-plotted, fall on a straight line.

Examples are shown in Figures 1 and 2, where I first used Excel to plot all 36 eigenvalues from data set E, and then followed by plotting the last 35 eigenvalues from the same data set. Figure 1 shows that a straight line provides a poor fit to the plot; Figure 2 indicates how the picture changes after the first eigenvalue is omitted from the plot – the straight line does a much better job of fitting the data.

Figure 1: scree plot for data set E (36 eigenvalues).

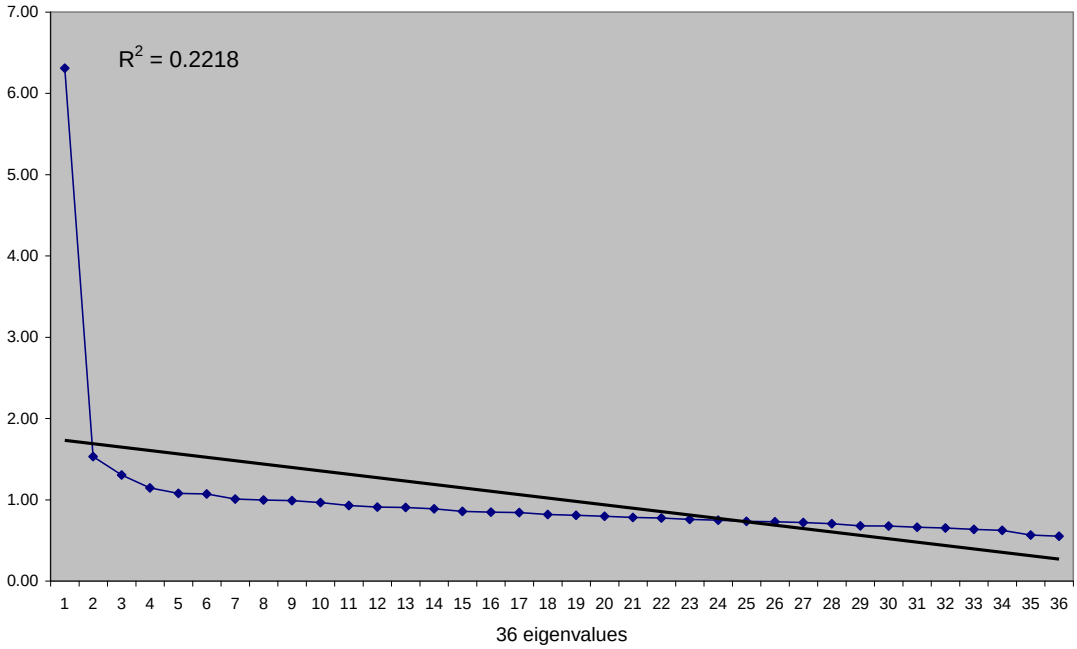
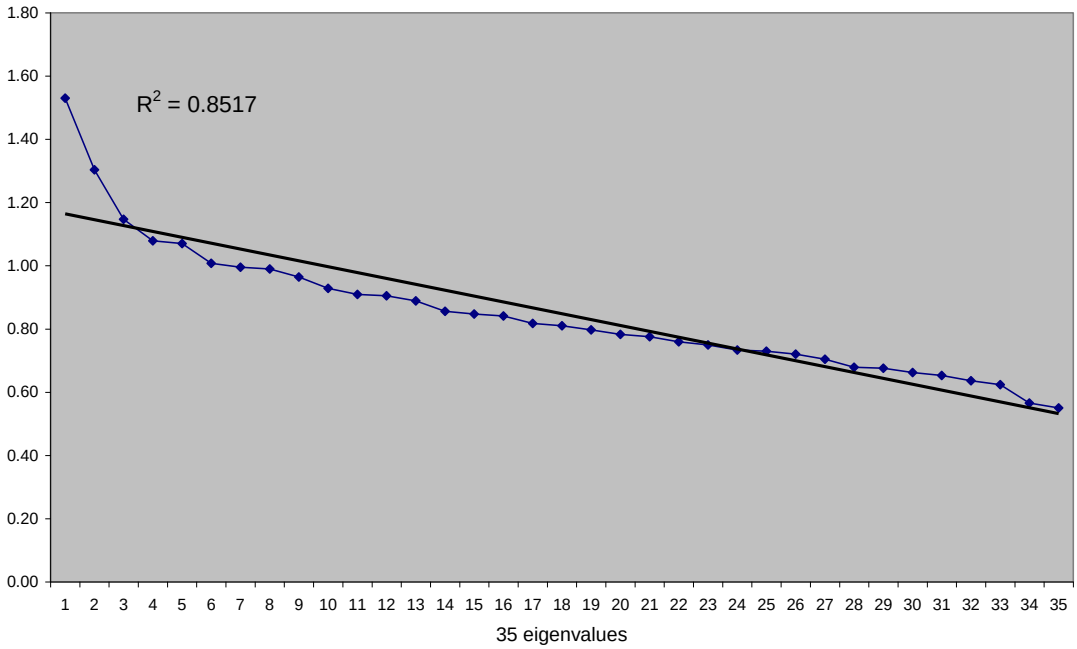


Figure 2: scree plot for data set E (last 35 eigenvalues).



Returning to Table 1, the R^2 figures for data set E reveal what happens after more eigenvalues are removed, one by one. By the time the first three eigenvalues are dropped from the plot, R^2 has increased to 0.97, indicating that a straight line fits the (residual) scree plot quite well.

My thought is that this procedure might have some generalizability, might help to interpret scree tests. Will it? Time will tell; others will have to try it. In the meantime, perhaps we could compare the use of R^2 to other attempts to remove the "subjectivity and ambiguity" mentioned by Hayton et. al. (2004).

For example, a paper by Dawber, Rogers, and Carbonaro (2004) interprets a scree plot in the following manner: "The shape of the scree plot yielded by a principal component analysis showed a dominant first component and the difference between the second and third components, third and fourth, and so forth were small in comparison to the difference between the first and second components, suggesting each examination was essentially unidimensional".

This interpretation strikes me as perhaps falling short of what is required – it is a purely verbal description of a scree plot, and seemingly not very complete. The paper itself provided neither a figure with the scree plot, nor a table with actual eigenvalues. We have no means of confirming the authors' interpretation for ourselves.

In Dawber (2004) we find a more quantitative summary of a scree plot; I believe the interpretive style used in Dawber's work follows that recommended by Hambleton et. al. (1991):

> A principal component analysis … yielded 14 components with eigenvalues greater than 1.0 for the English exam. The eigenvalue for the first component was 8.37, and accounted for 12% of the variance. The difference between the first and second eigenvalues was 7.01, while the differences between successive eigenvalues were small (0.11, 0.05, 0.04, 0.04). Further, the ratio between the first and second eigenvalues was 6.16, while the ratios of the remaining successive eigenvalues were close to 1 (i.e., <1.10). The results suggest that there is one dominant principal component for the English exam.

Compare this with my interpretation of the scree plot corresponding to data set F1 in Table 1:

> When plotted on the entire scree plot, with all 30 eigenvalues, a straight line fell short of providing a good fit (R^2=0.48). Removing the first eigenvalue from the scree plot, and again overlaying a straight line, raised R^2 to 0.94, indicating a good fit, suggesting that the scree began after the first eigenvalue, and supporting the likelihood of a single dominant underlying dimension.

Obviously, I found it necessary to use fewer words than did Dawber. But then, we're not talking about the same data set. It might well take me more words to summarize what happened in, for example, data sets A and B, although I could take a quick stab: in data set A the straight line started to snap in after the second eigenvalue (R^2>0.75), while in data set B the line started to snap in after the third eigenvalue (R^2>0.80).

Some other features of the statistics found in Table 1: there is an apparent inverse relationship between the % values, and corresponding R^2 figures: the higher the %, the lower R^2. And, in general, the straight line overlay seems to snap in place fairly rapidly, the R^2 values usually start low, but quickly switch to 0.80 or better.

I emphasize that I am not suggesting a general rule to the use of this R^2-based scree plot interpretation. At most, I suggest it might come to be a useful companion for those who use scree plots. It is certainly easy to compute in Excel: place the eigenvalues in a row, highlight the row, and use Excel's Chart Wizard to make a "Line" chart. Then, use the Chart drop-down menu from Excel's main toolbar to "Add a trendline". One of the options to the trendline is "Display R-squared value on chart".

Some readers will desire more information on the data sets shown in Table 1. All but two of them are authentic, that is, from real-life applications of cognitive and affective tests. For example, data set A is from a secondary school attitude test used in Puerto Rico. Data set B is from a widely used attitudes-towards-employment instrument developed in the United States. Sets H and I are from norming samples for a high-school aptitude test created in New Zealand. Data set L is from a basic skills competency test used in the United States.

Data sets C and J were made for use in test and measurement classes at the universities of Illinois and Southern California, respectively. The factorial structure of the data in these data sets was defined *a priori*.

All but one of the data sets reflect the work of professional test users, from colleagues with years of experience in our field. The exception is data set K, a cognitive test created by a graduate student with little experience.

In the process of preparing this paper, I looked at numerous other data sets. Common factor analysis operating guidelines, or "rules of thumb", suggest we have "large samples", and/or, samples where the ratio of N to Nits, sample size to number of items, is 3 or better. I used these guidelines to filter my pool of data sets.

Coefficient alpha and principal components

As the developer of the Lertap classical item and test analysis system (Nelson, 2000), I at times receive user questions having to do with coefficient alpha, and with possible methods to maximize it.

One of Lertap's standard reports, "Stats1f", indicates how much a test's alpha value will increase if an item were to be omitted from the test. Users sometimes employ this information to start removing items, one by one, noting the corresponding change in alpha.

This procedure often serves to increase a test's alpha value, but is labour-intensive, and generally quite time consuming.

Another way to go about improving a test's alpha value is to plot item discrimination figures, lopping off the items with the lowest values. Kehoe (1995) is one of numerous authors who report on the results of this method, finding "... a 30-item multiple-choice test administered by the author resulted in a reliability of .79, and discarding the seven items with item-test correlations below .20 yielded a 23-item test with a reliability of 0.88!".

I applied this approach to data set F1.

Table 2 displays the "item discrimination bands" for the 30 items used by data set F1, in a format common to the Lertap 5 Excel application. The table indicates that four questions, p10, p17, p26, and p29 had item discrimination indices, item-test correlations, less than .10. Nine questions had item-test correlations less than .20. Eliminating these nine items {p1, p14, p16, p22, p28 and p10, p17, p26, p29} saw alpha change from its original value of 0.79 to 0.81, a small increase.

Table 2: item discriminations for data set F1

item discrimination bands	
.00:	p10 p17 p26 p29
.10:	p1 p14 p16 p22 p28
.20:	p4 p11 p15 p18 p24 p25 p27 p30
.30:	p9 p13 p19 p20 p21 p23
.40:	p2 p3 p8
.50:	p6 p7 p12
.60:	p5
.70:	
.80:	
.90:	

Some would suggest that a better method of identifying weak items would involve the application of factor analysis. For reasons which will become apparent later, I prefer to use principal components for this task. I had Lertap 5 extract the first principal component from data set F1's corre-

lation matrix; the resultant item-component correlations are summarized in Table 3.

Table 3: item-component correlations for data set F1

P-Comp1 bands	
.00:	p10 p17 p26 p29
.10:	p16 p28
.20:	p1 p14 p18 p22 p24 p25 p30
.30:	p4 p11 p15 p27
.40:	p9 p13 p19 p20 p21 p23
.50:	p2 p3 p8
.60:	p12
.70:	p5 p6 p7
.80:	
.90:	

A common rule of thumb applied in factor analysis is to use correlations of .30 and above to retain items; those items whose correlations with the factor (or component) are less than .30 become candidates for omission from the test.

In the case under question, data set F1, there are 13 items with component correlations below .30. I eliminated these items from the test, had Lertap 5 produce new test statistics, and saved results as data set F2 (seen in Table 1).

Coefficient alpha once again came out to be 0.81.

In the case of data set F1, little was gained by deleting weak items, no matter which method was used to identify such items. I note, however, that Kehoe (1995) might well have a more positive comment; following the theme of his paper, one would think he might suggest that I have indeed gained: I now have a shorter, more reliable test.

I would take this point, and mention that often the benefit realised by eliminating weak items is greater than the 0.79 to 0.81 increase found here in the case of data set F1.

Consider, as another example, data set J, corresponding to an affective test. Table 4 indicates the item-test correlations for the four items belonging to data set J.

Table 4: item-test correlations for data set J

correlation bands	
.00:	Q1
.10:	Q2
.20:	
.30:	Q3 Q4
.40:	
.50:	
.60:	
.70:	
.80:	
.90:	

Putting aside the small number of items in this data set, the pattern seen in Table 4 is typical of tests with low alpha values. One might think there's not much hope for this test as half of its items have very low item-test correlations.

But look at the item-component correlations shown in Table 5.

Table 5: item-component correlations for data set J

P-Comp1 bands	
.00:	Q1 Q2
.10:	
.20:	
.30:	
.40:	
.50:	
.60:	
.70:	
.80:	
.90:	Q3 Q4

Table 5 should revive our interest in this test. The pattern it displays is almost remarkable – we've got two really good items, Q3 and Q4, and two weak ones.

Lertap's bands of item statistics are designed to provide a quick summary of a selected item or component statistic. At times the lowest of these bands may contain negative values. If a particular item statistic is less than zero, such as the item-test correlation, or the item-component correlation, Lertap will position the statistic in the **.00:** band. We do not know if an item found in the lowest band has a low positive value for the statistic, a zero value, or a negative value.

But all is not lost; the actual statistics are always provided somewhere in one of Lertap's reports. Table 6 shows an extract from the "IStats" report, where the complete principal components results are found:

Table 6: item-component statistics for data set J

	Q1	Q2	Q3	Q4
eigens	1.93*	1.31	0.65	0.11
percent	48%	32%	16%	3%
p-comp1	-0.31	-0.15	0.94	0.96

*corresponding alpha = 0.641

Table 6 indicates that Q1 and Q2 do, in fact, have negative loadings[1] on the first principal component. Those who regularly engage in factor analysis will know what to do next: form a test which consists of just the strong items, which, in this case, includes Q3 and Q4.

On doing this, the resultant test was found to have an alpha value of 0.94. Out of interest, I formed another test, one with items Q1 and Q2, finding it to have an alpha of 0.50, and a correlation with the other test, the one with Q3 and Q4, of -0.09.

Returning to the question underlying this discussion, how to improve a test's alpha figure: using item-test correlations, or discriminations, to eliminate weak items is a process with some promise, but, clearly, a better process is to apply a principal components analysis. In the two examples looked at here, data set F1 and data set J, the principal components process was superior in both cases.

True, for data set F1 the change in alpha was the same for both approaches. Both the item-test correlation (or discrimination) process and the principal components process saw alpha increase from its initial value of 0.79 to 0.81. However, the point could be made that the principal components process led to greater parsimony, that is, to a test with fewer items.

The benefit of the principal components process was clearer in the case of data set J, where there was a real difference in the patterns suggested by Lertap's bands of statistics.

I now draw attention to one more advantage of using principal components, and refer back to Table 6.

Notice, in Table 6, that the first eigenvalue has an asterisk attached, with a note to the effect that the "corresponding alpha" value was 0.641.

This output is again from Lertap. What it says is that, were we to apply the p-comp1 loadings to the scoring of the test items, we would end up with a new composite score having an alpha value of 0.641.

[1] Lertap does not rotate the component; as a result, an item's correlation with the component is equal to its loading on the component.

Compare this to the original test alpha of 0.39 for this data set. It would seem there would be much to gain by re-weighting the item scores by the p-comp1 loadings—the increase in alpha is considerable. (In fact I doubt anyone familiar with multivariate analysis would really go for this idea; we are in the same ballpark as regression weights, and using such weights to derive composite scores is a process which is seldom robust. Besides, we already know that we can obtain a much more reliable test by simply eliminating half of the test's items.)

The relationship between coefficient alpha and principal components is provided by the following equation:

$$\alpha_j = \frac{p}{p-1}(1 - \frac{1}{\lambda_j})$$

For the p principal components of a correlation matrix, the value of coefficient alpha for the j^{th} component is given by the equation above, where p represents the number of test items, and λ_j the corresponding eigenvalue.

Applying the equation to the first principal component of data set J, with p=4 and λ=1.93, produces 0.641, the result seen above in Table 6.

It can be shown that coefficient alpha for the first principal component is the maximum possible value which alpha can obtain for any linear composite of the p items comprising the test (see, for example, Hakstain and Bay (1972, p.18), Mulaik (1972, p.211), and Lord (1958)).

Table 7 compares original alpha values for each data set to the value of coefficient alpha corresponding to the data set's first principal component. In the table, α_o refers to the original alpha value, α_{pc} to alpha for the first principal component, and δ to the change.

Table 7: alpha comparisons

Data set	α_o	α_{pc}	δ
A	0.94	0.97	0.03
B	0.97	0.98	0.01
C	0.67	0.80	0.13
D	0.87	0.88	0.01
E	0.86	0.87	0.01
F1	0.79	0.82	0.03
F2	0.81	0.82	0.01
G	0.89	0.91	0.02
H	0.90	0.91	0.01
I	0.93	0.93	0
J	0.39	0.64	0.25
K	0.34	0.49	0.15
L	0.86	0.88	0.02

There are only three data sets with alpha changes of any consequence: C, J, and K. Of these three, only data set K represents an authentic testing situation, the other two data sets were made for class exercises, and had two factors built into them *a priori*.

I expected to see more in this table, to see greater differences in the alpha values. For the moment I can only note that most of these tests are strong ones, as far as alpha goes. The comparison of the alpha values seen in Table 7 indicates that, by and large, the process of adding item scores in a conventional, unweighted manner, is sound, one which could not be much improved on. This is, of course, assuming we are content to use all of the test's original items – it is still possible we could improve on the indicated original alpha values by eliminating items having low correlations with the first principal component. The chances of improving any of Table 7's α_o values are real, as the examples above with data sets F1 and J show. However, when the original alpha values are as high as most of those in Table 7, such work might have little payoff (unless the goal is to shorten the test).

Summary

The investigation of twelve data sets featured in this paper suggests that the interpretation of scree tests might be assisted, might be made more objective, by fitting a straight line to the plot of eigenvalues, and then using R^2 to indicate where the scree begins. It was found that such a line will often seem to "snap in", with R^2 suddenly jumping to values in excess of 0.80 as eigenvalues are removed, one by one, from the plot.

It was also suggested that the use of item-test correlations to eliminate items from a test, with an eye to improving alpha, was not as beneficial as was eliminating items on the basis of their correlation with the first principal component.

A comparison of alpha values was undertaken to investigate the difference between a test's original alpha value and that which would result from re-weighting item scores using item loadings on the first principal component. It was found that tests with high original alpha values would not benefit from such re-scoring.

References

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-76.

Dawber, T. (2004). *An investigation of the robustness of Lord's item difficulty and discrimination formulas.* Unpublished PhD dissertation, Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Dawber, T., Rogers, W.T., & Carbonaro, M. (2004, April). *Robustness of Lord's formulas for item difficulty and discrimination conversions between classical and item response theory models*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Hakstain, A.R. & Bay, K.S. (1972). *User's manual to accompany the Alberta general factor analysis program.* Edmonton, Alberta: Division of Educational Research Services, University of Alberta.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, California: Sage.

Hayton, J.C., Allen, D.G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods, 7* (2), 191-205.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. Washington, D.C.: *ERIC Clearinghouse on Assessment and Evaluation,* ERIC Identifier: ED398237.

Lord, F.M. (1958). Some relationships between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291-296.

Mulaik, S.A. (1972). *The foundations of factor analysis.* New York: McGraw-Hill.

Nandakamur, R. (1994). Assessing dimensionality of a set of item responses – comparison of different approaches. *Journal of Educational Measurement, 31*, 17-35.

Nelson, L.R. (2000). *Item analysis for tests and surveys using Lertap 5.* Perth, Western Australia: Faculty of Education, Social Work, and Language Studies, Curtin University of Technology.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27* (3), 159-203.