Coefficients Alpha and Omega, an Empirical Comparison

Larry R Nelson[1]
Curtin University, Western Australia
Document date: 18 February 2018
website: www.lertap.com
DOI: 10.13140/RG.2.1.1957.5929

Update 18 February 2018

This paper was originally published in June, 2016.

Since then, the special Omega1 macro found in Lertap 5 has been enhanced and is now substantially easier to use. The latest edition of the macro creates two files for use with "**R**", a free programming and data analysis system noted for, among other things, its psychometric utility – numerous free resources in R support a true variety of analysis tools useful for analysing results from tests and surveys.

Interested readers will rush to consult this companion paper.

The original paper:

The standard index of reliability used in Lertap 5 is a statistic commonly referred to as "alpha", "coefficient alpha", and/or "Cronbach's alpha".

Alpha is discussed under the "Reliability" section of Chapter 7 in the Lertap 5 manual (Nelson, 2000[2]).

A somewhat more general discussion of "reliability" and Lertap 5 is found at this webpage. A favourite reference of mine, recommended basic reading, is the text by Meyer (2010).

Alpha is an extremely common measure of reliability, perhaps particularly appropriate when applied in cognitive testing, as in, for example, when assessing student achievement in a particular subject area.

Coefficient omega is also used as a measure of reliability. It has gained prominence after several authors have suggested, and shown, that it is a better index of reliability. Here I refer readers to Dunn, Baguley, & Brunsden (2014), Geldhof, Preacher, & Zyphur (2013)[3], and Revelle & Zinbarg (2008).

In this paper I present results obtained from the application of four programs, using them to get alpha and omega figures for several selected datasets. The programs were: Lertap 5; SAS Proc CALIS; and two CRAN packages, MBESS and psych.

The datasets and results are shown below. In all cases I have reported results without rounding them. Not all programs were used with all datasets. MBESS, for example, proved to be somewhat problematic as I found it very slow when working with large datasets. For SAS, I used the free University version.

---

[1] Comments / questions may be sent to l.nelson@curtin.edu.au

[2] References are found at this webpage.

[3] The Geldhof et al. paper exhibits more balance than the others when it comes to the use of alpha.

Outcomes for selected datasets

1) Results for the **MathsQuiz** dataset with n=999 and items=15

Lertap 5:	**alpha**: 0.7964; alpha(pc1): 0.807
SAS CALIS:	**omega**: 0.7984
MBESS:	**omega**: 0.8075; **alpha**: 0.7964
psych:	**omega**: 0.80; **alpha**: 0.78; omega(h): 0.70
links:	the dataset; output from psych program

2) Results for the **UniAA** dataset with n=127 and items=30

Lertap 5:	**alpha**: 0.7376; alpha(pc1): 0.769
SAS CALIS:	**omega**: 0.7455
MBESS:	**omega**: 0.7445; **alpha**: 0.7361
psych:	**omega**: 0.76; **alpha**: 0.74; omega(h): 0.38
links:	the dataset; output from psych program

3) Results for the **UniBB** dataset with n=132 and items=34

Lertap 5:	**alpha**: 0.8162; alpha(pc1): 0.834
SAS CALIS:	**omega**: 0.8162
MBESS:	**omega**: 0.8190; **alpha**: 0.8162
psych:	**omega**: 0.83; **alpha**: 0.81; omega(h): 0.52
links:	the dataset; output from psych program

4) Results for the **Zmed** dataset with n=2,470 and items=100

Lertap 5:	**alpha**: 0.9522; alpha(pc1): 0.958
SAS CALIS:	**omega**: 0.9543
psych:	**omega**: 0.96; **alpha**: 0.95; omega(h): 0.76
links:	the dataset; output from psych program

5) Results for the **HalfTime** dataset with n=424 and items=100

Lertap 5	**alpha**: 0.9347; alpha(pc1): 0.941
SAS CALIS	**omega**: 0.9154
MBESS	**omega**: 0.9344; **alpha**: problems computing
psych	**omega**: 0.94; **alpha**: 0.93; omega(h): 0.68
links:	the dataset; output from psych program

6) Results for the **NorthernRivers** dataset with n=689 and items=40

Lertap 5	**alpha**: 0.9180; alpha(pc1): 0.920
SAS CALIS	**omega**: 0.9184
psych	**omega**: 0.92; **alpha**: 0.92; omega(h): 0.74
links:	the dataset; output from psych program

7a) Results for the **LenguaBIc** dataset with n=5,504 and items=50

Lertap 5	**alpha**: 0.8099; alpha(pc1): 0.846
SAS CALIS	**omega**: 0.8181
psych	**omega**: 0.83; **alpha**: 0.82; omega(h): 0.72
Note	three items had negative CTT discrimination (I21, I29, I39)
links:	the dataset; output from psych program

7b) Results for the **LenguaBIc** dataset with n=5,504 and items=47

Lertap 5	**alpha**: 0.8288; alpha(pc1): 0.846

psych	**omega**: 0.84; **alpha**: 0.83; omega(h): 0.72
Note	excluding the items with negative discrimination
links:	output from psych program

---

8) Results for the **Negocios** dataset with n=500 and items=60

Lertap 5	**alpha**: 0.8567; alpha(pc1): 0.873
SAS CALIS	**omega**: 0.8497
psych	**omega**: 0.87; **alpha**: 0.86; omega(h): 0.57
links:	the dataset; output from psych program

---

9) Results for the **DunnSES** dataset with n=212 and items=7

Lertap 5	**alpha**: 0.9364; alpha(pc1): 0.940
SAS CALIS	**omega**: 0.9358
MBESS	**omega**: 0.9376; **alpha**: 9363
psych	**omega**: 0.95; **alpha**: 0.94; omega(h): 0.93
links:	the dataset; output from psych program

A link to one of the scripts used with the CRAN *psych* package is here; in this case the script is that written to process the DunnSES data. A link to the corresponding Dunn csv file is here. A helpful, concise guide to the use of CRAN packages is found in Dunn, Baguley, & Brunsden (2014, pp. 407-408)[4].

The program script used to analyse the DunnSES data with SAS University and Proc CALIS may be seen here.

Comments

The main reason I undertook this work was to see if I ought to make an effort to incorporate coefficient omega in Lertap 5. This would, in all likelihood, be quite possible, but I would not expect it to be a piece of cake to program.

However, given the results above, bringing omega into Lertap 5 will not have much priority. There is a resounding indication (in this paper) that it wouldn't be worth it – coefficients alpha and omega are nearly identical in all of the datasets summarized here.

Appendix B indicates how much factor loadings can vary without seeming to markedly diminish the agreement between alpha and coefficient omega.

In Geldhof, Preacher, & Zyphur (2013) mention is made of a six-item scale with item loadings of 0.40, 0.40, 0.60, 0.60, 0.80 and 0.80, finding omega at 0.78 with alpha at 0.77; they report that differences between alpha and alternative reliability estimate are "*… relatively minor … in applied research….*"[5].

Another factor mitigating against bringing omega into Lertap 5 on a priority basis: it is very easy to take results from Lertap and pass them over to the CRAN *psych* package. The output from the psych package is well presented, formatted, and annotated; it includes a useful graphical summary of how items load on factors which I have not included here.

Caveat

---

[4] References are listed at this webpage.

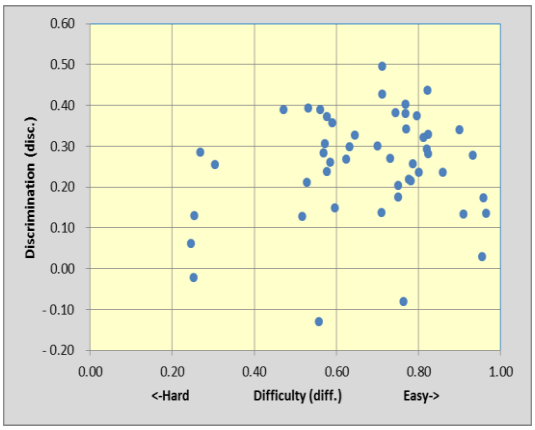
[5] Personal correspondence from Prof. Geldhof received 25 May 2016: "*Your results match my personal experience – alpha and omega give essentially the same results under most conditions*".

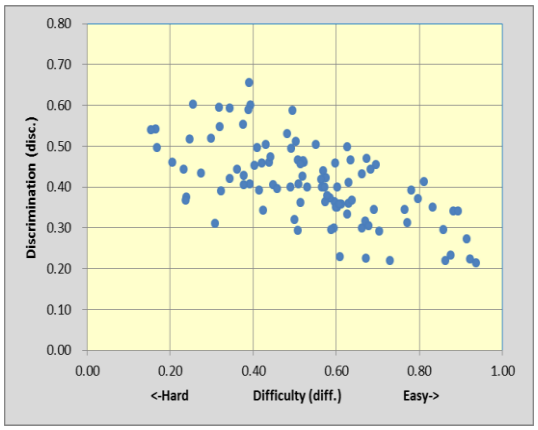
Eight of the datasets used in this study involve multiple-choice cognitive tests. For the most part, these tests were created by educators and professional test developers with substantial experience – by and large, the quality of the tests, from a statistical point of view, was quite good, a notable exception being the test at 7a) above where three items had negative discriminations.

The last dataset, 9) above, was a short affective scale. The seven items in the scale were similar in that their standard deviations did not vary much, ranging from 0.98 to 1.27 (in this case), and the correlations of each item with a composite formed by summing scores on the other eight items in the scale were also similar, going (in this case) from a low of 0.69 to 0.90[6]. This scale, in other words, was a good one, with similarly-behaving, highly inter-correlated items.

I have added these observations as a "caveat" to highlight the fact that the datasets used in this study feature quality instruments, basically free of items with undesirable characteristics.

Undesirable item characteristics would be: items with standard deviations out of line with the other items, and items whose correlations with the other items are noticeably lower. In particular, the correlations should not be negative – the LenguaBIc dataset, 7a above, had three items with negative correlations; at 7b these items were removed and both alpha and omega increased slightly. The LenguaBIc test had 50 items; the impact of removing items with very low correlations will be more substantial when the test or scale has fewer items.





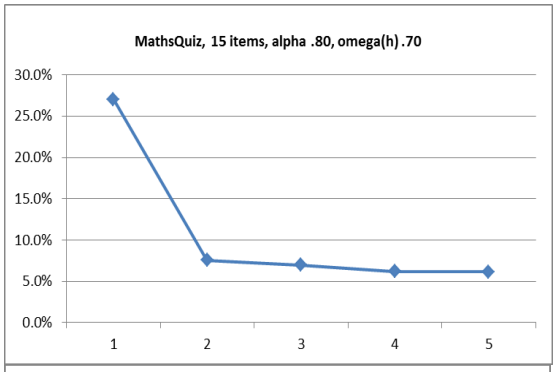
The scatterplots above display the difficulty / discrimination values for each item in the LenguaBIc (left) and Zmed (right) tests. Three LenguaBIc items had correlations (discriminations) below zero; we might call the lowest two of these "outliers", or undesirables. There do not appear to be any outliers among the Zmed items[7], allowing us to say, perhaps, that the Zmed test was found to be desirably free of undesirable items.

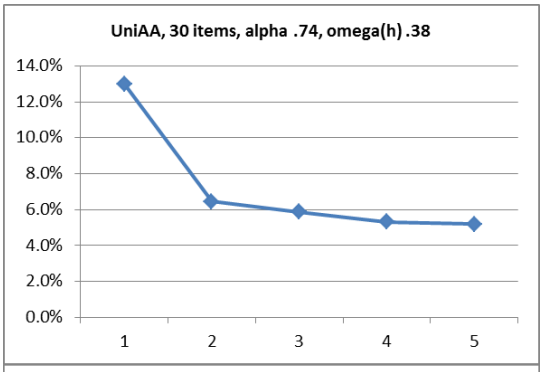
[6] The "s.d." and "cor." columns in a Lertap Stats1b report provide this information.

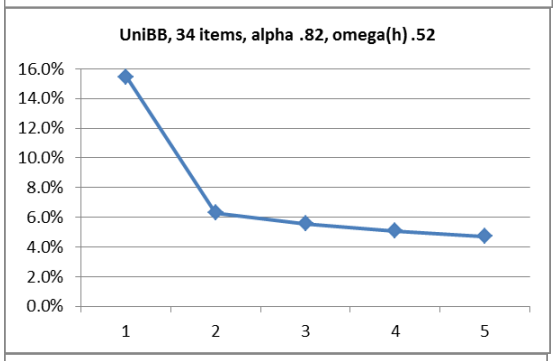
[7] These scatterplots are found in Lertap Stats1b reports for cognitive tests.

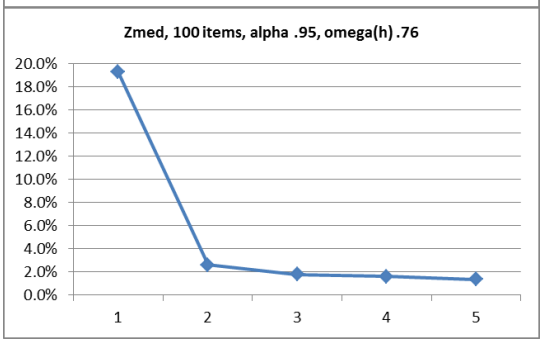
Appendix A: Scree plots

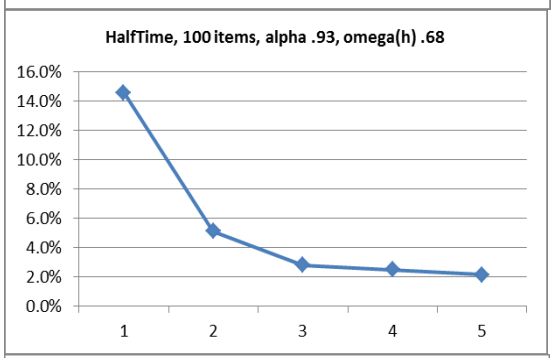
I thought it might be illuminating to see what scree plots of the eigenvalues for the first five principal components might look like, and present them below. I have included omega(h) here as I was looking for a possible trend between it and the size of the eigenvalue for the first principal component. (Trend not found.)

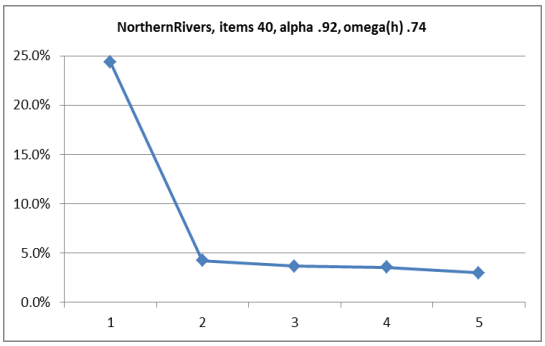


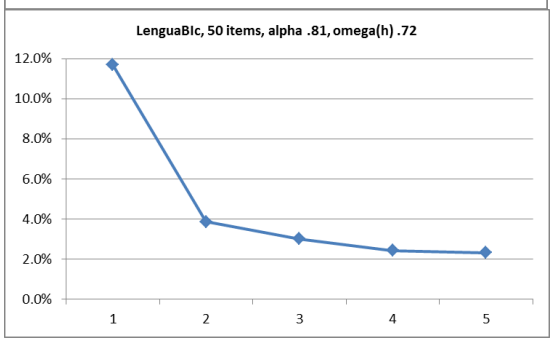


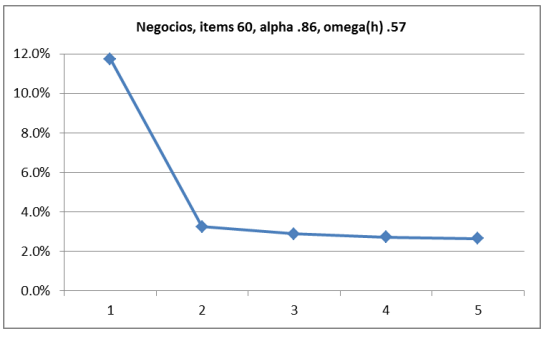


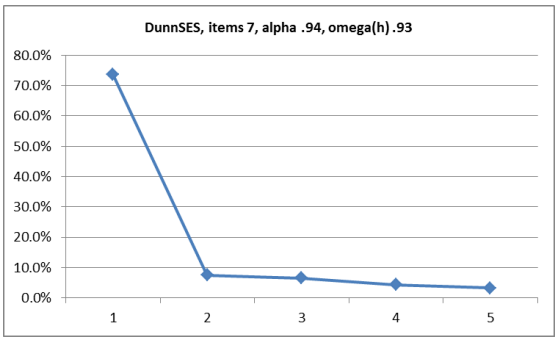












The "scree test" has, over many years, often been used as a test of "unidimensionality".

If a test or scale has just one common factor underlying it, the scree plot should look like that for DunSES above: a very sharp, steep drop from the first eigenvalue to the second eigenvalue, with subsequent eigenvalues decreasing in a very minor, gentle manner, with no bumps along the way.

Thompson (2004) wrote that the factor count ends " … where there is an 'elbow', or levelling of the plot." The matter of determining where the elbow is can be assisted, he wrote, by the "pencil" test: imagine laying a pencil over the scree plot, starting from the right-hand side of the plot. The factor count ends where the points in the plot are no longer covered by the pencil[8].

In the graphs above, the pencil test might suggest there to be only one factor in many of the scree plots, but not with HalfTime, and maybe not with UniAA, maybe not with LenguaBIc.

A problem with the pencil test is that it might not take into account the steepness of the plot, which is an important point. The steeper the plot, the bigger the difference between the first and second eigenvalues, the more likely there may be just one underlying common factor. More support for the single-factor possibility comes when the difference between the second and third eigenvalues is trivial[9].

Note: the scree plots presented above are based on the percentage of variance accounted for by each eigenvalue. This is not normal – users of the scree test generally base the plot on actual eigenvalue values. I suggest that using percentages tends to normalise the plots, making it possible to better compare plots from datasets having a varying number of items. And, using percentages makes it easier to quantify the size of the drop from the first eigenvalue to the second.

Nelson (2005) has more comments on the use of scree plots.

[8] No fair using a really fat pencil.

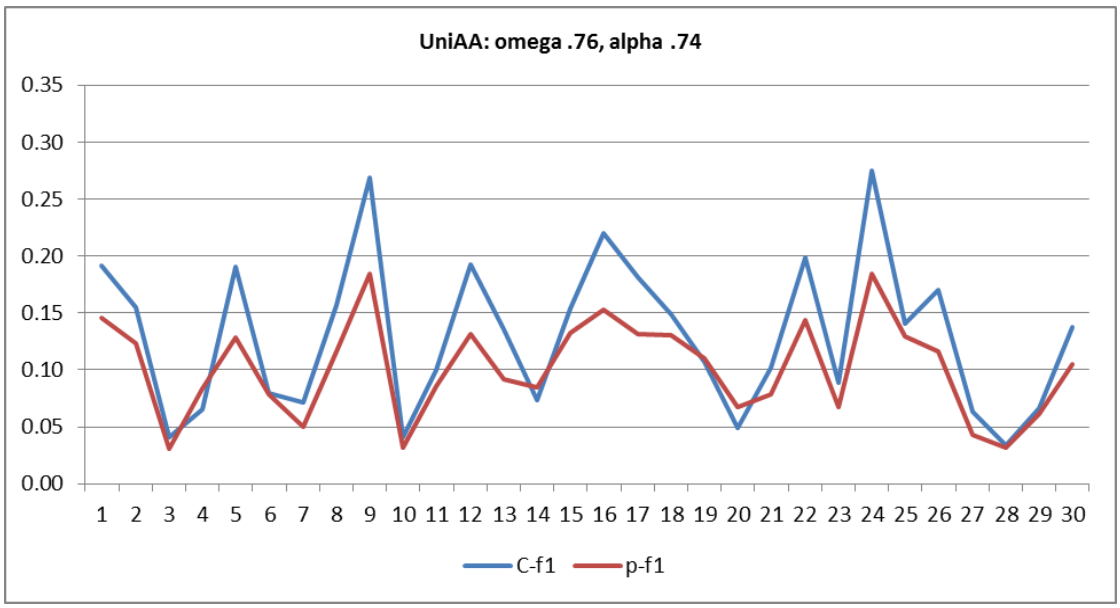
[9] The plots above do not show all eigenvalues, just the first five – the number of eigenvalues will equal the number of items (in a principal components analysis), so the complete plots would show values to the right of those plotted here, each one not greater than the previous one.

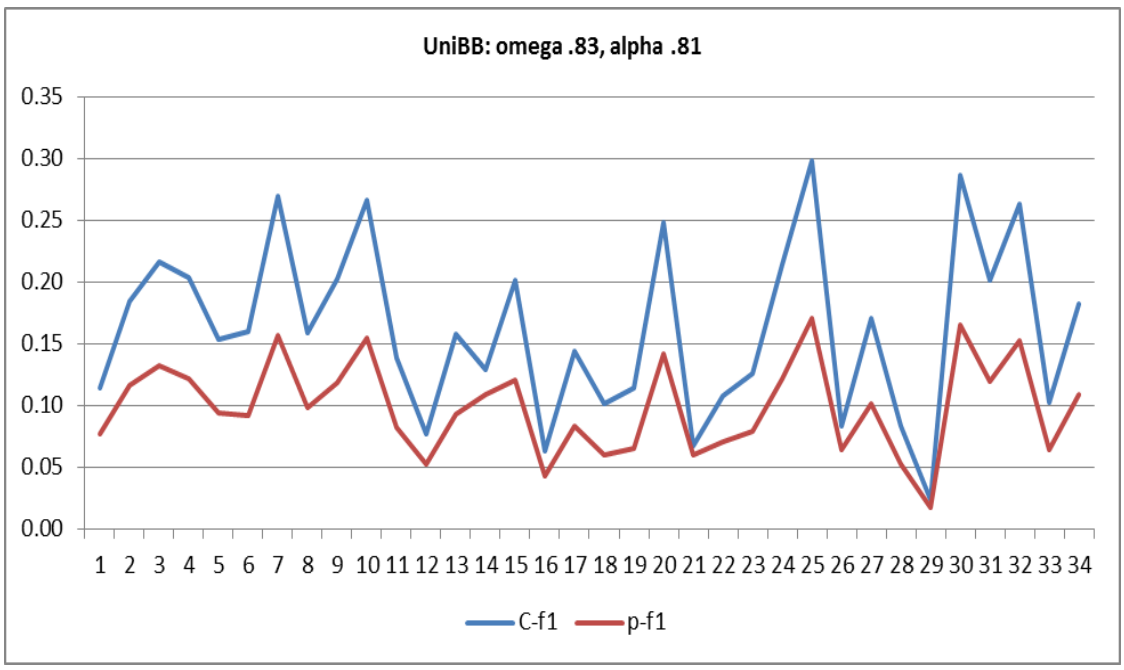
Appendix B: loadings / weights

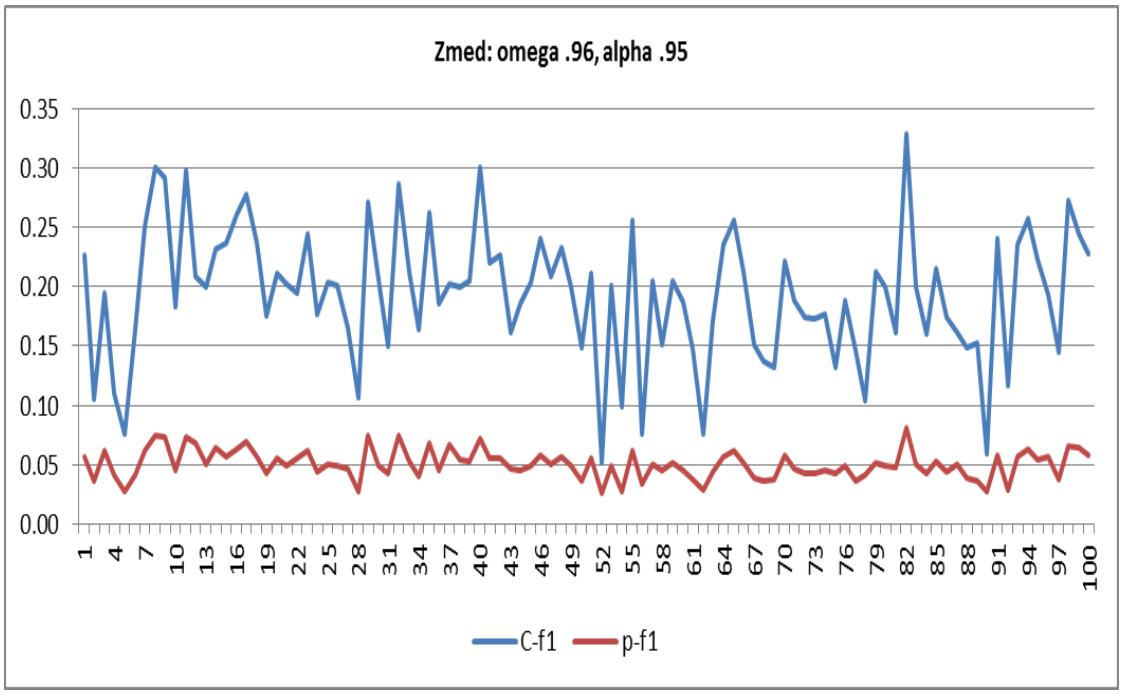
Below is a graphical comparison of the loadings/weights of test items on the first principal factor (unrotated), as output by Lertap 5, and on the latent variable from the SAS Proc CALIS.

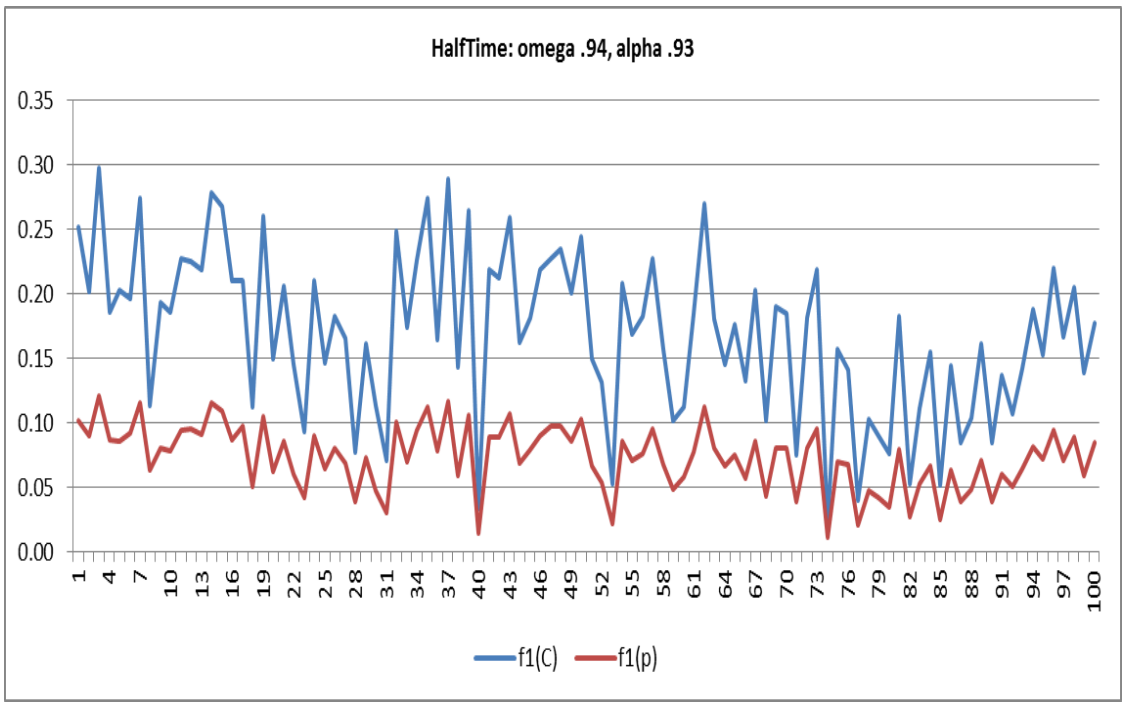
C-f1 is the weight from CALIS on the latent variable, while p-f1 is the loading on the first principal factor from Lertap 5.

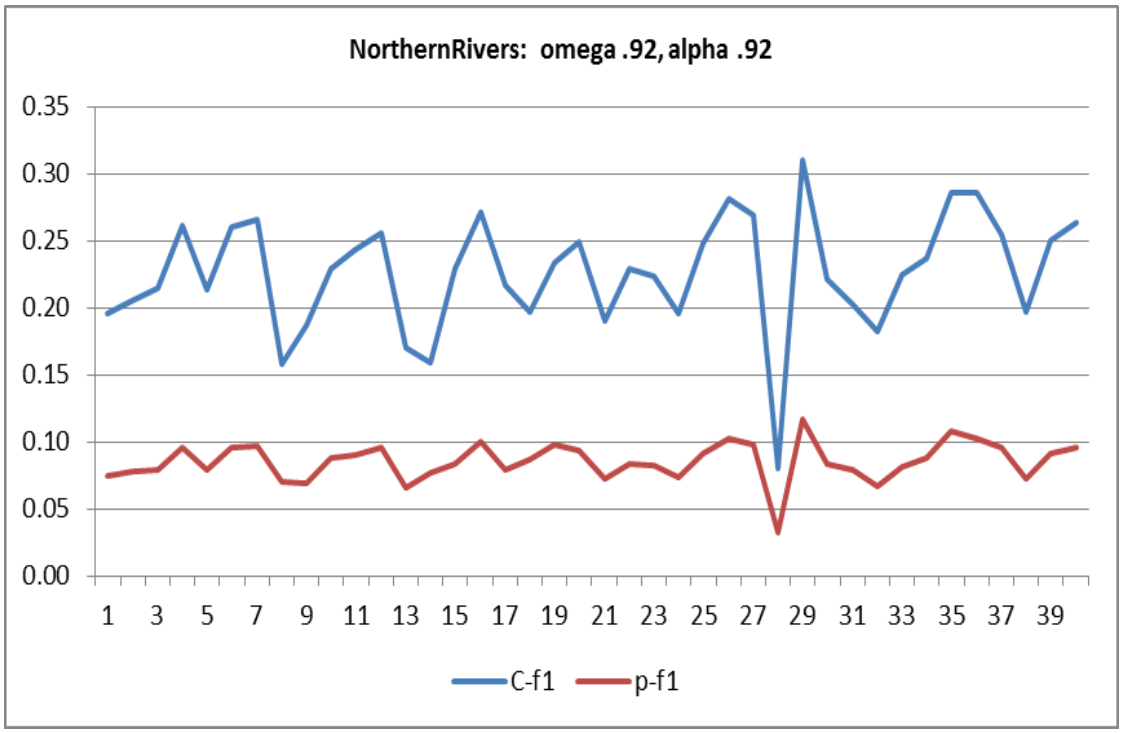
The pattern is similar; the two trace lines tend to rise and fall at the same items. What is interesting is to notice how much the item weights on the latent variable (C-f1) can swing without seeming to markedly serve to distance alpha from omega. C-f1 is the top line in all of these graphs (blue when seen in colour).

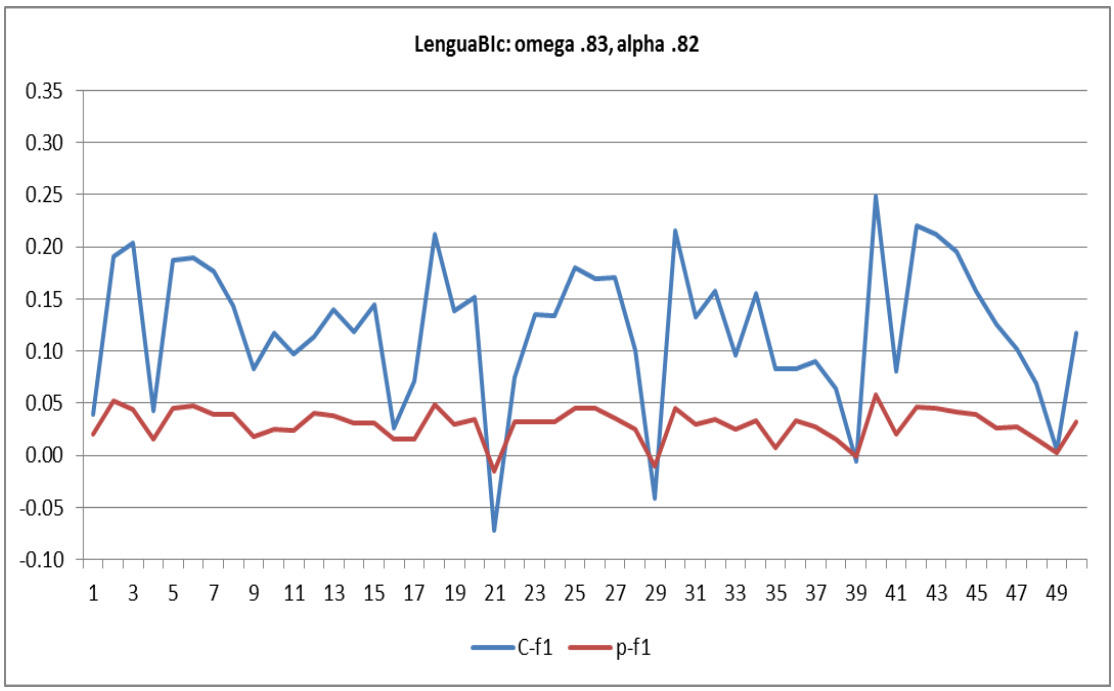





Zmed: omega .96, alpha .95
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
C-f1
p-f1


HalfTime: omega .94, alpha .93
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97 100
f1(C)
f1(p)





Note negative values at items 21, 29, and 39.

References are listed at [this webpage](#).