Europe's Journal of Psychology
ejop.psychopen.eu | 1841-0413

Research Reports

Coefficient Alpha

Interpret With Caution

Panayiotis Panayides*[a]

[a] Lyceum of Polemidia, Limassol, Cyprus.

Abstract

Heavy reliance on Cronbach's alpha has been standard practice in many validation studies. However, there seem to be two misconceptions about the interpretation of alpha. First, alpha is mistakenly considered as an indication of unidimensionality and second, that the higher the value of alpha the better. The aim of this study is to clarify these misconceptions with the use of real data from the educational setting. Results showed that high alpha values can be obtained in multidimensional scales or tests given a sufficient number of items. Therefore, alpha cannot be an indication of unidimensionality. At the same time, after a certain point, higher values of alpha do not necessarily mean higher reliability and better quality scales or tests. In fact very high values of alpha could be an indication of lengthy scales, parallel items or a narrow coverage of the construct under consideration. Researchers are advised to apply caution when reporting alpha.

Keywords: coefficient alpha, reliability, unidimensionality

Europe's Journal of Psychology, 2013, Vol. 9(4), 687–696, doi:10.5964/ejop.v9i4.653

Received: 2013-06-26. Accepted: 2013-08-04. Published (VoR): 2013-11-29.

Handling Editor: Maciej Karwowski, Academy of Special Education, Warsaw, Poland.

*Corresponding author at: Nikou Kavadia 1, K. Polemidia, 4152, Limassol, Cyprus. E-mail: p.panayides@cytanet.com.cy

In many studies on the construction and validation of psychometric scales heavy emphasis is placed on coefficient alpha (Cronbach, 1951). However, there seems to be some confusion as to the true meaning and proper interpretation of this well-known statistic. Many authors have pointed out that alpha is inappropriately used in studies mainly because satisfactory, or even high, values of alpha are persistently and incorrectly used as an indication of the unidimensionality of the scale (Cortina, 1993; Nunnally & Bernstein, 1994; Schmitt, 1996; Sijtsma, 2009; Streiner, 2003). The confusion arises from the use of *internal consistency* and *homogeneity* as if they were synonymous.

Internal consistency refers to the degree of interrelatedness between the items whereas homogeneity refers to the unidimensionality of a set of items. Internal consistency is a necessary, but not sufficient, condition for homogeneity (Green, Lissitz, & Mulaik, 1977).

The formula for alpha is given by

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i V_i}{V_t}\right) \quad \text{(Cronbach, 1951, p. 299)}$$

where n is the number of items, V_t is the variance of the total scores and V_t is the variance of the item scores.

Cronbach (1951) describes alpha as a generalization of the Kuder-Richardson's coefficient of equivalence (K-R20) that has the following important properties:

> (a) α is the mean of all possible split-half coefficients (b) α is the value expected when two random samples of items from a pool like those in the given test are correlated (c) α is the lower bound of the coefficient of precision … (d) α estimates, and is the lower bound to the proportion of test variance attributable to common factors among the items … (e) α is an upper bound to the concentration in the test of the first factor among the items (pp. 331-332).

Cortina (1993) cited a series of definitions and descriptions of alpha and then proceeded to incorporate these in the following description of alpha:

> It is a function of the extent to which items in a test have high communalities and thus low uniquenesses. It is also a function of interrelatedness, although one must remember that this does not imply unidimensionality or homogeneity. (p. 100)

The above description clarifies that high alpha does not necessarily mean unidimensionality. Provided that the items have high communalities (i.e., total variance explained by the factor(s), that is, high loadings of items with one or more factors) alpha will be high, especially so if the scale has a large number of items.

Cortina (1993) has demonstrated, by means of various hypothetical examples, that alpha can be high (greater than 0.7) in spite of low item intercorrelations and multidimensionality, provided there is a sufficient number of items. With average item intercorrelations of 0.50, he found alphas of 0.85 and 0.76 in a two-dimensional and three-dimensional scale respectively, with orthogonal (uncorrelated) dimensions and scale length of 18 items. When the average item intercorrelation was raised to 0.70 alpha values were increased to 0.90 and 0.84.

Values of Alpha and Test Items

Since alpha is influenced by the number of items and the inclusion of parallel items, many scale designers fall into the trap of including an unnecessarily high number of items in an attempt to achieve high alphas believing this to be an indication of a good psychometric scale.

High values of alpha may indicate item redundancy as a result of numerous items that relate weakly to the construct. It may also indicate items with high inter-item correlations which exhibit a narrow coverage of the construct under consideration, thus causing construct underrepresentation and lowering the validity of the scale (Boyle, 1991; 1985; Kline, 1979).

Boyle (1991) argues that in order to maximise the breadth of measurement of the construct one should select items with high loadings on the factors measured but at the same time with low inter-correlations. He suggests that a moderate to low item homogeneity is preferable particularly for the areas of motivation and personality.

Kline (1979) also suggests that each part of the test must be measuring something different (hence the moderate to low item homogeneity). He advises that correlations above 0.70 should be avoided as they make the scale too narrow and too specific. He states that "if one constructs items that are virtually paraphrases of each other, the result would be high internal consistency and very low validity" (p. 292). Nevertheless, in some instances perhaps narrow construct coverage is desirable. Such cases include if one wants to measure a highly specific attribute of a construct, or if the intention is a selection process which, with the use of a test, aims to isolate only the persons with the highest ability.

Having the aforementioned in mind, the practice of merely including additional items must be regarded as an unwise method of increasing the reliability of a scale. Nunnally (1978) recommends reliabilities of 0.70 or better (but not much beyond than 0.80) for basic research and between 0.90 and 0.95 in cases where important decisions are to be made on the basis of the test scores.

This Study

The aim of this study is to help resolve two misconceptions about Cronbach's alpha. First, that alpha is an indication of the unidimensionality of a scale or test and second, that the higher the value of alpha the better. These are explained through a brief review of the literature (presented in the short introduction) and by means of two examples with real data from the educational setting.

Example 1

In this example the effects of changes in dimensionality and number of items on coefficient alpha are displayed. In contrast to Cortina's (1993) report, where the two and three dimensions were made orthogonal, in this example the researcher used real data from the educational setting. For the first example data from Panayides (2009) were used with permission from the author. The data was comprised of 272 high school students' responses to a 20-item maths test, a 20-item language test and a 6-item maths self-esteem (MSE) scale. The two tests were diagnostic tests given to identify weaknesses in basic knowledge in the given subjects with the intention of providing remedial support to those needing it. The maths test contained exercises on algebra (operations with simple fractions, algebraic fractions, expansions, algebraic identities, factorizations, solutions of simple linear and quadratic equations with the use of factorization) and geometry (properties of angles, angles at a point, on a straight line, in a triangle, and in parallel lines). The Language test consisted of a reading comprehension, the text for which was unfamiliar to the students, and a grammar exercise. All test questions were of a familiar nature to the students. A marking scheme was provided for each test to ensure uniform marking by all the teachers involved. The MSE scale contained items such as "I am quite good in maths" and "I have generally done better in Mathematics courses than in other courses" answered on a 6-point Likert scale. The tests were administered at the beginning of the academic year to senior high school beginners.

Initial analyses on the data showed that the maths items had corrected item-total correlations from 0.420 to 0.794. The corrected item-total correlations for the language test varied from 0.204 to 0.507 and for the MSE scale from 0.501 to 0.729. Alpha was found to be 0.910, 0.751 and 0.842 for the maths and language tests and the MSE scale respectively.

The two factors of maths ability and language ability, as measured by the two tests, were moderately correlated. The maths self-esteem (MSE) factor was highly correlated with the math ability and weakly correlated with the language ability.

Furthermore, Confirmatory Factor Analysis (CFA) was performed on the data in order to confirm the factor structure (the three dimensions) of the data. Description of CFA is beyond the scope of this study. It is however worth referring to the goodness of fit (of the model to the data) statistics that are used in this study. Brown (2006) refers to three different groups of indices depending on the information they provide and advises researchers to use at least one from each group. For this reason, the researcher used the Standardised Root Mean Square Residual (SRMR) from the Absolute fit group of indices, the Root Mean Square Error of Approximation (RMSEA) suggested by Browne and Cudeck (1992) from the Parsimony correction fit group and the Comparative fit index (CFI) from the Comparative fit group.

Hu and Bentler (1999) suggested that reasonably good fit of the model under consideration to the data is obtained for

- SRMR values close to 0.08 or below.
- RMSEA values close to 0.06 or below.
- CFI values close to 0.95 or greater, in agreement with Bentler (1990) who adds that values between 0.90 and 0.95 may indicate acceptable model fit.

Table 1 below shows the fit indices for two of the sets of data used. First the two-factor model (maths and language) with data from the two tests only and then the three-factor model (maths, language and MSE) with the whole set of data.

Table 1

Fit Indices for the Two Models

Fit index	2-factor model	3-factor model
SRMR	0.050	0.049
RMSEA	0.031	0.027
CFI	0.938	0.945

The fit indices indicate a good fit of the two different models to the two sets of data. The disattenuated correlation (r_d) between the maths and language factors in the two factors model was 0.647 (corresponding to attenuated r = 0.538) indicating moderately correlated factors. Similarly, Table 2 shows the disattenuated correlations in the three-factor model.

Table 2

Factor Correlations in the 3-Factor Model

	Maths	Language	MSE
Maths	1.000		
Language	0.650	1.000	
MSE	0.839	0.528	1.000

As expected the maths factor and the MSE are highly correlated (r_d = 0.839 corresponding to r = 0.734) whereas there is a weaker correlation between the language factor and MSE (r_d = 0.528 corresponding to r = 0.422)

Alpha Calculations

Alpha was calculated for tests with a different number of items. The tests used included two moderately correlated factors (maths + language), two highly correlated factors (maths + MSE) and two more weakly correlated factors (language + MSE). In doing so, the researcher complements the study by Cortina (1993) who used orthogonal factors. Finally a three-factor test was used (maths + language + MSE). In order to eliminate bias in the item selection process a randomized design was used. The items used in each test were randomly selected from the data set, two or three at a time thus steadily increasing the test length and this can be considered as representative of the item set.

Table 3 shows the alpha values and the corresponding standard errors of alpha (ASE), for the various sample sizes in the case of the 2-factor test with the two moderately correlated factors of maths and language ability. It is obvious that, despite the two different and distinct factors alpha increases as the test length increases. In fact, with around 10 to 11 items alpha exceeds 0.7 (the minimum value considered satisfactory for psychometric scales), with around 18 to 20 items alpha exceeds 0.8 (a satisfactory value for educational tests) and with close to 40 items alpha exceeds even 0.90. As expected, ASE decreases as the number of items increases.

Table 3

Alpha Values for Various Test Lengths (2-Factor Test)

No. of items	Maths items	Lang. items	Alpha	ASE
6	3	3	0.492	0.048
8	4	4	0.613	0.035
10	5	5	0.698	0.027
12	6	6	0.757	0.022
14	7	7	0.773	0.020
16	8	8	0.790	0.019
18	9	9	0.796	0.018
20	10	10	0.802	0.017
24	12	12	0.843	0.014
28	14	14	0.863	0.012
32	16	16	0.881	0.010
36	18	18	0.897	0.009
40	20	20	0.909	0.008

Table 4 shows the alpha values for the various sample sizes in the case of the other two 2-factor tests. First a test with the two highly correlated factors (maths and MSE) and then a test with the two more weakly correlated factors (language and MSE). It is obvious that in the case of the two highly correlated factors alpha reaches 0.8 in a test length of around 7 items and 0.9 at around 16 items. In contrast, in the case of the more weakly correlated factors, alpha reaches 0.80 in a test length of about 16 items. It is clear from the three presented examples that in the case of a 2-dimensional test, given a sufficient number of items alpha will exceed 0.7, or 0.8, or even 0.9, regardless of the dimensionality of the test (this is true only for positively correlated factors). It is also evident that the higher the correlation between the two factors the smaller the number of items required for alpha to exceed 0.80 or 0.90 and the smaller the standard error of alpha.

Table 4

Alpha Values for Various Test Lengths (2-Factor Tests)

2 – factor test (Maths + MSE)					2 – factor test (Language + MSE)				
No. of items	Maths items	MSE items	Alpha	ASE	No. of items	Lang. items	MSE items	Alpha	ASE
2	1	1	0.570	0.052	2	1	1	0.115	0.110
4	2	2	0.637	0.036	4	2	2	0.399	0.060
6	3	3	0.772	0.021	6	3	3	0.594	0.038
8	4	4	0.816	0.017	8	4	4	0.683	0.029
10	5	5	0.831	0.015	10	5	5	0.728	0.025
12	6	6	0.877	0.011	12	6	6	0.785	0.019
16	10	6	0.894	0.009	16	10	6	0.800	0.018

Table 5 shows the alpha values for the various sample sizes in the case of the 3-dimensional test. It is clear that, despite the three different and distinct factors alpha again increases as the test length increases. In fact, with a test length of around 8 items alpha exceeds 0.7, with around 13 items alpha exceeds 0.8 and with around 36 items it exceeds 0.9.

Table 5

Alpha Values for Various Test Lengths (3-Factor Test)

No. of items	Maths items	Lang. items	MSE items	Alpha	ASE
6	2	2	2	0.603	0.037
9	3	3	3	0.714	0.026
12	4	4	4	0.797	0.018
15	5	5	5	0.828	0.015
18	6	6	6	0.856	0.013
26	10	10	6	0.878	0.011
36	15	15	6	0.902	0.009
46	20	20	6	0.925	0.007

Example 2

This example demonstrates why higher values of alpha do not necessarily mean higher reliability and a better scale or test. For the purposes of this example the researcher refers to the findings of Panayides and Walker (2013) whose study was on the psychometric properties of the Foreign Language Classroom Anxiety Scale (FLCAS).

They report five different studies, including the original one on the construction of the scale, in which alpha values were indeed very high ranging from 0.90 to 0.96. Panayides and Walker (2013) investigated the reasons for these high values. Of the 33 items only 5 had corrected item-total correlations below 0.5; in fact 15 (almost half) of these correlations were above 0.70, not very desirable values according to Kline (1979) because these could suggest a very narrow construct coverage. The range of the construct coverage was investigated by Panayides and Walker (2013) with the use of the Rasch models. These models have a unique advantage over all other models in Item Response Theory, the placement of persons and items on the same axis, with the same measurement units, the logits. Figure 1 displays the distribution of person measures on the FLCA axis and item estimates.

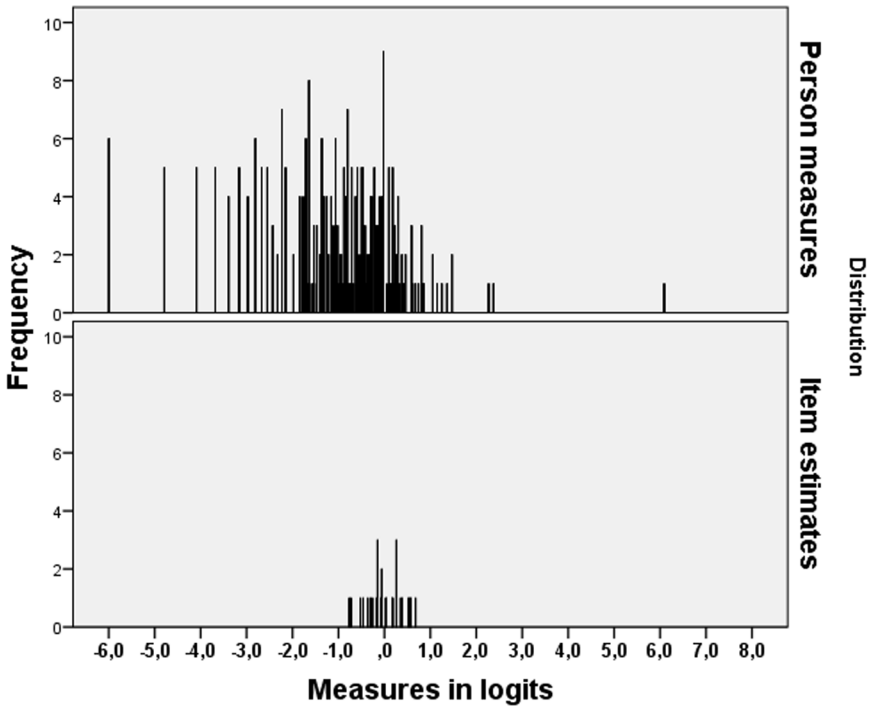

Figure 1. Distributions of person measures and item estimates.

The coverage of the construct by the items of the scale was indeed too narrow. The distribution of person estimates varies from about – 5.0 to 3.0 logits, a range of about 8.0 logits (log-odds units, onto which the raw scores are transformed by the Rasch models, thus achieving interval-level measurement from ordinal-level scores). The one person with a measure of approximately 6.0 is the one with the maximum possible score on the scale. Similarly the six persons with an estimate of -6.0 are the ones with the minimum possible score on the scale. In a very notable contrast the item difficulties only covered a range of just 1.44 (from –0.76 to 0.68) logits. Figure 2 shows the Test Information Function (TIF) for the scale used. This curve shows the range of values of the measures for which information about person estimates is high (i.e., where the standard errors of the estimates are small).

The ideal shape for a TIF is rectangular, meaning high information all over the construct continuum. In this case however, there is a very narrow peak indicating that high information is obtained for about a range from -2.0 to 2.0 logits.

As a consequence of this narrow coverage of the construct a large proportion of person estimates, the ones positioned far from the narrow range covered by the items, have larger errors of estimate. Thus the precision of these person estimates is jeopardized.

Panayides and Walker (2013) also reported that the scale included many parallel items including a group of six parallel items with very similar statistics (item measures from -0.15 to 0.26 and point measure correlations from 0.63 to 0.76) and high inter-correlations (from 0.46 to 0.69). When any five of these six items are removed, leaving one as a representative of the whole group of the six items, alpha changes from 0.963 to a range from 0.953 to 0.955, depending on which five are removed. Therefore, the removal of five of the six parallel items does not significantly affect the value of alpha.

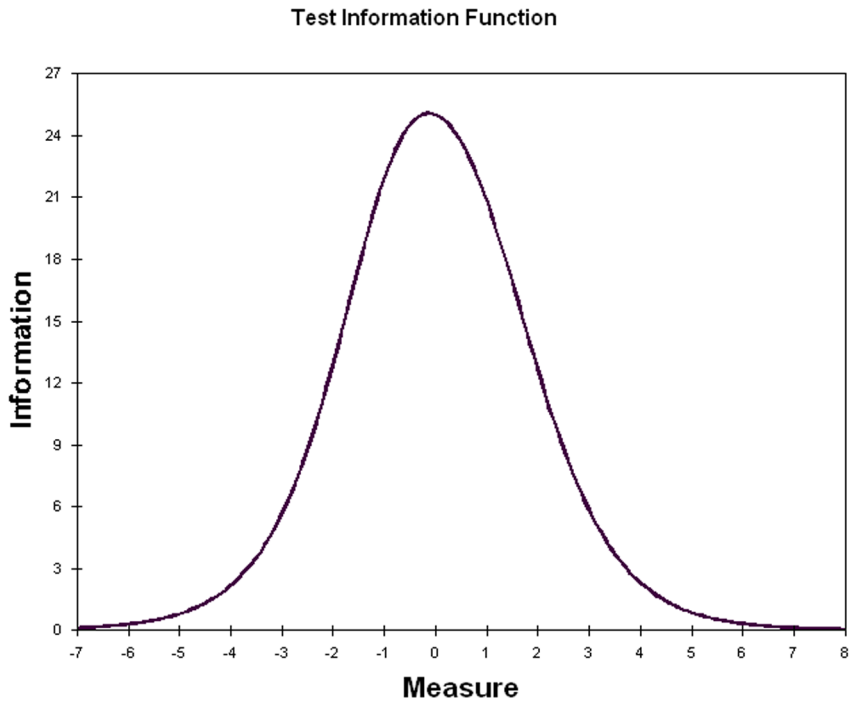

Figure 2. Test information function.

Conclusions

Given the heavy reliance on Cronbach's alpha in many validation studies, the aim of this study is to clarify possible misconceptions on the following two questions with the use of real data from the educational setting.

Question 1: Is Alpha an Indication of the Unidimensionality of a Scale?

Extending on Cortina's (1993) examples with hypothetical data with two or three orthogonal dimensions, the researcher used data from two educational tests, a maths and a language test, and from a 6-item math self-esteem (MSE) scale. The conclusion drawn from these examples is clear. Given any combination of a 2-dimensional test (whether the two factors are highly, moderately or weakly correlated) with a sufficient number of items alpha can exceed 0.70 or 0.80 or even 0.90. The researcher suggests that no one can argue that using maths items together with MSE items, or language items with MSE items will constitute a unidimensional test and yet high alphas can be achieved. In fact, in these specific examples, with only 16 items alpha reached 0.90 in the first case, where the two factors were highly correlated and 0.80 in the second where the two factors were more weakly correlated. A similar conclusion can be drawn from the 3-factor dataset. Given a sufficient number of items alpha will reach very high values.

Therefore, the answer to the first question is that alpha is not an indication of the unidimensionality of the scale since high alphas can be achieved with multidimensional scales, given a sufficient number of items. What alpha indicates is merely the length and the interrelatedness between the items of a scale.

Question 2: Are Higher Values of Alpha Always Better?

It is made clear in the literature that very high values of alpha could mean lengthy scales, parallel items, item redundancy or narrow coverage of the construct (or construct underrepresentation). Whichever of these reasons applies, it is generally accepted that this lowers the validity of the scale (Kline, 1979).

In the example referred to in this study the reason for high alpha values was a combination of the aforementioned reasons. The researcher argues that the narrow construct coverage intuitively leads to an oxymoron: after a certain value (say above 0.90), higher values of alpha represent possibly higher item reliability but lower person reliability, in the sense of accuracy of measurement. The explanation is quite simple. Narrow construct coverage leads to a larger proportion of person estimates outside the item targeting range, which in turn leads to a larger proportion of higher errors of estimates in the person measurement process. Therefore, the precision, and thus the reliability, of person measures decreases.

Concluding Remark

Researchers are advised to apply caution in reporting alpha and to bare in mind two important facts. First, as reported in the literature and demonstrated in this study, high values of alpha do not necessarily mean that the scale is unidimensional. Evidence regarding the factor structure of the data has to be collected through factor analytic methods. Second, alpha "should not be too high (over 0.90 or so). Higher values may reflect unnecessary duplication of content across items and point more to redundancy than to homogeneity" (Streiner, 2003, p.102). Furthermore, higher values may reflect a narrow coverage of the construct which jeopardizes the precision of a large proportion of person measures.

References

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246. doi:10.1037/0033-2909.107.2.238

Boyle, G. J. (1985). Self-report measures of depression: Some psychometric considerations. *The British Journal of Clinical Psychology, 24*(1), 45-59. doi:10.1111/j.2044-8260.1985.tb01312.x

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*(3), 291-294. doi:10.1016/0191-8869(91)90115-R

Brown, T. A. (2006). *Confirmatory Factor Analysis for applied research.* New York, NY: The Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230-258. doi:10.1177/0049124192021002005

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement, 37*, 827-838. doi:10.1177/001316447703700403

Hu, L., & Bentler, P. M. (1999). Cutoff crieteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. doi:10.1080/10705519909540118

Kline, P. (1979). *Psychometrics and psychology*. London, United Kingdom: Academic Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994) *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill, Inc.

Panayides, P. (2009). *Exploring the reasons for aberrant response patterns in classroom maths tests* (Doctoral dissertation, Durham University, Durham, United Kingdom). Retrieved from http://ethos.bl.uk/OrderDetails.do?did=3&uin=uk.bl.ethos.492990

Panayides, P., & Walker, M. J. (2013). Evaluating the psychometric properties of the Foreign Language Classroom Anxiety Scale for Cypriot senior high school EFL students: The Rasch measurement approach. *Europe's Journal of Psychology, 9*(3), 493-516. doi:10.5964/ejop.v9i3.611

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. doi:10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse and the very limited usefulness of Cronabch's alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/s11336-008-9101-0

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

About the Author

Panayiotis Panayides holds a BSc in Statistics with Mathematics (Queen MaryCollege, University of London), an MSc in Educational Testing (Middlesex University, UK) and a PhD in Educational Measurement (University of Durham, UK). He is currently an assistant headmaster and head of the Mathematics department at the Lyceum of Agros in Cyprus. His research interests include educational and psychological measurement and research into mathematics education.