Reliability of assessments in engineering education using Cronbach's alpha, KR and split-half methods

Stephen O. Ekolu & Harry Quainoo

University of Johannesburg
Auckland Park, South Africa

ABSTRACT: An attempt to apply reliability measurements to module assessments in engineering degree programmes is presented in this article. Three techniques for estimating internal consistency - the Cronbach's alpha, KR 21 and split-half methods - were employed in the study. Ten-year data of examination marks were used. The data comprised 13 modules of small- to medium-size classes involving 723 students taught at the university to BSc/BEng engineering degree level. Overall, a majority of modules gave acceptable reliability coefficients of 0.4 to 0.8, based on results obtained from all three methods. A strong correlation was found between Cronbach's alpha and the split-half method. Good correlation of KR 21, with alpha coefficient and split-half methods, occurred only for alpha values exceeding 0.3. It was shown that internal consistency measurement can provide meaningful evaluations of module examinations or assessments in engineering study programmes.

Keywords: Engineering education, assessments, Cronbach's alpha, KR 21, split-half, reliability coefficient

INTRODUCTION

There are different types of reliability measurement methods [1][2]. Of interest in the present work are the internal consistency measurement methods of *Cronbach's alpha*, *Kuder-Richardson formulae 20/21*, and the *split-half method*. Equation (1) gives the formula for Cronbach's alpha (α). The method can be used for both the dichotomous and polytomous scoring of items, of which the latter employs a Likert scale [3].

$$\alpha = \frac{N.\bar{C}}{\bar{v} + (N-1).\bar{C}} \tag{1}$$

where: N - number of test items or questions;

$\bar{v}$ - the average of all variances of the test items;

$\bar{C}$ - average of all covariances between the paired test items.

Of the two types of Kuder-Richardson (KR) formulae, i.e. KR 20 and KR 21, the latter gives a direct estimation of reliability using a minimal set of data, requiring only the number of test items, mean and variance, as given by Equation (2) [4].

$$KR21 = \frac{N}{N-1}\left(1 - \frac{\bar{X}(N-\bar{X}}{N.\sigma^2}\right) \tag{2}$$

where: $\bar{X}$ - the mean of all results or scores;

N - the number of test items or questions;

σ^2 - variance of all results or scores.

In applying the KR formula, it is assumed that all the test items are of the same level of difficulty. KR 21 gives reliability index values lying between 0 and 1, as does Cronbach's alpha. For the split-half method, a set of measurements are divided

into two tests, typically by splitting the number of test items into even items for one test half and odd items for the other test half. Reliability is then estimated as the Pearson-moment correlation coefficient (PMCC), r_{xy}, between the scores or results of the even and odd items. Equation (3) gives the PMCC formula for calculating the split-half reliability. Since the number of test items/questions is less than the full length of the test, e.g. three hours, an adjustment is made in the split-half test calculations, to allow for the shorter test length. The Spearman-Brown formula given in Equation (4) is used to calculate the adjusted reliability r_p.

$$r_{xy} = \frac{\sum xy}{N-1(SD_x)(SD_y)} \tag{3}$$

where:

$$SD_x = \sqrt{\frac{\sum x^2}{N-1}},\ SD_y = \sqrt{\frac{\sum y^2}{N-1}},$$

N - the number of respondents;
x - the residual (score-mean) on all even items for each respondent;
y - the residual (score-mean) on all odd items for each respondent;
SD_x and SD_y - standard deviations for the even and odd items, respectively;

and

$$r_p = \frac{2.r_{xy}}{1+r_{xy}} \tag{4}$$

ISSUES ON RELIABILITY MEASUREMENT

The reliability methods described in the foregoing above are used to measure the internal consistency of test items, basically indicating how interrelated the test items or questions may be [5]. The more interrelated (*unidimensional*) the items are, the higher the calculated reliability coefficient. However, there is no clear agreement on the specific criteria for interpreting Cronbach's alpha [6]. A common interpretation of the coefficient is $\alpha < 0.5$ for low reliability, $0.5 < \alpha < 0.8$ for moderate (acceptable) reliability, $\alpha > 0.8$ for high (good) reliability. A low alpha value may result from:

- a small number of test items or questions;
- heterogeneity of items which measure more than one concept, construct or knowledge area;
- poorly interrelated items.

It may be noted that aiming for high reliability would make the items more similar and less unique in assessing different knowledge areas of the domain. Consequently, content validity can be adversely affected by high reliability coefficients. This aspect can be an important consideration in preparing examination item questions for engineering modules.

Tau-equivalent is a condition requiring that all test items must measure the same trait, aversion or ability. It is also referred to as *a unidimensional assumption* [7]. However, it has been shown that violation of this assumption may not severely affect reliability estimates. Actually, multidimensional tests can have an alpha value that is similar or higher than results from the same form of unidimensional tests. These observations have led researchers to consider Cronbach's alpha as a measure that may not be confined to internal consistency only [5]. Since most test measurements are inherently heterogeneous, the tau-equivalent assumption is often violated. As such, the alpha value calculated is considered to be the lower limit estimate of reliability [5].

The length of a test influences the value of alpha calculated. Longer test lengths, such as those with a large number of test items, give higher alpha values. A small number of test items would violate tau-equivalence and give a lower reliability coefficient [5][8]. Some researches [6][7][9] indicate 20 or more test items as the required minimum number for internal consistency measurements. Also, it can be expected that each group of test-takers would have different characteristics and performance levels, each of which should give a different reliability value [10]. Therefore, it is evident that reliability is not only a property of the test items but also of the test group. In some cases, it may be useful to have a control test group as a reference. It is preferred that a heterogeneous test group be used in measurements as this leads to an improved reliability coefficient [1][5][7].

It has been suggested that the difficulty of items selected should be such that 40% to 60% of respondents would give a correct response [1]. While deciding on item difficulty - which may be easier in psychometric tests - it has to be carefully considered with respect to formative or summative assessments in academic studies, such as engineering.

EXAMINATION ASSESSMENTS IN ENGINEERING PROGRAMMES

In an earlier associated article, a detailed discussion was given on the structure of examination assessments for engineering modules [2]. The structure of these assessments tends to conflict with the requirements for internal

consistency measurement. While an internal consistency test evaluates a close association among test items, engineering examination assessments are heterogeneous as they usually cover diverse knowledge areas. Also, essay-type test items of varied levels of difficulty and lengths are typically used. These test items may also attract different marks.

Four carefully selected test questions are usually sufficient for a summative assessment or examination lasting three hours. This small number of test items used in the examinations falls short of the minimum 20 questions recommended (in some literature) for psychometric tests [6][9]. The examination assessments for engineering modules rarely use multiple choice items or a large number of test questions. Accordingly, summative assessments for engineering modules usually violate the homogeneity assumption required in internal consistency tests. Fortunately, heterogeneity may not lower the calculated reliability value. This is consistent with suggestions in the literature that Cronbach's alpha also appears to measure characteristics other than internal consistency. Tan showed that the alpha coefficient of a heterogeneous test measurement would be incorrect if calculated as a composite [7]. They found that the KR 20 for a set of 237 test questions, made of seven subtests, gave a reliability coefficient higher than those calculated for each subtest. Yet according to the nature of internal consistency measurement, using the composite to calculate KR 20 should give a lower coefficient than those obtained using subtests. Other similar studies have reported the same result [11][12].

Indeed, various researches have shown that using a large number of test items has an overriding effect of increasing alpha, whether the items are homogeneous and unidimensional or not [13-15]. Panayides conducted a study involving 272 high school students on a 20-item mathematics test, 20-item English test and 6-item mathematics self-esteem test [13]. By calculating the alpha coefficient for a varied number of test items and for various two-factor or three-factor models, Panayides found that all the models gave higher alpha coefficients as the number of test items increased [13]. Interestingly, other researches have also indicated that a small number of test items can, too, give high alpha coefficients [16]. These observations underscore the lack of consensus relating to the suitable number of test items [17].

METHODOLOGY

The data for the present study consisted of summative examination marks of students who undertook the four-year BSc/BEng degree programmes in civil engineering. The data were acquired over a 10-year period. Altogether, the summative assessment data were taken from 13 second- to fourth-year modules involving 723 students, as shown in Table 1. The class sizes for each module varied from 15 to 106 students. These classes fall within the category of small- to-medium size. There is no strictly standardised grouping of class sizes, so various researches use different class size groupings [18-20]. For the purposes of the present study, class sizes that were under 20 students were considered small, those with 20 to 90 students were medium, and those with more than 90 students were considered large class sizes [18]. Heterogeneity of the class groups was evident in the assessment results [2][21]. The marks always showed a normal distribution, thereby indicating a properly composed group. Reliability coefficients were calculated for each module assessment.

It may be noted that the modules had the same test length, i.e. the same number of questions; however, the test items/ questions were not of the same level of difficulty. Normal distribution characteristics and linearity are a requirement for data that is used for alpha reliability estimation. It has been shown that alpha is affected by even small deviations from normal distributions [17]. This emanates from the effect which tails in skewed distributions have on variance.

Table 1: Data of summative examination assessments for various engineering modules.

Module	S414	S415	M215	S423	S312	S313	S424	M214	MN04	U315	V314	U314	V315
Class size	65	56	15	60	79	71	58	106	38	48	42	46	39
Number of questions	4	4	4	4	4	4	4	4	4	4	4	4	4
Marks per question (%)	25	25	25	25	25	25	25	25	25	25	25	25	25

All the test items/questions employed in the module assessments were essay-type questions, as mentioned earlier. The data were employed in estimating internal consistency using each of the three methods, viz *Cronbach's alpha*, *split-half* and the *KR 21* formula. The type of scoring and mark allocations used in the module assessments was not dichotomous but rather a continuous variation (polytomous) assigned from 0 to 25 marks per test item. The Likert type scale was used in the Cronbach's alpha calculations. Two different scale widths, the 25-level and 5-level Likert scales were used and their effects assessed. The 5-level Likert scale consisted of marks assigned as 1 = 0 to 5 marks, 2 = 6 to 10 marks, 3 = 11 to 15 marks, 4 = 16 to 20 marks, 5 = 21 to 25 marks.

RESULTS AND DISCUSSION

Cronbach's Alpha

Cronbach's alpha was calculated for the 25-level and 5-level Likert scales. Figure 1a shows the effects of different Likert scale widths. There was a slight decrease in the alpha coefficient for values $\alpha > 0.40$, when the Likert scale width was reduced from 25 to 5. When the alpha values were either negative or lower than 0.40, the 5-level Likert scale gave

higher reliability coefficients than did the 25-level Likert scale. However, the differences in results obtained using the two scale widths, were small and negligible. These results are consistent with the findings of Voss et al who reported that wider scales give greater variance, which increases the alpha value [22]. The values of Cronbach's alpha were found to fall between 0.40 and 0.70 for a majority of the modules - see Figure 1a.

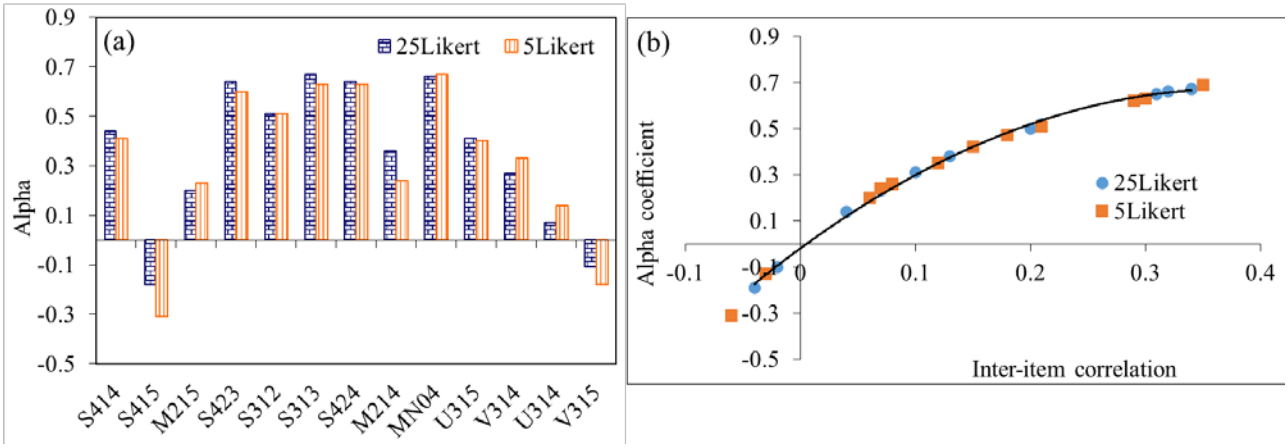

Figure 1: Cronbach's alpha coefficients; a) calculated using different Likert scale levels; b) showing the influence of item interrelatedness.

In the literature, alpha values $0.50 < \alpha < 0.80$ are regarded to be of moderate reliability [7]. Considering that engineering examination assessments are not homogeneous, as discussed earlier, the alpha values for properly constructed test items in these examinations should be low or moderate. As seen in Figure 1a, the modules M215, M214, V314 and U314 gave alpha values that were lower than 0.40, while modules S415 and V315 gave negative alpha values which, in turn, are not meaningful.

Figure 1b shows the inter-item correlation coefficient to be below 0.20 for all modules with an alpha value less than 0.30. These observations do not necessarily suggest that the assessments were not properly constructed but rather they imply that the test items of these modules were less interrelated, which is possible in modules that may include knowledge domains that are not closely linked. Spiliotopoulou indicated that for tests that measure broad constructs of an assessment, the inter-item correlation should be between 0.15 to 0.20, while for narrow or unidimensional constructs, it should lie between 0.4 and 0.5 [17]. In the present study, the positive inter-item correlation varied from 0.1 to 0.3, which underscores the multidimensionality of the data used. It has also been suggested that using low alpha values should be considered acceptable provided it is based on an informed understanding of the data characteristics, rather than applying perfunctory benchmarking, such as mere adoption of $\alpha > 0.7$ [17][21].

Figure 1b also shows that the inter-item relatedness increases non-linearly with an increase in the alpha coefficient. Again, the low or negative alpha coefficient does not imply that these were flawed assessments but rather, it may indicate that these modules had some topics that were largely independent of each other, as previously discussed. It is also evident in Figure 1b that the width of Likert scale used in reliability estimation had no effect on the relation between pairwise inter-item correlation and alpha. Similarly, class size, i.e. the size of test group, had no influence on the alpha coefficient.

Kuder-Richardson's KR 21

The KR 21 coefficient was calculated for each of the modules, using the reliability formula given in Equation (2). Unlike Cronbach's alpha which involves applying the Likert scale for polytomous scoring, the KR 21 does not require the scores for each test item/question to be known, rather the overall mark/score achieved by the test-taker in an assessment is sufficient. KR 21 is the most direct statistic of internal consistency. It uses a simple, basic statistical procedure to generate a reliability coefficient. In the present study, values of KR 21 for each of the modules were found to lie between 0.4 and 0.8, except for one module - M215 - that gave a very high KR 21 > 0.97, a value which appears to be misleading. This module had a small class size of 15 and also gave a poor Cronbach's alpha of 0.23. Clearly, there appear to be some distinct factors that differently influence the two methods. This observation is further seen with modules U314, S415, V315, which had very low or negative alpha coefficients while the corresponding KR 21 values 0.53, 0.58, 0.60 were reasonably good.

Split-half Method

Calculation of the split-half reliability coefficient was done for each module using the PMCC formula given in Equation (3) and the Spearman-Brown formula given in Equation (4). Here too, only the total scores obtained from test items/questions are required in conducting the split-half reliability calculations. Most modules gave an r_{xy} coefficient falling between 0.40 and 0.80. However, low or negative coefficients of 0.09, -0.03, -0.04 were obtained for the

modules S415, U314, V315, respectively. Similar results of the Cronbach's alpha were also obtained for these modules. Based on these results, it may be deduced that the split-half method is sensitive to similar factors as Cronbach's alpha, unlike the KR 21 coefficient, the behaviour of which is quite different. The factors responsible for the different behaviour of KR 21 relative to the alpha and split-half methods are not fully understood and require further investigation.

Comparison of Results

The relationships between results obtained using all three reliability methods are given in Figure 2. Generally, there is good agreement between KR 21 and Cronbach's alpha, but only for $\alpha > 0.3$. It appears that KR 21 is more stable and less affected by various factors, compared to the Cronbach's alpha. Some researches indicate that for normally distributed data, KR 20/21 should be used rather than Cronbach's alpha [17].

The split-half method is strongly correlated with Cronbach's alpha across the full range of values, while KR 21 only correlates with $\alpha > 0.30$ (Figure 2). It is known that Cronbach's alpha is strongly affected by the number of test items/questions. For the civil engineering programmes evaluated, the examinations consisted of four essay-type questions. Interestingly, all the three methods gave meaningful reliability coefficients, despite the few test items used in the examination assessments. The low reliability coefficients obtained for some modules appear to be explained by the very low inter-item relatedness found in the test items.

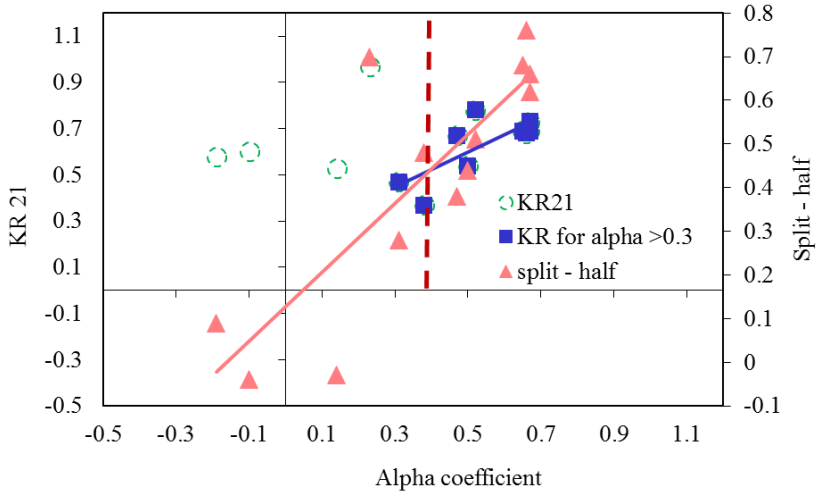

Figure 2: Relationship between KR 21, split-half and Cronbach's alpha.

CONCLUSIONS

The study was conducted to explore the possibility of measuring the reliability of summative examination assessments usually given in the BSc/BEng engineering degree programmes. The internal consistency measurement techniques employed comprised the Cronbach's alpha, KR 21 and split-half methods. Ten-year summative examination data were used in the study.

It was found that despite the heterogeneity and small number of test items in the summative examinations, all three methods exhibited meaningful estimates of reliability coefficient, giving values between 0.4 and 0.8. However, the Cronbach's alpha and split-half methods also gave low or negative coefficients for some modules. The low values obtained are attributed to poor inter-item relatedness of the test items.

The alpha and split-half coefficients were found to be strongly correlated. The KR 21 and alpha had a strong correlation but only for alpha values exceeding 0.30. The relationship between the inter-item correlation and alpha coefficient is non-linear. An inter-item relatedness coefficient exceeding 0.2 was required to give Cronbach's alpha coefficient exceeding 0.40. It appears Cronbach's alpha values exceeding 0.3 indicate an assessment that has balanced and valid test items/questions.

ACKNOWLEDGEMENTS

The work presented in this article was funded by the National Research Foundation (NRF) of South Africa, IPRR Grant No. 96800. The authors are grateful for the financial support given by the NRF.

REFERENCES

1. Hale, C.D. and Astolfi, D., *Measuring Learning and Performance: a Primer*. (3rd Edn), Florida, USA: Saint Leo University, 33574 (2014).

2. Ekolu, S.O., Cronbach's alpha reliability coefficient in engineering assessments - a preliminary study on possibilities and precautions. *Proc. 6th African Engng. Educ. Assoc. Conf.*, CUT, Bloemfontein, Free State, South Africa, 7-11 (2016).
3. Gwet, K.L, *Measures of Association and Item Analysis*. In: Handbook of Inter-Rater Reliability (4th Edn), Gaithersburg, MD, ebook, Advanced Analytics, LLC, (12, 343-65) (2012), 28 March 2016 http://www.agreestat.com/book4/
4. Kuder, G.F. and Richardson, M.W., The theory of the estimation of test reliability. *Psychometrika*, 2, **3**, 151-160 (1937).
5. Tavakol, M. and Dennick R., Making sense of Cronbach's alpha. *Inter. J. of Medical Educ.*, 2, 53-55 (2011).
6. Streiner, D.L., Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. of Personality Assess*., 80, **1**, 99-103 (2003).
7. Tan, Ş., Misuses of KR-20 and Cronbach's alpha reliability coefficients. *Educ. and Sci*., 34, **152**, 101-112 (2009).
8. Graham, J., Congeneric and (essentially) Tau-equivalent estimates of score reliability: what they are and how to use them. *Educational and Psychological Measur.*, 66, **6**, 930-944 (2006).
9. Cortina, J.M., What is coefficient alpha? An examination of theory and applications. *J. of Appl. Psych.* 78, **1**, 98-104 (1993).
10. Allen, K., Reed-Rhoads, T., Terry, R.A., Murphy T.J. and Stone, A.D., Coefficient alpha: an engineer's interpretation of test reliability. *J. of Engng. Educ.*, 97, **1**, 87-94 (2008).
11. Klein, S., Sollereder, P. and Gierl, M., Examining the factor structure and psychometric properties of the test of visual-perceptual skills. *Occup. Therapy J. of Research*, 22, **1**, 16-24 (2002).
12. Lane, A. and Ziviani, J., Assessing children's competence in computer interactions: preliminary reliability and validity of the test of mouse proficiency. *OTJR: Occup., Partic. and Health*, 23, **1**, 18-26 (2003).
13. Panayides, P., Coefficient alpha. *Europe's J. of Psych.*, 9, **4**, 687-696 (2013).
14. Schmitt, N., Uses and abuses of coefficient alpha. *Psych. Assess.*, 8, **4**, 350-353 (1996).
15. Kopalle, P.K., Alpha inflation? The impact of eliminating scale items on Cronbach's Alpha. *Organ. Behav. and Human Dec. Proc.*, 70, **3**, 189-197 (1997).
16. Helms, J.E., Henze, K.T, Sass, T.L. and Mifsud, V.A, Treating Cronbach's alpha reliability coefficients as data in counseling research. *The Couns. Psych*., 34, **5**, 630-660 (2006).
17. Spiliotopoulou, G., Reliability reconsidered: Cronbach's alpha and paediatric assessment in occupational therapy. *Austr. Occup. Ther. J.*, 56, **3**, 150-155 (2009).
18. Sapelli, C. and Illanes G., Class size and teacher effects in higher education. *Economics of Educ. Review*, 52, 19-28 (2016).
19. Koenig, L.B., Gray, M., Lewis, S. and Martin, S., Student preferences for small and large class sizes. *Inter. J. of Hum. and Soc. Sci.*, 5, **1**, 20-29 (2016).
20. Keil, J. and Partell, P.J., The Effect of Class Size on Student Performance and Retention at Binghamton University. Binghamton University, Binghamton, NY, USA (1998).
21. Ekolu, S.O., Correlation between formative and summative assessment results in engineering studies. *Proc. 6th African Engng. Educ. Assoc. Conf.*, CUT, Bloemfontein, Free State, South Africa, 12-16 (2016).
22. Voss, K.E., Stem, D.E., Jr. and Fotopoulos, S., A comment on the relationship between coefficient alpha and scale characteristics. *Marketing Letters*, 11, **2**, 177-191 (2000).

BIOGRAPHIES

Stephen O. Ekolu is an Associate Professor of concrete materials and structures, former Head of the School of Civil Engineering and the Built Environment at the University of Johannesburg, South Africa. He holds an MSc (Eng) with Distinction from the University of Leeds, UK, and a PhD from the University of Toronto, Canada. Prof. Ekolu is a professionally registered engineer and a rated researcher with more than 19 years of academic/industry research experience. His research interests include concrete materials and structures, cementitious materials, durability of concrete, service life modelling, environmental science and engineering education.

Harry Quainoo is a Senior Lecturer of transportation engineering, urban planning and construction management at the University of Johannesburg, South Africa. He obtained his MSc and PhD degrees from the University of the Witwatersrand, and has more than 15 years of teaching and research experience. His areas of expertise are transportation engineering and project management, with interest in engineering education.