Cronbach's Alpha Reliability Coefficient in Engineering Assessments – a Preliminary Study on Possibilities and Precautions

Stephen O. Ekolu

Department of Civil Engineering Science, University of Johannesburg,
Auckland Park 2006, South Africa
E-mail: sekolu@uj.ac.za, sekolu@gmail.com

***Abstract* - This paper attempts to apply the Cronbach's alpha to engineering studies. There is hardly any available literature or research on application of this method to engineering course assessments. Alpha coefficient is commonly used in psychometric tests, as a measure of estimating internal consistency. The data used in this preliminary study consisted of five modules taught over five years by different instructors.**

It is found that alpha reliability coefficient values were generally 0.4 to 0.7 but others gave low or negative alpha values which raises the need for precautions. These preliminary findings highlight an underlying potential regarding estimation of reliability in engineering assessments using alpha coefficient but further research is needed to understand how the values determined, relate to internal structure of the assessments.

***Keywords* – Cronbach's alpha, summative assessment, reliability coefficient, validity, internal consistency**

I. INTRODUCTION

Engineering studies at higher education institutions (HEIs) are carefully constructed to follow highly structured curriculum to progressively provide learning and knowledge over a planned period of study from entry to completion stage of mastery. The full study program is a complex structure typically comprising subsets of knowledge areas covering basic sciences, mathematical sciences, engineering sciences and design while complementary skills such as computing etc. are spread across different stages of learning to provide support skills. These knowledge areas are built within academic modules taught at different stages of the program, with lower level modules typically serving as pre-requisites to higher level modules, implying the increase in module difficulty towards higher levels. The instruction of each module is conducted following a Teaching–Assessment Cycle (TAC), shown in Fig. 1 [1], and done across the semester. According to TAC, the instructor conducts continuous assessment of learning acquired by students during the course of instruction. This requires conduct of formative assessment typically in form of assignments, tests and projects. Formative assessment is intended to inform the instructor of the effectiveness of his/her instructional methods and accordingly adjust, if necessary. More importantly, the results of formative assessment enables students to improve their learning progression.

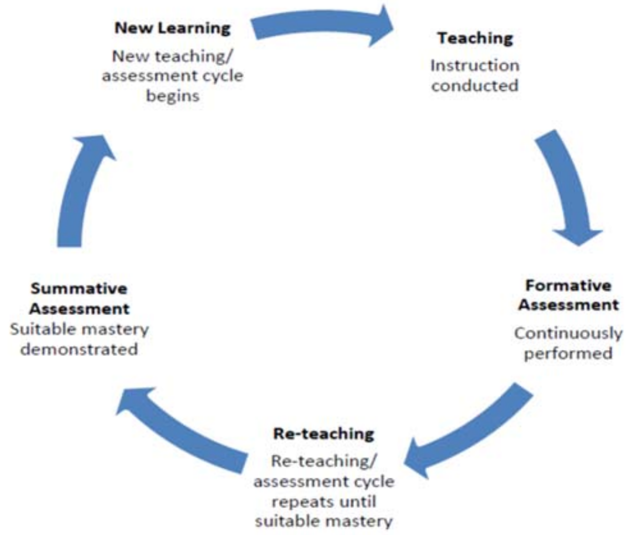
New Learning
New teaching/ assessment cycle begins
Teaching
Instruction conducted
Formative Assessment
Continuously performed
Re-teaching
Re-teaching/ assessment cycle repeats until suitable mastery
Summative Assessment
Suitable mastery demonstrated

Fig.1 Teaching–assessment cycle [1]

The modern instruction methods used at HEIs are informed by *constructivism*, a theory of human learning which advances the concept that humans learn effectively by experiencing and interacting with the elements within their learning environment. In the process, the minds of individuals form new understanding by comparing the newly observed or experienced knowledge with the present understanding s/he has about the subject. This is followed by replacing the present understanding with the new knowledge or rejecting the new understanding in favour of the present knowledge. The use of different teaching practices are intended to align with the constructivist theory, thus, the constructivist class in engineering studies would typically be interactive and student-centred while the instructor conducts moderation. Such engineering classes involve students to actively engage in discussions, projects, field trips, experiments etc. as opposed to the traditional non-constructivist class at which the instructor is authoritative and directs the module instruction rather than moderating the learning process. Non-constructivist instruction may strictly follow the textbook approach with tight adherence to a fixed curriculum [2]-[4].

In engineering study programs, marks obtained from formative assessments typically contribute to the final mark. Accordingly, it is often the case that formative assessment marks obtained by students would influence his/her preparation towards summative assessment. Often, students are expected to achieve a minimum requirement in formative assessment in order to qualify for summative assessment. As lecturing continues alongside formative assessment, the instruction of different knowledge components or topics of the module are scheduled such that students are expected to master the knowledge domain, by the time lecturing for the module is complete. It is this level of mastery which is assessed through summative examinations.

Summative assessments are generally high stakes exams for most students as it strongly contributes to the failure or promotion of a student to the next level of the study program. While weighting of summative assessment is only a proportion of the final mark, typically no less than 50%, it is a final opportunity for the student to progress. Students who have shown weakness during formative assessment aim to raise their academic performance level during the final examination. For most students who finance their studies through loans or sponsorships, failing a final exam may mean loss of sponsorship while the funding expenditure of their studies increases, if a student has to repeat the module. Accordingly, instructors hold the responsibility to ensure that summative assessment of a module has adequate levels of *reliability* and *validity* as measures of the knowledge areas covered during the module instruction. Reliability and validity are two different concepts that refer to 'precision' and 'accuracy'. Considering a bathroom scale, or example, if the correct weight of a person is 70 kg but the scale reading gives 55 kg, each time the measurement is made, then the scale is reliable but inaccurate, implying that the results are invalid [1].

This consideration is important as it ensures that students are not disadvantaged by exams that may be unintentionally skewed towards particular dimensions while neglecting others. In the fields of education and psychology, measurement of reliability is conducted using psychometric tests. Such tests are not commonly encountered in engineering studies and research, apart from perhaps some rare questionnaire type surveys or evaluations. This paper is an exploratory attempt to use reliability measurement approaches that are often employed in psychometric tests, to consider how they may relate to assessment of engineering modules. The study is limited to the Cronbach's alpha coefficient method, subsequently discussed. It is applied to summative assessment results of civil engineering modules followed by interpretation of the results obtained.

II. RELIABILITY

A. Classical test theory

All test measurements contain errors. The Classical test theory, recognizes that each test taker or examinee has a *true* score, upon which measurement error is added to give the observed score or mark. Hence the classical test theory [5-6], can be written as

$$X_i = T_i + E_i \qquad (1)$$

Where X_i is the score/result obtained by a particular test taker i, during a given test measurement or exam event. T_i is the test takers theoretical 'true' score, and E_i is an error responsible for the difference between X_i and T_i. According to the Classical test theory, reliability is defined as the ratio of variance of true score/mark to variance of observed score /marks and is conveniently written as given in equation 2

$$r_x = \frac{\sigma^2{}_T}{\sigma^2{}_X} = 1 - \frac{\sigma^2{}_E}{\sigma^2{}_X} \qquad (2)$$

Where r_x is reliability of the observed score/result, and σ_X^2, σ_T^2, σ_E^2, are the variance of observed score/result, variance of true score/result, and variance of error respectively. In practice, however, the true score is unknown, making it impossible to theoretically calculate reliability. For this reason, reliability is estimated using test measurements.

B. Internal consistency measurement

Cronbach's alpha is one of the most frequently used methods of estimating internal consistency reliability. The alpha formula is written as:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_i{}^2}{\sigma_T{}^2}\right) \qquad (3)$$

Where, N is the number of test items or questions, σ_i^2 is the variance for each test item, and σ_T^2 is the total variance. Cronbach's alpha can also be written in an alternative standardized expression as:

$$\alpha = \frac{N.\bar{C}}{(\bar{\nu} + (N-1).\bar{C})} \qquad (4)$$

Where, N is the number of test items or questions, $\bar{\nu}$ is the average of all variances of the test items, and $\bar{C}$ is the average of all covariances between the paired test items. Equation (4) resolves the problem of test items measured using different units [7]. The Cronbach's alpha method can be used in tests for both dichotomously (non-continuously) and polytomously (continuously) scored items, the former being 'agree /disagree', 'right/wrong', 'correct /incorrect' type of responses, while in the latter, responses entail ascendency/descendency in agreement i.e. attitude scale such as 'agree', 'strongly agree', 'disagree', 'strongly disagree'. In applying Cronbach's alpha to dichotomous tests, the responses are assigned binary (0,1) numerals for computation. For polytomous responses, the level of Likert scale is defined based on the number of items, such as a five level scale: 1 = 'strongly disagree', 2='disagree', 3=undecided, 4='agree', 5='strongly agree' [7].

III. ASSESSMENTS IN ENGINEERING STUDIES

In considering the potential application of internal consistency reliability measurement in summative assessments, it is crucial to take into account the nature and structure of assessments in engineering studies, and how they relate or conflict with the assumptions employed in reliability methods, specifically the Cronbach's alpha coefficient used in this study. The critical concepts requiring detailed consideration in structuring of summative assessments in engineering studies are the *inter-relatedness of test items/questions*, *dimensionality*, *and homogeneity of the test/assessment* as a whole. In psychometric tests, the primary purpose of internal consistency test is to measure the association that exists between different test items or questions, such that each item contributes significantly to the same knowledge domain or base being assessed. As earlier mentioned, engineering program studies involve learning the diverse knowledge areas of basic sciences, mathematical sciences, engineering sciences, design and synthesis. These knowledge areas are usually compartmentalized as components of different modules taught over the duration of a four year bachelor's undergraduate degree. For example, a structural design module would involve knowledge areas of:- basic sciences (materials, physics, chemistry), engineering sciences (structural analysis, physics, maths), design (conception, imagination, creativity, use of code procedures, maths). While the primary construct is engineering design mastery, it cannot be attained without the knowledge areas that are pre-requisite to its mastery. Accordingly, it is necessary for the assessment of a module to evaluate different knowledge areas as part of the design knowledge domain, in order to ensure validity or representatives of the course content. These knowledge areas that are embedded within the structural design knowledge domain can be considered as sub-tests.

In summative assessments, the test items/questions are selected randomly across the full set of module topics and knowledge areas. It should also be noted that module topics are presented as building blocks, so that knowledge areas covered in earlier topics would be needed in the topics presented at later stages towards knowledge domain mastery. This example illustrates how highly structured the typical engineering modules can be, while ensuring inter-relatedness as a critical requirement. As mentioned in the foregoing, summative assessments have to be *valid* and have to accordingly, involve multiple outcomes. As a result, most test items are essay type questions, each of them covering different concepts or knowledge areas and would normally be divided into sections. For example, one question on beam design can examine basic science (materials), engineering science (structural analysis, maths), and design (conception, technical procedure, maths). Another question on column design would examine similar or different knowledge principles from the beam question but this time, the knowledge area is applied to column design. Therefore, inter-relatedness between the test items can be expected with respect to knowledge areas. But because these knowledge areas are diverse, it is possible that different test items may be used to measure different skills, which renders the test assessment to be heterogeneous both in the type of knowledge areas assessed and in the level of item difficulty. It can be appreciated that examiners in engineering assessments use test items of different difficulties across the assessment but may or may not maintain the same score /mark allocation for each test item. Also associated with item difficulty is the length of test item/question. A summative assessment may use questions of different lengths and that require different time periods to complete. Accordingly, different marks may be allocated to different test questions based on their difficulty and time required to complete the item. This too brings heterogeneity into the assessment. For most engineering modules, however, it is common practice to try as much as possible to provide test questions of same mark allocation. It is also common to introduce variations of balancing out the presence of difficult test questions by including a relatively easier question(s) so as to give the test taker (student) a comprehensive assessment overall.

Some engineering modules are structured to cover two or three different knowledge domains. For example, a module on strength of materials may be divided into two parts presenting material science and mechanics. In such modules, the test questions for material science can be completely unrelated to those in mechanics, making the assessment inherently heterogeneous. In assessing such modules, it is common to divide the assessment paper into sections, each section covering test items of a different knowledge domain e.g. material science, mechanics domain, amongst others.

IV. PRELIMINARY STUDY

A preliminary study was conducted using results from summative assessments of BEng/BSc degree in engineering. Data were taken from five modules of civil engineering study program. The modules designated as S414 and S415 was a structural engineering course that was offered at different academic years for fourth-year students, M215 was a strength of materials module for second-year students, S423 was third-year civil engineering theory course, and S312 was construction materials course for fourth-year students.

The class sizes for each module varied from 56 to 79 students, except one module M215 which had 15 students. This range of classes generally fall within the category of small to medium size classes [8]. There is no strictly standardized grouping of class sizes, so various researchers typically apply different ranges of class size groupings in their studies [8]-[10]. For purposes of this study, class sizes with student numbers under 20 = small, 20 to 90 = medium, over 90 = large [8].

Summative assessment marks from final exams were used in this investigation. Heterogeneity of the class groups is evident in their assessment results, a sample of which are shown in Fig. 2 for modules S415, S423, S423. The results show normal distribution behaviour, which represents the typical characteristics of a properly composed group. It is also seen that the average performance of the class groups lies between 40 to 65%, depending on the module. Similar observations are exhibited by the other modules, M215 and S312.

In this study, alpha reliability coefficient was determined for each module of the same test length i.e. same number of questions but the test items would not be of the same level of difficulty. All assessments /exams consisted of four essay type items or questions, each being worth 25 marks. It should also be mentioned that all questions were compulsory.

V. RESULTS AND DISCUSSION

In this study, Cronbach's alpha was estimated using the formulae given in equations (3) and (4), for both the 25-Likert and 5-Likert scales. The 5-level Likert scale was defined 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.

Table 1 and Fig. 3 show the effects of Likert scale level and the two Cronbach's alpha formulae on reliability estimations. When the different equations (3) and (4) were used, it is seen that for all positive alpha coefficient, equation (4) consistently gives reliability values that are similar or slightly higher than those determined using equation (3). But the difference is small, occurring at 1/100th of magnitude and is negligible. Similarly, changing the Likert scale level from 25 to 5, generally gave a small decrease in alpha coefficient. The reason for this behavior is not clear, however, the difference is small and negligible, occurring at 1/100th of magnitude.

As seen in Fig. 3, a majority of the modules gave Cronbach's alpha falling between 0.40 to 0.70 which is consistent with interpretation of $0.50 < \alpha < 0.80$ as moderate reliability [11]. In engineering studies, high reliability is not desirable as it depletes course content and diminishes validity of the assessment. Some modules, however, gave alpha coefficient that is lower than 0.40 and even negative values. As given in Table 1, Module S415 gave negative alpha coefficients ranging from -0.04 to -0.31, while the coefficients for M215 were positive but also generally low. S423 gave the highest coefficients with a value of 0.66.

While the negative alpha coefficients do not make sense, it is observed that all modules that gave $\alpha < 0.30$ had low inter-item correlation coefficients of less than 0.20, which indicates that these modules had test items that had very small inter-relatedness. Modules with low or negative alpha coefficient do not necessarily imply flawed assessments but could mean that these modules had some items or topics that were completely independent of others, as discussed earlier.

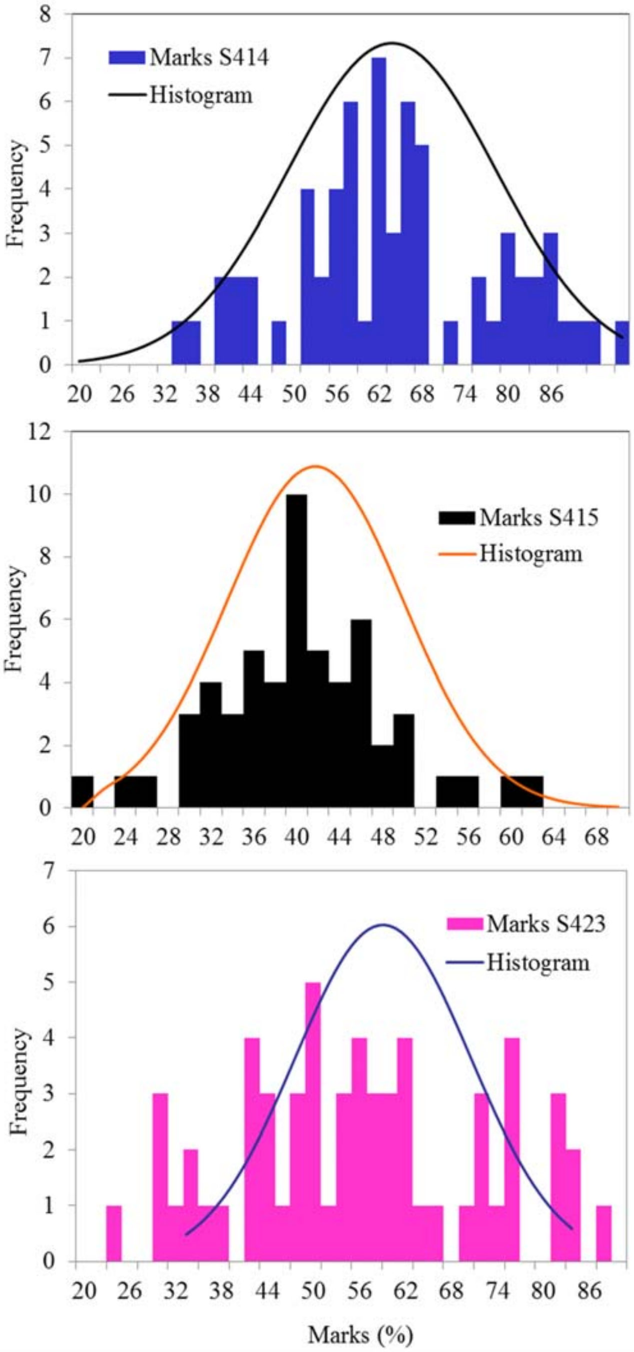

Fig. 2 Summative assessment marks for engineering modules S414, 415, 423.

Accordingly, there is need for precaution in interpreting the alpha coefficients, when applied to comprehensive test items. Also, alpha coefficients usually apply to measurements involving a large number of test items, suggested to be at least 20 direct questions. It is interesting to note, however, that the method shows robustness by giving good response to a small number but comprehensive test items typically used in engineering assessments.

TABLE 1
ALPHA COEFFICIENT CALCULATED USING DIFFERENT FORMULAE AND LIKERT SCALES

Module	Class size	alpha, 25-Level Likert			alpha, 5-Level Likert		
		Eqn 3	Eqn 4		Eqn 3	Eqn 4	
		α-coeff	Item corr	α-coeff	α-coeff	Item corr	α-coeff
S414	65	0.44	0.18	0.47	0.41	0.15	0.42
S415	56	-0.18	-0.04	-0.19	-0.31	-0.06	-0.31
M215	15	0.2	0.07	0.23	0.23	0.08	0.26
S423	60	0.64	0.32	0.66	0.6	0.29	0.62
S312	79	0.51	0.21	0.52	0.51	0.21	0.51

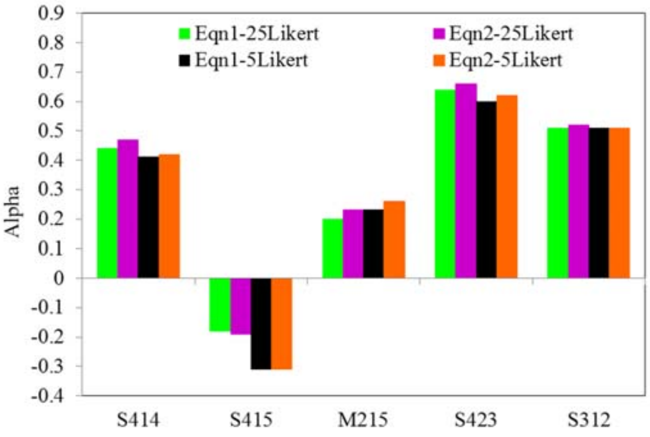

Fig. 3 Comparison of alpha coefficient calculated for 25 and 5-Likert scales

VI. CONCLUSION

A preliminary study was conducted to explore the possibility of using Cronbach's alpha to estimate reliability of summative assessments in engineering bachelors degree programs. It was found that despite the heterogeneity and small number of test items in engineering modules, the alpha coefficient, gave estimation of reliability coefficients to be between 0.4 to 0.7 but it also gave low or negative coefficients for some modules. The factors responsible for the low /negative alpha in some modules are not clear but appears to be associated with the inter-item relatedness.

The Cronbach's alpha method also shows robustness demonstrated by its good response to a small number of test items which are comprehensive questions, the type commonly used in engineering assessments or exams. Further research is needed to understand how alpha coefficients relate to the internal structure of comprehensive test items/questions.

ACKNOWLEDGEMENTS

The study presented in this paper was funded by the National Research Foundation (NRF) of South Africa, IPRR Grant No. 96800. The author is grateful for the financial support given by NRF.

REFERENCES

[1] C.D Hale and D. Astolfi, *Measuring learning and performance: a primer*, 3rd Ed., 2014, Saint Leo University, P. O. Box 6664, St. Leo, Florida 33574. Ebook: http://charlesdennishale.org/

[2] J. C. Le Coze, "Towards a constructivist program in safety", *Safety Science*, Vol. 50, Issue 9, November 2012, 1873-1887

[3] S. Cristea, "The fundaments of constructivist pedagogy", *Procedia - Social and Behavioral Sciences*, Vol. 180, 5 May 2015, 759-764.

[4] İ.Tuncel and A. Bahtiyar, "A case study on constructivist learning Environment in Content Knowledge Courses in Science Teaching", *Procedia - Social and Behavioral Sciences*, Vol. 174, 12 February 2015, 3178-3185.

[5] M. Meadows and L. Billington, *A review of the literature on marking reliability*, May 2005, National Assessment Agency, 89p.

[6] B.W Junker, *Some aspects of classical reliability theory and classical test theory*, Department of Statistics, Carnegie Mellon University, Pittsburgh PA 15213, 1 March 2012, 16p.

[7] K. L. Gwet, *Measures of association and item analysis* (chapter 12, p343-65), In Handbook of Inter-Rater Reliability (4th Edition), 2012, ebook, Advanced Analytics, LLC, PO Box 2696, Gaithersburg, MD 20886-2696. http://www.agreestat.com/book4/. Accessed 28 March 2016.

[8] C. Sapelli, and G. Illanes, "Class size and teacher effects in higher education", *Economics of Education Review*, 52, 2016, 19–28.

[9] L.B. Koenig, M. Gray, S. Lewis and S. Martin, "Student preferences for small and large class sizes", *International Journal of Humanities and Social Science*, Vol. 5, No. 1; January 2015, 20-29.

[10] Jack Keil and Peter J. Partell, *The effect of class size on student performance and retention at Binghamton University*, Binghamton University, PO Box 6000, Binghamton, NY 13902-6000.

[11] S. Tan, "Misuses of KR-20 and Cronbach's Alpha Reliability Coefficients", *Education and Science*, 2009, Vol. 34, No. 152, 101-112.