TECHNOLOGY ENGINEERING ONLINE LEARNER METRICS TO ANALYZE INSTRUCTIONAL EFFICACY

By

JAMES EDWARD OSLER * MAHMUD A. MANSARAY **

* North Carolina Central University
** North Carolina Central University

ABSTRACT

The online deployment of Technology Engineered online Student Ratings of Instruction [SRIs] by colleges and universities in the United States has dynamically changed the deployment of course evaluation. This research investigation is the fourth part of a post hoc study that analytically and psychometrically examines the design, reliability, and validity of the items used in the delivery of an SRI instrument online during the 2012 Spring Semester at a United States Historical Black College and University [HBCU]. The sample under analysis consisted of 7,919 distributed Student Ratings Instruments with a Grand Total of 56451 analyzed instrument items. The application of several statistical measures was used to validate the Technologically Engineered SRI instrument used in the study. Results of the study included: a Cronbach's Alpha Test that yielded an extraordinarily high internal consistency of the student rejoinders to the items on the rating scale; and a Tri–Squared Test supported the research findings attributed to the Cronbach's Alpha Test results. This research is the continuation of in–depth investigations into Technology Engineering as discussed in i-manager's Journal on School Educational Technology, further explored in i-manager's Journal of Educational Technology, and data analyzed with the Tri–Squared statistical measure first introduced in i-manager's Journal on Mathematics.

Keywords: Advanced Statistics, Construct Validity, Cronbach's Alpha, Efficacy, Goodman & Kruskal's Lambda, Principal Component Factor Analysis, Reliability, Technology Engineering, Statistics, Student Ratings of Instruction [SRI], Tri–Squared Test, and Validity.

INTRODUCTION

Rationale

This study is a continuation of a research investigation into the practical importance of determining teaching effectiveness in terms of student learning outcomes in colleges and universities (particularly in the United States [US]). This research study used an online psychometric tool to Technologically Engineer a Student Ratings of Instruction Instrument which was deployed to determine student perceptions of faculty teaching. The determination of teaching excellence still appears to be exigent for researchers (Roche & Marsh, 2000; Young & Shaw, 1999). There is hardly any measurement, attribute, conduct, or classroom strategies, which completely grasps the meaning of an effective teacher. Barry (2010), for example, noted teaching effectiveness "involves a deep understanding of subject matter, learning theory and student differences, planning, classroom instructional strategies, knowing individual students, and assessment of student understanding and proficiency with learning outcomes" (pp. 4). Hassel (2009) also said the underlying specification of teacher effectiveness must be the student learning outcomes; that is, the extent to which students learn, and more valued effects. Vogt (1984), on the other hand, associated effective teaching to the potential to provide tutoring to diverse students of disparate talents while integrating teaching goals and evaluating the existing knowledge aspect of the students.

Technology Engineering ™ ©

"Technology Engineering" is the combination of Learner–Based Tools, Educational Games, Educational Systems, Relevance, and Collaborative Learning Strategies to create an interactive and dynamic cognitive economy. Technology Engineering leads to Product–Based Inquiry. An example of "Technology Engineering" as a methodology is the combination of innovative technology tools such as Instructor–Authored Interactive Metametric Learning

Modules with effective teaching strategies and dynamic distance education tools such as Course Management and Learning Management Systems (Osler, Hollowell, and Nichols, 2012). Technology Engineering is an innovative form of interactive technology used as a method of teaching science and math content and curriculum through collaboration with teachers in the classroom (Osler, Hollowell, and Palmer, 2008). In the SRI study, Technology Engineering is applied through the design and delivery of the SRI instrument online to the research participants.

Determining Technology Engineered Instructional Efficacy

Establishing "Instructional Efficacy" or "Teaching Effectiveness" is the major emphasis in higher education. Determining teacher effectiveness may not be uncommon to academics since various studies on teaching excellence and student learning outcomes have already been completed (Hunsaker, Nielsen, & Bartlett, 2010; Perry as cited in Keeley, Furr, & Buskist, 2010; Schrodt, Witt, Myers, Turman., Barton, & Jernberg, 2008). After all, a practice that is prevalent in many universities and colleges in the United States and elsewhere is the use of student ratings to evaluate teaching effectiveness and student success (Cohen as cited in Donnon, Delver, & Beran, 2010; Kneipp, Kelly, Biscoe, & Richard, 2010; Stowell, Addison, & Smith, 2012). Marsh and Roche (as cited in Zhao & Gallant, 2012) noted that, the use of student ratings of faculty in colleges and universities is the provision of a constructive comment to faculty for teaching enhancement, as well as a cursory evaluation of teaching effectiveness for personnel or administrative decisions. Thus, this study continues to further investigate the efficacy of the SRI instrument in determining teaching (instructional) effectiveness through an in–depth investigation of item analysis. In addition, the study will examine whether ratings completed by student ratings guarantee the validity and reliability of the instrument, in measuring what it is designed to measure: teaching effectiveness. In addition, the study will also assess for internal consistency as a means of establishing the validity and reliability of the instrument used for rating the effectiveness of the courses and programs. The study made use of existing student data from research previously conducted during the 2012 academic year at a HBCU (including the use of a quantitative approach to determine the impact of the evaluation model on teaching excellence and student success).

Supportive Research for the Validity and Reliability of SRIs that can be Technology Engineered

There is a growing list of literature on the uses, validity and reliability of the student ratings of faculty. Several journals on the evaluation of university faculty on teaching effectiveness and student success make use of the student ratings of instruction survey (Clayson, 2009; Jones, 2010; Zhao et al., 2012). The faculty uses student ratings to obtain a student response about their courses and register growth in their instruction parts and accountabilities (Donnon et al., 2010), which may have a significant effect on their professions (Sprinkle, 2008). Aleamoni (as cited in Zhao et al., 2012) had suggested the use of student ratings because students can offer information on the attainment of critical educational goals; affinity with the instructor; and, fundamentals of a classroom, such as instructional supplies, assignment, and instructional procedures. The Student ratings are also used to express knowledge to the students and to establish administrative resolutions, such as giving life–term tenure and advancement (Marsh, 2007; McKeachie, 2007). In addition, studies have revealed that students incline to consider teaching evaluations earnestly, and are enthusiastic to contribute and offer expressive response when they consider and can realize that their contributions are being reflected and assimilated by their teachers and the university (Agbetsiafa, 2010).

Notwithstanding, the growing list of journal articles on the usage of student ratings does not preclude the controversy surrounding their reliability, validity, generalizability, and their assessment of university teaching (Marsh, 2007). In general, the reliability of student ratings is concerned with the consistency [internal], stability and dependability of the assessment over time (Chen & Watkins, 2010; Zhao et al., 2012). McMillan (as cited in Zhao et al., 2012) also noted a dependable resolution is one that has comparable performance at diverse times. Despite the controversy surrounding the reliability and validity of student ratings of instructors, several studies have found that student ratings of instructor are reliable, steady across items, raters, and session (Beran, Violato, Kline, & Frideres, 2009; Beran et al., 2009; Kneipp et al., 2010; Zerbini & Abbad, 2009; Spooren, P. P., Mortelmans, D. D., & Denekens, J. J., 2007).

Several research studies have applied different statistical models to determine the reliability of the rating tools used to evaluate faculty. Most of these tests centered on the internal consistency [stability] of the ratings. The more prevalent is the Cronbach's Alpha statistics, with the alpha varying from 0 to 1; the 1 being the highest reliability score. Beran et al. (2009), for example, conducted a study to determine what students find useful about student ratings. With the application of survey responses from (N = 1229) students at a leading Canadian university, the authors developed a psychometrically substantial measure of the usefulness of student ratings. Using the Cronbach's Alpha model, the results established the internal consistency reliability of the 16 items on the rating scale at .93, thus indicating a high level of internal consistence for the student ratings. Kneipp et al. (2010) also reported an inter–rater reliability estimation of above .91 in the Cronbach's Alpha model, in their study, to determine the effect of the instructor's characteristics on the quality of instruction. In addition, Zuberi, Bordage, and Norman (2007) also noted the internal consistency [reliability] for student ratings obtained in their analysis of student evaluation of teaching in outpatient clinics was .98. Donnon et al. (2010) similarly realized a high internal consistency with a Cronbach's Alpha coefficient of .93 in their research on student ratings of instructors in medical sciences graduate programs. This only serves to highlight the high internal consistency of the student ratings in evaluating instructors.

Apart from the dispute surrounding their reliability, the student ratings are also faced with the problem of validity. In all–purpose, validity of student ratings denotes the degree to which student assessments of faculty instruction essentially evaluate what they are planned to evaluate (Zhao et al., 2012). Agbetsiafa (2010) also noted if student ratings are to be used to evaluate teaching effectiveness and student learning outcomes, the tools must be exposed to challenging validity trials and exploration.

So much has been said about the various validity studies of the student ratings of faculty. However, an often overlooked, but significant theoretical framework, applied in the validity analysis of student ratings of teaching excellence is the generalizability theory. Shavelson and Webb (as cited in Bo, Johnson, & Kilic, 2008) noted generalizability theory (G theory) describes the defect in conventional test theory in which the error term is an indistinct combination of different error bases. Bo et al, (2008) also said the analysis of variance (ANOVA) methods can be used to obtain estimates in the generalizability theory. Zuberi et al. (2007), for example, employed a repeated measure generalizability device to determine the validation of the student rating instrument in the evaluation of teaching in an outpatient clinic. Using the analysis of variance (ANOVA) and variance components, the results established a reliability score of .98 and a validity score of about 57%, indicating high reliability and validity of the student rating instrument used to evaluate teaching in outpatient clinics.

Research Questions

The research questions listed below were developed to examine the validity and reliability of the Student Ratings of Instruction [SRI] instrument used in the study to evaluate teaching quality.

Q_1: Do ratings completed by students engender internal reliability [consistency] in their measurement of teaching effectiveness?

This question calls for a quantitative research design. The ratings from the survey data at HBCU were used to determine the reliability of the question. This is specified in hypothesis H_{10} and H_{1a}. The statistical tool used was Cronbach's Alpha to verify the reliability [consistency] of the instrument used to evaluate teaching effectiveness.

Q_2: Do ratings completed by student ratings produce augmented validity in measuring teaching effectiveness?

The question aforementioned question calls for a quantitative research design. Again, the research made use of existing data at HBCU to determine this question. This is again outlined in hypothesis H_{20} and H_{2a}. The study will apply the factor analysis to determine the validity [construct validity] of the instrument used to determine teaching effectiveness.

Hypotheses

The below hypotheses were used to assess the research questions one and two. Each research question addresses a null hypothesis with anticipation of a non–significant association, and an alternative hypothesis that suggests that a significant association does occur between the variables.

H_{10}: The student ratings do not increase the reliability of the instrument used to assess teaching effectiveness.

H_{1a}: The student ratings significantly increase the reliability of the instrument used to assess teaching effectiveness.

H_{20}: The ratings completed by students do not create any validity of the rating instrument used in evaluating teaching effectiveness.

H_{2a}: The ratings completed by students generate increased validity of the rating instrument used in evaluating teaching effectiveness.

Tri–Squared Test Mathematical Hypotheses

The first sets of Mathematical Hypotheses used in the study in terms of Tri–Squared to determine Technology Engineered SRI item efficacy, validity, and reliability were as follows:

H_0: $Tri^2 = 0$

H_1: $Tri^2 \neq 0$

Cronbach's Alpha [α] Mathematical Hypotheses

The second sets of Mathematical Hypotheses used in the study in terms of Cronbach's Alpha to determine reliability were as follows:

H_{10}: $\alpha \leq 0$

H_{1a}: $\alpha > 0$

The second sets of Mathematical Hypotheses used in the study in terms of Cronbach's Alpha to determine validity were as follows:

H_{20}: $\alpha \leq 0$

H_{2a}: $\alpha > 0$

Statistical Mathematical Models

Tri – Squared Test [Tri^2]

Tri–Squared comprehensively stands for "The Total Transformative Trichotomous–Squared Test" (or "Trichotomy –Squared"). The Total Transformative Trichotomous–Squared Test provides a methodology for the transformation of the outcomes from qualitative research into measurable quantitative values that are used to test the validity of hypotheses. It is based on the mathematical "Law of Trichotomy". The Total Transformative Trichotomous –Squared Test provides a methodology for the transformation of the outcomes from qualitative research into measurable quantitative values that are used to test the validity of hypotheses. The advantage of this research procedure is that it is a comprehensive holistic testing methodology that is designed to be static way of holistically measuring categorical variables directly applicable to educational and social behavioral environments where the established methods of pure experimental designs are easily violated. The unchanging base of the Tri–Squared Test is the 3 × 3 Table based on Trichotomous Categorical Variables and Trichotomous Outcome Variables. The emphasis the three distinctive variables provide a thorough rigorous robustness to the test that yields enough outcomes to determine if differences truly exist in the environment in which the research takes place (Osler, 2013).

Tri–Squared is grounded in the combination of the application of the research two mathematical pioneers and the author's research in the basic two dimensional foundational approaches that ground further explorations into a three dimensional Instructional Design. The aforementioned research includes the original dissertation of optical pioneer Ernst Abbe who derived the distribution that would later become known as the chi square distribution and the original research of mathematician Auguste Bravais who pioneered the initial mathematical formula for correlation in his research on observational errors. The Tri–Squared research procedure uses an innovative series of mathematical formulae that do the following as a comprehensive whole: (i) Convert qualitative data into quantitative data; (ii) Analyze inputted trichotomous qualitative outcomes; (iii) Transform inputted trichotomous qualitative outcomes into outputted quantitative outcomes; and (iv) Create a standalone distribution for the analysis possible outcomes and to establish an effective-research effect size and sample size with an associated alpha level to test the validity of an established research hypothesis. Osler (2012) defined Tri–Squared as:

$$Tri^2 = T_{Sum}\left[\left(Tri_x - Tri_y\right)^2 : Tri_y\right]$$

Cronbach's Alpha [α]

One of the significant statistical models in this research is the Cronbach's Alpha [α]. It is a valuable coefficient for examining the internal consistency and has been named after Lee Cronbach who first developed it in 1951. Bland and Altam (1997) defined Cronbach's Alpha as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} s_i^2}{s_T^2}\right)$$

Where, k is the amount of objects, s_i^2 is the variance of the i^{th} object and ith object and S_T^2 is the variance of the final total created by adding all the objects. In addition, they also said if the objects were not simply added to make the total, but were initially multiplied by weighting coefficients, then the object must be multiplied by its coefficient ahead of the analysis of the variance s_i^2. Certainly, the formula must contain at least two objects, that is, $K<1$ or α cannot be distinct. Field (2009) also defined the Cronbach's Alpha somewhat differently from that stated by Bland et al. (1997), even when the ideas are similar. Field defined the Cronbach's Alpha as:

$$\frac{N^2 Cov}{\sum s_{item}^2 + \sum Cov_{item}}$$

The author noted that, for every object on the scale, two things can be computed: the variance contained in the object, and the covariance amongst an explicit object and any additional object on the scale. Thus, a variance–covariance matrix of the whole objects can be computed. In addition, the author also said, in the matrix, the diagonal rudiments establish the variance contained in an exact object, and the off–diagonal rudiments comprise covariances amid sets of objects. The upper half of the formula is the quantity of objects (N) squared multiplied by the mean covariance amongst objects. The lower half is only the total of all the object variances and object covariances. The arrays of the alpha statistic are between zero and one. The greater the coefficient, the better the select items organized together in evaluating the instrument construct, and thus the better the statistical reliability of the assessment tool. An alpha of 1.00 would imply a seamlessly consistent instrument, while a coefficient of zero would imply an untrustworthy tool.

Factorial Analysis

The factorial model used in this study is derived from Agbetsiafa (2010), and Field (2009). Concisely, factor analysis allows the delineation of an essential or hidden configuration in a data set. It accelerates the analysis of the configuration of the associations (correlation) among an outsized number of variables by describing a set of shared essential measurements, commonly termed factors (Agbetsiafa, 2010). Field (2009) noted that, factorial is a mathematical model, resembling a linear equation but without the intercept because the lines intersect at zero and, therefore, the intercept is also zero. Field (2009) defined factorial as:

$$Y_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni} + I_i$$

The values of b are the loading factors.

Agbetsiafa's (2010) was more detailed in his description of the factorial model in his research than Field (2009). According to Agbetsiafa, it is conceivable to reorient the data to allow the first small number of measurements to explain for much of the existing data. Assuming there is any idling in the data set; it is also conceivable to explain for most of the evidence in the original data with a significantly condensed amount of measurements. Adapting his template, this study also assumes that the 15 items on the student evaluation survey instrument bears relationships with a series of functions working linearly, and they may be represented by the following mathematical formulas:

$$Y_1 = \alpha_{10+}\, \alpha_{11}X_1 + \ldots + \alpha_{1n}X_n + \alpha_i$$

$$Y_2 = \alpha_{20+}\, \alpha_{21}X_1 + \ldots + \alpha_{2n}X_n + \alpha_i$$

$$Y_3 = \alpha_{30+}\, \alpha_{31}X_1 + \ldots + \alpha_{3n}X_n + \alpha_i$$

...

$$Y_{15} = \alpha_{150+}\, \alpha_{151}X_1 + \ldots + \alpha_{15n}X_n + \alpha_i$$

Where: Y = a variable with recognized data; α = constant; X_i = the fundamental factors; and, ε_i = the error terms, which help to point out the conjectured associations, are not exhaustive. Thus, applying the technique to the recognized 15 items on the student rating survey instrument, factor analysis describes the unidentified X utilities. The developing loadings from the analysis are the constants, and the factors are the X utilities. The scope of the individual loading for every utility assesses the degree to which the definite utility is associated with the explicit variable (Y). Thus, for any of the 15 variables in equation one of the proposed study, the model may be written as: $Y_1 = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots + \alpha_n X_n + \varepsilon_i$, where X_is denote factors, and α_is signify the loadings.

Conducting Technology Engineered SRI Research at an HBCU

The backdrop for this research study is the Research, Evaluation, and Planning Department (REP) of Historic Black College and University. The study will make use of the student rating responses conducted during the spring semester of 2012 by REP. This is a cross–sectional data set. Arguably, REP has the responsibility to coordinate all student, faculty and administrative surveys on behalf of the university. 7,919 students responded to the spring semester survey. The students were graduates and undergraduates and were from the various departments, schools and colleges within the university, with the exclusion of the Law School. The ratings from the Law School have different components, which are incompatible with the ratings of the rest of the schools and colleges in the university.

The Technology Engineered SRI Measurement Scale

The student ratings of instruction survey is employed to evaluate course instructors and is administered online during the spring and fall semesters of each academic year, with the CourEval Assessment Tool in a 5–point Likert scale, to all registered students of the university. The rating survey requires students to assess their instructors on 15 items in the assessment tool. The instrument has two subscales. Items 1 to 3 measure the student's efforts in the course, where the scale comprises the following: 1 = never, 2 = not much of the time, 3 = about half of the time, 4 = most of the time and, 5 = all of the time. Items 4 to 15 evaluate the instructor, where the scale comprises the following: 1 = strongly disagree, 2 = disagree, 3 = no opinion, 4 = agree and, 5 = strongly agree. This research considers the evaluation of the instructor as an assessment of teaching effectiveness or teaching quality. A sample of the assessment tool is presented in Appendix C of page 49. Also, the instrument has a section where students can make open–ended statements about the instructors, when these are requested by the individual colleges or departments within the university. For the 2012 spring semester ratings survey, the 7,919 responders evaluated instructors on 18,817 courses and course sections offered at HBCU. In addition, the only variables included in this study are the 15 items on the survey instrument with their respective ratings.

Results

Hypothesis 1

The Cronbach's Alpha was computed to test hypothesis one on whether the ratings completed by the students, to evaluate faculty on the 15 items, significantly increase the reliability of the rating scale. The results of the study are revealed in Tables 1 –4. Table 1 is the Reliability Statistics for the Test of Internal Consistency. In general, the unstandardized alpha is usually used to interpret the internal consistency of the rating scale. In this regard, the Cronbach's Alpha of .954 was close to unity, and was significantly higher for the 15 items on the rating scale. Table 2 descriptive statistics for each item on the rating scale. Item one [A1] (M = 4.57, SD = .670) on the rating scale, for example, required to state the level of effort a student had exerted on the course. Also, Table 3 matrix unveiling the inter–item correlations of each item on the scale with every other item. According to the Table, 2 (The

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.954	.949	15

Table 1. Reliability Statistics

	Mean	Std. Deviation	N
A1	4.57	.670	18817
A2	4.58	.606	18817
A3	4.51	.729	18817
B1	4.37	.884	18817
B2	4.23	1.011	18817
B3	4.27	.997	18817
B4	4.28	1.005	18817
B5	4.20	1.009	18817
B6	4.19	1.013	18817
B7	4.26	.975	18817
B8	4.27	.938	18817
B9	4.21	1.000	18817
B10	4.29	.979	18817
B11	4.29	.962	18817
B12	4.30	.967	18817

Table 2. Item Statistics

	A1	A2	A3	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
A1	–	.496	.670	.295	.258	.264	.247	.332	.290	.257	.239	.262	.238	.235	.260
A2	.496	–	.458	.209	.170	.176	.169	.204	.184	.175	.168	.173	.163	.158	.165
A3	.670	.458	–	.260	.234	.235	.211	.277	.269	.217	.219	.242	.206	.213	.219
B1	.295	.209	.260	–	.799	.817	.687	.718	.744	.705	.752	.699	.693	.683	.743
B2	.258	.170	.234	.799	–	.841	.702	.721	.746	.711	.742	.699	.695	.694	.748
B3	.264	.176	.235	.817	.841	–	.750	.745	.756	.743	.742	.689	.728	.717	.794
B4	.247	.169	.211	.687	.702	.750	–	.741	.698	.730	.665	.637	.758	.714	.799
B5	.332	.204	.277	.718	.721	.745	.741	–	.771	.722	.696	.676	.717	.694	.770
B6	.290	.184	.269	.744	.746	.756	.698	.771		.736	.792	.752	.708	.689	.745
B7	.257	.175	.217	.705	.711	.743	.730	.722	.736	–	.731	.699	.707	.689	.762
B8	.239	.168	.219	.752	.742	.742	.665	.696	.792	.731	–	.772	.690	.681	.726
B9	.262	.173	.242	.699	.699	.689	.637	.676	.752	.699	.772	–	.671	.654	.699
B10	.238	.163	.206	.693	.695	.728	.758	.717	.708	.707	.690	.671	–	.836	.844
B11	.235	.158	.213	.683	.694	.717	.714	.694	.689	.689	.681	.654	.836		.835
B12	.260	.165	.219	.743	.748	.794	.799	.770	.745	.762	.726	.699	.844	.835	–

Table 3. Inter Item Correlation Matrix

subject matter of this course is well organized), for example, is positively correlated with B3 (The instructor clearly presents his/her subject matter) (r = .841).

Table 4 is the Item Total Statistics table. This is, perhaps, the most significant table of the Cronbach's Alpha model. The table provides five pieces of information for each item on the scale. The two most significant portions of the table are the Corrected –Item Total Correction and the Alpha if Item Deleted. The earlier is the correlation of each detailed item with the complete total of the remaining items on the scale. If it is assumed the correlation of this to be discreetly high of say .40 or greater, then it can be surmised the precise item to be, at least temperately correlated with most of the remaining items and will generate a worthy module of this conjectured rating scale. In the case of Table A4, almost all of the items on the scale had significant inter–correlations with each other. The scale mean for item B4, for example, was 60.57, and a significant positive correlation (r = .81) with other items on the scale. Its coefficient alpha – the extent to which the item is consistent or inter–correlated for the scale was 0.95.

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
A1	60.28	111.566	.369	.513	.957
A2	60.27	113.559	.256	.279	.959
A3	60.34	111.653	.328	.476	.958
B1	60.48	101.518	.839	.748	.949
B2	60.62	99.358	.837	.768	.949
B3	60.58	99.131	.862	.807	.948
B4	60.57	99.921	.812	.713	.949
B5	60.65	99.419	.835	.719	.949
B6	60.66	99.117	.847	.757	.948
B7	60.59	100.295	.819	.697	.949
B8	60.58	100.850	.823	.743	.949
B9	60.64	100.397	.791	.677	.950
B10	60.56	100.073	.828	.785	.949
B11	60.56	100.672	.811	.762	.949
B12	60.55	99.423	.875	.835	.948

Table 4. Item-Total Statistics

A1 – I have put a great deal of effort into this course.

A2 – I have attended classes regularly.

A3 – I have completed the required readings for this course.

B1 – The stated goals and objectives for this course are consistent with what was actually taught.

B2 – The subject matter of this course is well organized.

B3 – The instructor clearly presents his/her subject matter.

B4 – The instructor is enthusiastic and arouses interest in his course.

B5 – My power to think, criticize, and/or create have been improved as a result this course.

B6 – The texts and other readings assigned for this course have been helpful.

B7 – The instructor uses instructional approaches (for example, discussions, lectures, audio–visuals, field work, demonstrations, computer programs, etc.) which effectively enhance learning in this course.

B8 – The examinations are consistent with the course objectives and the instruction.

B9 – Quizzes, examinations and/or written assignments are provided frequently enough to help me evaluate my progress.

B10 – The instructor is genuinely concerned with students' progress.

B11 – I am able to get help from the instructor when I need it.

B12 – This instructor is effective in promoting learning.

According to the results for all 15 items in the table, students offered optimistic assessments of the course teachings they received. Notwithstanding, the standard deviations for the 15 items were also less than unity, indicating that the scale means thoroughly mirrored the responses of most students, and there was not an extensive disparity of responses to the rating items. Therefore, the excessive coefficient alpha for the individual items on the scale revealed that, the items were consistent and highly inter–correlated and together, they indicated good internal consistency reliability. This is, in addition to Table 1, where the alpha for all 15 items was .95, thus indicating that, all 15 items produced a scale that had realistic internal consistency reliability. Therefore, the null hypothesis that student ratings do not increase the reliability of the instrument used to assess teaching quality is rejected. The instrument does show strong internal reliability.

Hypothesis 2

The study also conducted the principal component factor analysis (PCA) with varimax rotation, to determine the underlying structure for the 15 items on the rating scale. It is also to check the validity of the rating instrument. The initial step in this analysis is the establishment of a correlation matrix for the 15 items on the scale. Non–significant correlations between the items would suggest they are distinct and will, therefore, not be expected to create one or more factors. If this is realized, it would not be realistic to forge ahead with the construction of a factor analysis. Table 5 below shows the estimated correlation and the level of significance of the 15 items on the rating scale. All of the items were significant (p = .000) and in correlations with one another, thus suggesting they may establish one or more factors.

Next, is the Kaiser–Meyer–Olkin (KMO) statistic, a measurement of "sampling suitability". Meaning that it tells whether there are enough items predicted by each factor. A zero value shows that the summation of fractional correlations is greater than the totality of the correlations, representing dispersal in the configuration of correlations; and; therefore, factor analysis is probably unsuitable. A value near one point to a configuration of correlations that is reasonably condensed and, therefore, factor analysis ought to produce unique and dependable factor (Agbetsiafa, 2010). Field (2009) suggested that, the KMO evaluation must be > .70, and is lacking if < .50. Leech, Barrett, and Morgan (2011) also said KMO must be > .70, to justify adequate items for every factor. Thus, in the below Table 6, the KMO statistic for the 15 items on the scale was .960 ('superb', as stated by Field, 2009), confirming the sampling adequacy for the analysis. In addition, the entire KMO values (Table 6) for individual items were > .77, which is beyond the tolerable margin of .50 (Field, 2009). The Barlett's test of sphericity followed the KMO measurement. This is a measure of the null hypothesis that, the preliminary correlation matrix is an identity matrix. Thus, for factor

*Correlation Matrix*ᵃ

		A1	A2	A3	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
	A1	1.000	.496	.670	.295	.258	.264	.247	.332	.290	.257	.239	.262	.238	.235	.260
	A2	.496	1.000	.458	.209	.170	.176	.169	.204	.184	.175	.168	.173	.163	.158	.165
	A3	.670	.458	1.000	.260	.234	.235	.211	.277	.269	.217	.219	.242	.206	.213	.219
	B1	.295	.209	.260	1.000	.799	.817	.687	.718	.744	.705	.752	.699	.693	.683	.743
	B2	.258	.170	.234	.799	1.000	.841	.702	.721	.746	.711	.742	.699	.695	.694	.748
	B3	.264	.176	.235	.817	.841	1.000	.750	.745	.756	.743	.742	.689	.728	.717	.794
	B4	.247	.169	.211	.687	.702	.750	1.000	.741	.698	.730	.665	.637	.758	.714	.799
Correlation	B5	.332	.204	.277	.718	.721	.745	.741	1.000	.771	.722	.696	.676	.717	.694	.770
	B6	.290	.184	.269	.744	.746	.756	.698	.771	1.000	.736	.792	.752	.708	.689	.745
	B7	.257	.175	.217	.705	.711	.743	.730	.722	.736	1.000	.731	.699	.707	.689	.762
	B8	.239	.168	.219	.752	.742	.742	.665	.696	.792	.731	1.000	.772	.690	.681	.726
	B9	.262	.173	.242	.699	.699	.689	.637	.676	.752	.699	.772	1.000	.671	.654	.699
	B10	.238	.163	.206	.693	.695	.728	.758	.717	.708	.707	.690	.671	1.000	.836	.844
	B11	.235	.158	.213	.683	.694	.717	.714	.694	.689	.689	.681	.654	.836	1.000	.835
	B12	.260	.165	.219	.743	.748	.794	.799	.770	.745	.762	.726	.699	.844	.835	1.000
	A1		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	A2	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	A3	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	B1	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	B2	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	B3	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000	.000
	B4	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000	.000
Sig. (1-tailed)	B5	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000	.000
	B6	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000	.000
	B7	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	B8	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000	.000
	B9	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000	.000
	B10	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000	.000
	B11	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000		.000
	B12	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	

a. Determinant = .001

Table 5. Correlation Matrix

Kaiser–Meyer–Olkin Measure of Sampling Adequacy.		.960
Bartlett's Test of Sphericity	Approx. Chi–Square	271591.204
	df	105
	Sig.	.000

Table 6. KMO and Barlett's Test

analysis to function, it requires certain associations between variables, and if the R–matrix is an identity matrix, it follows that the entire correlation coefficients would be zero (Field, 2009). The Barlett's assessment must, therefore, be significant to be considered worthy. Leech et al. (2011) also noted the Barlett's test of sphericity must be significant (p < .05) to establish that, the correlation matrix is significantly dissimilar to an identity matrix, where correlations between variables are entirely zero. Therefore, according to the above Table 6, the Bartlett's test of sphericity χ^2 (105) = 271591.204, p < .001 was significant, demonstrating that correlations between items were adequately strong for PCA. In addition, a preliminary examination was prompted to acquire eigenvalues for each element [component] in the data. This is revealed in the below Table 7. Two

Anti-image Matrices

	A1	A2	A3	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
A1	.487	-.162	-.275	-.015	.003	.004	.004	-.040	-.002	-.006	.012	-.005	.003	.006	-.006
A2	-.162	.721	-.118	-.019	.005	.002	-.008	.001	.009	-.008	-.006	.002	-.007	.001	.008
A3	-.275	-.118	.524	-.002	-.003	-.002	.002	-.002	-.019	.009	.005	-.012	.005	-.011	.008
B1	-.015	-.019	-.002	.252	-.053	-.060	.003	-.016	-.014	-.004	-.039	-.018	-.004	-.003	-.008
B2	.003	.005	-.003	-.053	.232	-.080	-.007	-.014	-.015	-.006	-.021	-.020	.003	-.013	-.002
B3	.004	.002	-.002	-.060	-.080	.193	-.031	-.011	-.013	-.021	-.007	.009	.000	-.001	-.024
B4	.004	-.008	.002	.003	-.007	-.031	.287	-.050	-.004	-.047	.005	.006	-.040	.002	-.044
B5	-.040	.001	-.002	-.016	-.014	-.011	-.050	.281	-.065	-.025	.005	-.009	-.009	-.002	-.027
B6	-.002	.009	-.019	-.014	-.015	-.013	-.004	-.065	.243	-.024	-.066	-.051	-.009	-.002	-.001
B7	-.006	-.008	.009	-.004	-.006	-.021	-.047	-.025	-.024	.303	-.038	-.034	-.004	-.004	-.027
B8	.012	-.006	.005	-.039	-.021	-.007	.005	.005	-.066	-.038	.257	-.085	-.004	-.010	-.003
B9	-.005	.002	-.012	-.018	-.020	.009	.006	-.009	-.051	-.034	-.085	.323	-.015	-.005	-.007
B10	.003	-.007	.005	-.004	.003	.000	-.040	-.009	-.009	-.004	-.004	-.015	.215	-.090	-.051
B11	.006	.001	-.011	-.003	-.013	-.001	.002	-.002	-.002	-.004	-.010	-.005	-.090	.238	-.061
B12	-.006	.008	.008	-.008	-.002	-.024	-.044	-.027	-.001	-.027	-.003	-.007	-.051	-.061	.165
A1	.799[a]	-.273	-.545	-.043	.010	.013	.012	-.108	-.006	-.014	.035	-.012	.010	.019	-.020
A2	-.273	.878[a]	-.192	-.046	.013	.006	-.017	.002	.022	-.016	-.013	.003	-.017	.001	.023
A3	-.545	-.192	.795[a]	-.006	-.009	-.007	.005	-.006	-.054	.023	.013	-.029	.014	-.030	.027
B1	-.043	-.046	-.006	.974[a]	-.220	-.273	.009	-.059	-.055	-.015	-.154	-.063	-.018	-.013	-.037
B2	.010	.013	-.009	-.220	.966[a]	-.379	-.028	-.055	-.063	-.021	-.086	-.073	.015	-.054	-.010
B3	.013	.006	-.007	-.273	-.379	.960[a]	-.130	-.046	-.060	-.087	-.031	.035	-.001	-.004	-.136
B4	.012	-.017	.005	.009	-.028	-.130	.976[a]	-.176	-.015	-.160	.020	.019	-.163	.009	-.204
B5	-.108	.002	-.006	-.059	-.055	-.046	-.176	.977[a]	-.250	-.087	.018	-.029	-.035	-.006	-.127
B6	-.006	.022	-.054	-.055	-.063	-.060	-.015	-.250	.971[a]	-.089	-.264	-.180	-.041	-.008	-.005
B7	-.014	-.016	.023	-.015	-.021	-.087	-.160	-.087	-.089	.984[a]	-.135	-.109	-.015	-.013	-.119
B8	.035	-.013	.013	-.154	-.086	-.031	.020	.018	-.264	-.135	.966[a]	-.294	-.018	-.042	-.014
B9	-.012	.003	-.029	-.063	-.073	.035	.019	-.029	-.180	-.109	-.294	.974[a]	-.057	-.019	-.032
B10	.010	-.017	.014	-.018	.015	-.001	-.163	-.035	-.041	-.015	-.018	-.057	.958[a]	-.400	-.270
B11	.019	.001	-.030	-.013	-.054	-.004	.009	-.006	-.008	-.013	-.042	-.019	-.400	.957[a]	-.310
B12	-.020	.023	.027	-.037	-.010	-.136	-.204	-.127	-.005	-.119	-.014	-.032	-.270	-.310	.962[a]

npling Adequacy(MSA)

Table 7. Anti-image Matrices

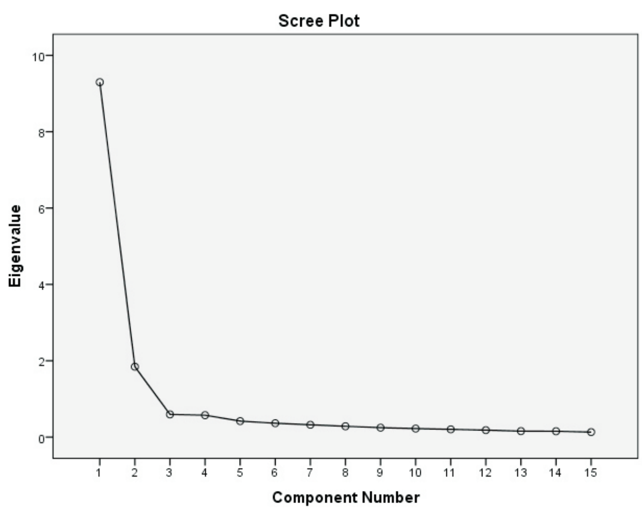


Figure 1. Scree Plot

elements had eigenvalues above Kaiser's condition of 1 (61.98%; 12.30%) and together, explained 74.28% of the variance. The scree plot (Figure 1) showed that, following the first two components, variances amongst the eigenvalues declined (the curve flattened), and were < 1.0. The graph again supported a two–component resolution.

Arguably, with the outsized sample size, and the merging of the scree plot and Kaiser's standard on two components, these were the quantities of components that were

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	9.298	61.984	61.984	9.298	61.984	61.984	9.245
2	1.844	12.291	74.276	1.844	12.291	74.276	2.985
3	.595	3.966	78.242				
4	.574	3.828	82.070				
5	.420	2.798	84.868				
6	.364	2.427	87.295				
7	.322	2.148	89.442				
8	.284	1.893	91.335				
9	.247	1.648	92.984				
10	.223	1.487	94.471				
11	.203	1.354	95.825				
12	.183	1.223	97.048				
13	.157	1.048	98.096				
14	.153	1.022	99.118				
15	.132	.882	100.000				

Extraction Method: Principal Component Analysis.
a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Table 8. Total Variance

engaged in the final analysis. Table 8 displays the items and factor loadings after rotation. The first factor, which appeared to index the promotion of learning and teaching effectiveness, had strong loadings on the first twelve items. Even when quizzes and examinations also had strong loadings (>= .82), they had the least value of the twelve items in the first factor. In addition, the second factor, which appeared to index students' efforts also had high loadings on the remaining three items in the table. The item "I have put a great deal of effort into this course", had the highest loading from the students' efforts perspective. Notwithstanding, the first factor which had information on effective learning advancement, goals and objectives, student–teacher relationship, among others, were found to be significantly related to teaching effectiveness, thus affirming the construct validity of the instrument used to evaluate teaching quality. Therefore, the null hypothesis that the assessment instrument has no validity is rejected.

The Cronbach's Alpha Test statistic of .95 (Table 1) illustrates an extraordinarily high internal consistency of the student rejoinders to the items on the rating scale. It thus confirms the reliability of the instrument used in evaluating teaching effectiveness. The result is in consonance with similar studies on the reliability of student ratings with the application of the Cronbach's Alpha statistic (Agbetsiafa, 2010; Anastasiadou, 2011; Beran, T., Violato, C., Kline, D., & Frideres, J., 2009; Donnon et al., 2010; Kneipp et al., 2011; Zerbini et al., 2009; Zuberi et al., 2007). Additionally, it is also realized that, the 15 questions on the rating scale considered as variables, entirely disintegrated into two modules or factors: Teaching Effectiveness consisting of 12 questions with factor loads ranging between .82 and .90; and, Students' Efforts consisting of three questions with factor loads ranging between .76 and .86 (See the above Table B5). The results also confirm that, the scale offered evidence of construct validity. Indeed, the factor loads of teaching effectiveness are significantly high, thus endorsing the construct validity of teaching excellence at HBCU, even with the exclusion of end–of–course grades. Again, the results of the study are consistent with similar results on the validity of student assessment using PCA (Agbetsiafa, 2010; Zerbini et al., 2009; Beran et al., 2009; Skowronek, J., Friesen, B., & Masonjones, H., 2011; Donnon et al., 2010; Safavi, S., Bakar, K., Tarmizi, R., & Alwi, N., 2012; Spooren et al., 2007; Beran, T., Violato, C., & Kline, D., 2007).

Arguably, student ratings are crucially necessary for summative purposes at HBCU. The ratings have been extensively applied in administrative decisions in connection with faculty retention, promotion, salary increase, and tenure approval. There are studies which have suggested the unease of faculty to use the ratings by the administration for summative [personnel] purposes (Abrami as cited in Beran et al., 2007; Nasser & Fresko as cited in Beran et al., 2009). However, this study cannot confirm the perception of the faculty on the use of student ratings for summative purposes since no personal interview with the faculty was conducted. That the ratings continued to be applied by the administration at HBCU for summative purposes, is aptly justified by the results of the research, even when this is not a trend study. Again, even when this is inconclusive, still, there are indications that some instructors carry a negative view of the student ratings out of unease that the ratings may be subjective by individualities of the instructors and courses as, for example, assignment grades, among others. These indications are consistent with the findings of Eiszlet, who said the instructors were concerned that student ratings may be prejudiced by the distinctiveness of the instructors (as cited in Beran et al., 2007). Even so, there is hardly any doubt about the internal consistency and construct validity of the rating instrument used to assess teaching quality, at least with respect to the findings of this study. There are also indications that HBCU faculty has found the ratings useful for formative purposes like, for example, instructional improvement, even when some have questioned their validity.

Even with the validation of the rating instrument used in assessing teaching quality, the research is not without its limitations. There were 7,919 students who responded to the survey. However, the research made no attempt to distinguish between undergraduate and graduate students; and, between online and distance learning and class–based [traditional] students. It is conceivable; assessing online and classroom–based students independently may have resulted in different outcomes from the existing outputs. However, a similar study found no significant difference in the mean ratings between online and class–based student evaluations of instruction (Stowell et al., 2012). The only significant difference between online and class–based student evaluations of instruction was the low response rate of online students (Anderson, Brown, & Spaeth, 2006; Stowell et al., 2012). Against this backdrop, it is likely distinguishing between online and class–based students' ratings would not have had any significant effect on the findings of this research.

In addition, the study confirms a strong association between student ratings and teaching effectiveness. The inter–correlations between the 12 items on teaching effectiveness are significantly positive, and ranges between .64 and .84 on both the Cronbach's Alpha scale (Table 3), and the correlation matrix result of the factorial analysis model (Table 5). It is likely the sound inter–correlations between the 12 items may have been influenced by the size of the classes. This is so because a single class at HBCU must have at least three enrolled students before the students are allowed to evaluate the instructor. White, for example, noted a positive association between class size and comprehensive analysis of teaching (as cited in Ibrahim, 2011). Ibrahim also summarized class size had a positive influence on the dependability of student ratings of instruction. However, this cannot be independently confirmed since class–size was never a factor in this study. Class size may be considered in future research studies. It is also likely that, the strong inter–correlation between the 12 items is because each has a relationship with teaching excellence.

Now, the research also did not consider the study of generalizability. That is; how confident are we that the student ratings perfectly mimic the instructor's comprehensive teaching ability, not just how efficiently he or she was in the 2012 spring semester only. Generalizability could have been resolved if the study had compared the same instructors teaching the same courses but in different semesters; or, the same instructors teaching various courses; or, different instructors teaching the same course. Though the investigations confirmed construct validity of the rating instrument for the 2012 spring semester, it is doubtful if a similar result would have dominion for the same instructors across semesters, for example. Studies, however, have shown a higher correlation between different courses taught by the same instructor, thus concluding it was the instructor, not the course, who was the leading cause of the student rating objects (Gillmore, Kane, & Naccaranto, 1978; Hogan, 1973; Marsh, 1981).

Even when the assessment indicator has shown strong reliability and validity in the completed study, there is the unease that the psychometric properties of the student evaluation of instruction may be antiquated and would need to be updated for significant formative purposes. Arguably, while there are studies that have had three or more factor loadings produced after factorial analysis (Agbetsiafa, 2010; Anastasiadou, 2011), this study only had two psychometric properties after factorial analysis, as already realized: student efforts and teaching effectiveness. The two together formed the 15 items on the rating scale. It would have been worthwhile to ensure the student ratings of instruction evaluate the following areas as well: course composition, organization, objectives, and the attributes, complexity, and significance of a course; instructor's readiness, passion, and course understanding; and, the instructor's purpose of arranging classroom exercises and alluring students in scholastic quests. Still, the earlier properties qualify for the application of factorial analysis.

Additionally, the inter–item correlation matrixes [correlation matrix] between the three items on student efforts are significantly positive and ranges between .46 and .67 (one–tailed greater than 0). At the same time, the inter–item correlations between the three items on student efforts and the 12 items on teaching effective even though significant at $p = 0.001$ (one–tailed greater than 0) are, however, weak (Table 3; and Table 5 in the Appendices). The low correlation is an indication the items on student efforts would require remodeling to establish a stronger association when compared with the items on teaching excellence.

In spite of the limitations of the study, the validity and reliability of the rating instrument is conclusive. Still, the study may not meet the skeptics on the validity and reliability of student ratings, even with the growing statement about their reliability. In any event, with regards to HBCU, administrators and faculty should consider reviewing or build upon the current survey instrument to include the following, to give the student ratings more useful basis for teaching quality and student learning achievement:

Student Section: (i) Comparative effort in the course; (ii) The comparative effort to do well in the course; (iii) Comparative participation in class activities; (iv) Comparative academic challenge of the course; (v) Hours per week spent in class; (vi) Average hours per week utilized in studying; (vii) Seeking instructor's assistance; and (viii) Significant course–work–hours available per week.

Course Section: (i) The course in its entirety; (ii) Course material and structure; (iii) Course objectives; (iv) Attributes, complexity and significance of the course; (v) Instructor's participation in the course; (vi) Instructor's delivery model; (vii) Expected course grade; (viii) Instructor's course understanding; (ix) Assignments in relation to course objectives; and (x) Exams and group discussions.

Instructor's Section: (i) Instructor's readiness and passion; (ii) Instructor's personality and idiosyncrasy; (iii) Motivating interest; (iv) Interaction with students; (v) Instructor's role in understanding the subject matter; (vi) Challenging students, and the use of reading assignments; (vii)

Instructor's support during the learning epoch; (viii) Learning presentation skills; and (ix) Feedback provided.

The above are only limited recommendations which could be strengthened or unconditionally overhauled by the administration and faculty. At the same time, it must be realized the unease of some faculty members to use the student ratings for summative purposes is complete, and cannot be easily dissuaded against such risk even with the strength of the findings of this research. If anything, the unpredictability of scholars to reach a settlement on the validity and reliability of student ratings to assess faculty, has encouraged some to endorse the improvement of the model to encompass other measurements of assessing teaching quality (e.g. peer reviews, self–rating, student interviews, etc.) (Berk, 2005; El Hassan, 2009; Kulik as cited in Donnon, T., Delver, H., & Beran, T., 2010; Marsh, Ginns, Morin, Nagengast, & Martin, 2011). Thus, it would be beneficial to include student ratings with other measurements of assessing teaching effectiveness (e.g. peer reviews, self–rating, instructor's portfolio, student reviews, sophomore and graduating senior surveys, etc.) when considering instructors for tenure and salary increase.

Conclusion

The specific item analysis conducted by the authors using the Tri–Squared Test resulted in the analysis of a massive amount of items first validated through the use of Cronbach's Alpha Test statistic of .95 (Table 1). As previously mentioned, the Cronbach's Alpha Test yielded an extraordinarily high internal consistency of the student rejoinders to the items on the rating scale. It thus confirms the reliability of the instrument used in evaluating teaching effectiveness. The result is in consonance with similar studies on the reliability of student ratings with the application of the Cronbach's Alpha statistic as stated by the following educational researchers: Agbetsiafa, 2010; Anastasiadou, 2011; Beran et al., 2009; Donnon et al., 2010; Kneipp et al., 2010; Zerbini et al., 2009; and Zuberi et al., 2007. The Tri–Squared Test yielded the following results: Tri^2 Critical Value Table = 18.467 (with d.f. = 4 at $\alpha = 0.001$). For d.f. = 4, the Critical Value for $p > 0.001$ is 18.467. The Calculated Tri–Square value is 92.531, thus, the null hypothesis (H_0) is rejected by virtue of the hypothesis test which yields the following: Tri–Squared Critical Value of 18.467 < 92.531 Calculated Tri–Squared Value. This further supported the research findings attributed to the initial research Cronbach's Alpha Test results. Additionally, Tri–Squared Residuals supported SRI validity with some negative outcomes by respondents to items regarding efficacy and consistency. However, the overwhelming majority of responses supported all three factors pertaining to validity, efficacy and reliability of the SRI. As result, the initial research finds have additional support through this study. It is was again determined that 15 questions on the rating scale considered as "variables", entirely disintegrated into two modules or factors: "Teaching Effectiveness" consisting of 12 questions with factor loads ranging between .82 and .90; and, "Students' Efforts" consisting of three questions with factor loads ranging between .76 and .86 (Table 9). The results also confirm that, the scale offered evidence of overall SRI construct validity. Indeed, the factor loads of teaching effectiveness are significantly high, thus endorsing the construct validity of teaching excellence at this particular HBCU [Note: This data

Item	Factor Loading 1	2	Communality
This instructor is efficient in promoting learning.	.904		.83
The instructor clearly presents his/her subject matter.	.886		.80
The instructor is genuinely concerned with students' progress.	.865		.76
The subject matter of this course is well organized.	.864		.76
The texts and other readings assigned for this course have been helpful.	.860		.77
The examinations are consistent with the course objectives and the instruction.	.854		.74
The stated goals and objectives for the course are consistent with what was actually taught.	.850		.76
I am able to get help from the instructor when I need it.	.849		.73
The instructor uses instructional approaches (e.g. discussions, lectures, audiovisuals field work, demonstrations, computer programs, etc.) which effectively enhance learning in this course.	.849		.74
The instructor is enthusiastic and arouses interest in this course.	.846		.73
My power to think, criticize, and/orcreate have been improved as a result of this course.	.839		.75
Quizzes, examinations and/or written assignments are provided frequently enough to help evaluate my progress.	.815		.69
I have put a great deal of effort into this course.		.857	.77
I have completed the required readings for this course.		.847	.74
I have attended class regularly.		.762	.59

Note. General note goes here.

Table 9. Factor Loadings for the Rotated Factors

was acquired during the 2012 Academic Spring Semester at the HBCU with the exclusion of end–of–course grades]. Again, the results of the study are consistent with similar results on the validity of student assessment using PCA (Agbetsiafa, 2010; Zerbini et al., 2009; Beran et al., 2009; Skowronek et al., 2011; Donnon et al., 2010; Safavi et al., 2012; Spooren et al., 2007; Beran et al., 2007).

Tri–Squared Test Table

The Application of the Cronbach's Alpha Reliability Model on Three Factors as Qualitative Outcomes to Determine SRI Efficacy Using the Tri–Squared Test

Data Analyzed Using the Trichotomous–Squared Three by Three Table designed to analyze the research questions and data extracted from an Inventive Investigative Instrument designed with the following Trichotomous Categorical Variables: a_1 = Is the Student Rating of Instruction Instrument effective?; a_2 = Is the Student Rating of Instruction Instrument valid?; and a_3 = Is the Student Rating of Instruction Instrument consistent? The 3 × 3 Table has the following Trichotomous Outcome Variables: b_1 = Yes; b_2 = No; and b_3 = No Opinion. The Inputted Qualitative Outcomes are reported as follows (for $56451_{[\text{Total Items}]} \in 7919_{[\text{Grand Total SRI}]}$):

$n_{Tri} = 56451_{[\text{Total Items}]}$

$\alpha = 0.001$

TRICHOTOMOUS TESTING INPUT VARIABLES

TRICHOTOMOUS RESULTS OUTPUT VARIABLES	a_1	a_2	a_3
b_1	15629	15751	15125
b_2	1222	1058	1406
b_3	1966	2008	2286

$$Tri^2\ d.f. = [C-1][R-1] = [3-1][3-1] = 4 = Tri^2_{[\bar{x}]}$$

The Tri–Square Test Formula for the Transformation of Trichotomous Qualitative Outcomes into Trichotomous Quantitative Outcomes to Determine the Validity of the Research Hypothesis:

$$Tri^2 = T_{Sum}\left[\left(Tri_x - Tri_y\right)^2 : Tri_y\right]$$

Tri^2 Critical Value Table = 18.467 (with d.f. = 4 at $\alpha = 0.001$). For d.f. = 4, the Critical Value for $p > 0.001$ is 18.467. The Calculated Tri–Square value is 92.531, thus, the null hypothesis (H_0) is rejected by virtue of the hypothesis test which yields the following: Tri–Squared Critical Value of 18.467 < 92.531 Calculated Tri–Squared Value.

Results

1.) Tri–Squared Calculated Value = 92.531

2.) Tri–Squared Degrees of Freedom = 4

3.) Tri–Squared Probability = 0.0096

4.) Tri–Squared Alpha Level = 0.001 [for $n_{Tri} = 56451_{[\text{Total Items}]}$ Maximized Test Critical Value]

Tri–Squared Percentage Deviations

	a_1	a_2	a_3
b_1	+0.8%	+1.6%	–2.4%
b_2	–0.5%	–13.8%	+14.4%
b_3	–5.8%	–3.8%	+9.6%

Percentage deviation and standardized residual are both measures of the degree to which an observed Tri–Squared cell frequency differs from the value that would be expected on the basis of the null hypothesis.

Goodman & Kruskal's Lambda (λ) Tri–Squared Results

Goodman & Kruskal's Lambda (λ) is a cross tabulation analysis measure of proportional reduction in error (Table 10, Table 11). Lambda indicates the extent to which the modal categories and frequencies for each value of the independent variable differ from the overall modal category and frequency. The Goodman–Kruskal Values for Lambda range from zero (indicating that there is "no association" between independent and dependent variables) to one (indicating a "perfect association" between independent and dependent variables). It is calculated with the following equation:

$$\lambda = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$$

Where,

ε_1 = is the overall non–modal frequency; and

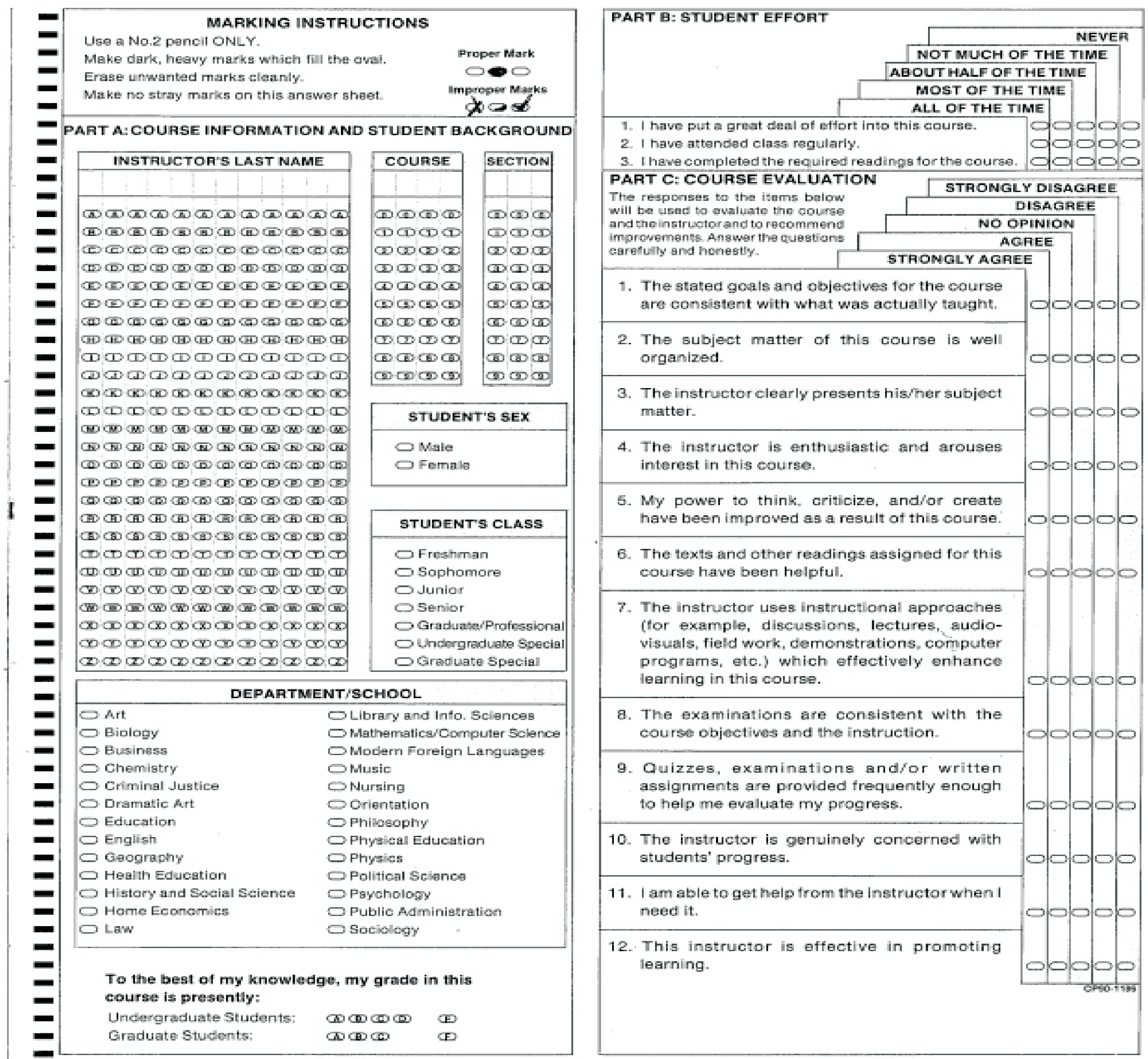
ε_2 = is the sum of the non–modal frequencies for each value of the independent variable.

Crosstabulation of Variables		Independent Variables			Results:
		Categorical Variable 1 = a_1	Categorical Variable 2 = a_2	Categorical Variable 3 = a_3	=
Dependent Variables	Outcome Variable 1 = b_1	15629	15751	15125	***46505***
	Outcome Variable 2 = b_2	1222	1058	1406	***3686***
	Outcome Variable 3 = b_3	1966	2008	2286	***6260***
Results:	=	***18817***	***18817***	***18817***	***56451***

Table 10. Goodman & Kruskal's Lambda (λ) Tri–Squared

Lambda (λ)or Predicting		Standard Error	.95 Confidence Interval Limits Upper	Lower
afromb	0	—	—	—
bfroma	0.0166	0.0052	0.0065	0.0267
Estimated Probability of Correct Prediction when Predicting				
a without knowledge of b:				0.8238
a from b:				0.8238
b without knowledge of a:				0.3333
b from a:				0.3444

Table 11. Goodman & Kruskal's Lambda (λ) Tri–Squared Results: Index of Predictive Association


MARKING INSTRUCTIONS

Use a No.2 pencil ONLY.
Make dark, heavy marks which fill the oval.
Erase unwanted marks cleanly.
Make no stray marks on this answer sheet.

Proper Mark

Improper Marks

PART A: COURSE INFORMATION AND STUDENT BACKGROUND

INSTRUCTOR'S LAST NAME | COURSE | SECTION

STUDENT'S SEX
- Male
- Female

STUDENT'S CLASS
- Freshman
- Sophomore
- Junior
- Senior
- Graduate/Professional
- Undergraduate Special
- Graduate Special

DEPARTMENT/SCHOOL
- Art
- Biology
- Business
- Chemistry
- Criminal Justice
- Dramatic Art
- Education
- English
- Geography
- Health Education
- History and Social Science
- Home Economics
- Law
- Library and Info. Sciences
- Mathematics/Computer Science
- Modern Foreign Languages
- Music
- Nursing
- Orientation
- Philosophy
- Physical Education
- Physics
- Political Science
- Psychology
- Public Administration
- Sociology

To the best of my knowledge, my grade in this course is presently:

Undergraduate Students: A B C D F
Graduate Students: A B C F

PART B: STUDENT EFFORT

NEVER / NOT MUCH OF THE TIME / ABOUT HALF OF THE TIME / MOST OF THE TIME / ALL OF THE TIME

1. I have put a great deal of effort into this course.
2. I have attended class regularly.
3. I have completed the required readings for the course.

PART C: COURSE EVALUATION

The responses to the items below will be used to evaluate the course and the instructor and to recommend improvements. Answer the questions carefully and honestly.

STRONGLY DISAGREE / DISAGREE / NO OPINION / AGREE / STRONGLY AGREE

1. The stated goals and objectives for the course are consistent with what was actually taught.
2. The subject matter of this course is well organized.
3. The instructor clearly presents his/her subject matter.
4. The instructor is enthusiastic and arouses interest in this course.
5. My power to think, criticize, and/or create have been improved as a result of this course.
6. The texts and other readings assigned for this course have been helpful.
7. The instructor uses instructional approaches (for example, discussions, lectures, audio-visuals, field work, demonstrations, computer programs, etc.) which effectively enhance learning in this course.
8. The examinations are consistent with the course objectives and the instruction.
9. Quizzes, examinations and/or written assignments are provided frequently enough to help me evaluate my progress.
10. The instructor is genuinely concerned with students' progress.
11. I am able to get help from the instructor when I need it.
12. This instructor is effective in promoting learning.

Source: REP

Appendix Student Ratings of Instruction Survey Instrument

References

[1]. **Agbetsiafa, D. (2010).** Evaluating effective teaching in college level Economics using student ratings of instruction: A factor analytic approach. *Journal of College Teaching & Learning*, 7(5), 57–66.

[2]. **Anastasiadou, S. D. (2011).** Reliability and validity testing of a new scale for measuring attitudes toward learning statistics with technology. *Acta Didactica Napocensia,* 4(1), 1–10.

[3]. **Barry, R. A. (2010).** *Teaching effectiveness and why it matters.* Marylhurst University and the Chalkboard Project, Retrieved from http://chalkboardproject.org/wp–content/uploads/2010/12/teacher–effectiveness–and–why–it–matters.pdf

[4]. **Beran, T., Violato, C., Kline, D., & Frideres, J. (2009).** What do students consider useful about student ratings?. *Assessment & Evaluation in Higher Education,* 34(5), 519–527.

[5]. **Beran, T., Violato, C., & Kline, D. (2007).** What's the "Use" of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, 37(1), 27–43.

[6]. **Berk, R. A. (2005).** Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17(1), 48–62.

[7]. **Bland, J., & Altman, D. G. (1997).** Cronbach's Alpha . *BMJ: British Medical Journal (International Edition),* 314(7080), 572.

[8]. **Bo, Z., Johnston, L., & Kilic, G. (2008).** Assessing the reliability of self– and peer rating in student group work. *Assessment & Evaluation in Higher Education,* 33(3), 329–340. Doi:10.1080/02602930701293181

[9]. **Chen, G., & Watkins, D. (2010).** Stability and correlates of student evaluations of teaching at a Chinese university. *Assessment & Evaluation in Higher Education,* 35(6), 675–685. Doi:10.1080/02602930902977715

[10]. **Clayson, D. E. (2009).** Student evaluations of teaching: Are they related to what students learn?. *Journal of Marketing Education*, 31(1), 16–30.

[11]. **Donnon, T., Delver, H., & Beran, T. (2010).** Student and teaching characteristics related to ratings of instruction in medical sciences graduate programs. *Medical Teacher*, 32(4), 327–332. Doi:10.3109/01421590903 480097

[12]. **El Hassan, K. (2009).** Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research & Development,* 28(3), 319–333. Doi:10.1080/07294360902839917

[13]. **Field, A. (2009).** *Discovering statistics using spss.* (3rd ed.). Thousand Oaks, CA: SAGE Publications Inc.

[14]. **Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978).** The generalizability of student ratings of Instruction: Estimation of the teacher and course components. *Journal of Educational Measurement,* 15, 1–13.

[15]. **Hassel, B. C. (2009).** *How should state define teacher effectiveness?.* Center for American Progress, Retrieved fromhttp://publicimpact.com/publications/PublicImpact–How_Should_States_Define_Teacher_Effectiveness.pdf

[16]. **Hogan, T. P. (1973).** Similarity of student ratings across instructors, courses, and time, *Research in Higher Education,* 1, 149–154.

[17]. **Hunsaker, S. L., Nielsen, A., & Bartlett, B. (2010).** Correlates of teacher practices influencing student outcomes in reading instruction for advanced readers. *Gifted Child Quarterly,* 54(4), 273–282. Doi:10.1177/ 00169862 10374506

[18]. **Ibrahim, A. (2011).** Using generalizability theory to estimate the relative effect of class size and number of items on the dependability of student ratings of instruction. *Psychological Reports,* 109(1), 252–258. Doi:10. 2466 /03.07.11.PR0.109.4.252–258

[19]. **Jones, B. (2010).** An examination of motivation model components in face–to–face and online instruction. *Electronic Journal of Research in Educational Psychology,* 8(3), 915–944.

[20]. **Keeley, J., Furr, R., & Buskist, W. (2010).** Differentiating psychology students' perceptions of teachers using the teacher behavior checklist. *Teaching of Psychology*, 37(1), 16–20. Doi:10.1080/00986280903426282

[21]. **Kneipp, L. B., Kelly, K. E., Biscoe, J. D., & Richard, B. (2010).** The impact of instructor's personality characteristics

on quality of instruction. *College Student Journal,* 44(4), 901–905.

[22]. Leech, N. A., Barrett, K. C., Morgan, G. A. (2011). *IBM: SPSS for intermediate statistics use and interpretation.* (4th ed.). New York, NY: Taylor & Francis Group.

[23]. Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects in student rating's of instructional effectiveness. *Applied Psychological Measurement*, 6, 47–59.

[24]. Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and usefulness. In R. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence–based perspective* (pp. 319–383). Netherlands: Springer.

[25]. Marsh, H. W., Ginns, P., Morin, A. S., Nagengast, B., & Martin, A. J. (2011). Use of student ratings to benchmark universities: Multilevel modeling of responses to the Australian Course Experience Questionnaire (CEQ). *Journal of Educational Psychology,* 103(3), 733–748. Doi:10.1037/a0024221

[26]. McKeachie, W. J. (2007). Good teaching makes a difference—And we know what it is. In R. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence–based perspective* (pp. 457–474). Netherlands: Springer.

[27]. Osler, J. E., Hollowell, P. G. & Nichols, S. M. (2012). Technology Engineering in Science Education: Where Instructional Challenges Interface Nonconforming Productivity to Increase Retention, Enhance Transfer, and Maximize Student Learning. July–September i-manager's *Journal of Educational Technology,* 9 (2), pp. 31–39.

[28]. Osler, J. E., Hollowell, G. P., & Palmer, C. E. (2008). Technology Engineering Science Instruction in the Information Age: Integrating Instructional Technology in K12 Education, June–August, i-manager's *Journal on School Educational Technology,* 4 (1), pp. 12–17.

[29]. Roche, L. A., & Marsh, H. W. (2000). Multiple dimensions of university teacher self concept. *Instructional Science,* 28(5), 439–468.

[30]. Safavi, S., Bakar, K., Tarmizi, R., & Alwi, N. (2012). The role of student ratings of instruction from perspectives of the higher education administrators. *International Journal of Business and Social Science,* 3(9), 233–239.

[31]. Schrodt, P., Witt, P. L., Myers, S. A., Turman, P. D., Barton, M. H., & Jernberg, K. A. (2008). *Communication Education,* 57(2), 180–200. Doi:10.1080/ 03634520701840303

[32]. Skowronek, J., Friesen, B., & Masonjones, H. (2011). Developing a statistically valid and practically useful student evaluation instrument. *International Journal for the Scholarship of Teaching and Learning*, 5(1), 1–19.

[33]. Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom–based student evaluations of instruction. *Assessment & Evaluation in Higher Education,* 37(4), 465–473. Doi:10.1080/02602938 . 2010.545869

[34]. Vogt, W. (1984). Developing a teacher evaluation system. *Spectrum,* 2(1), 41–46.

[35]. Young, S., & Shaw, D. G. (1999). *Profiles of effective college and university teachers.* Higher Education, 70(6), 670–686.

[36]. Zerbini, T., & Abbad, G. (2009). Reação aos procedimentos instrucionais de um curso via internet: Validação de uma escala. *Estudos De Psicologia,* 26(3), 363–371. Doi: 10.1590/S0103–166X2009000300009

[37]. Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, 37(2), 227–235. Doi:10.1080/02602938. 2010.523819

[38]. Zuberi, R. W., Bordage, G., & Norman, G. R. (2007). Validation of the SETOC instrument – student evaluation of teaching in outpatient clinics. *Advances in Health Sciences Education,* 12(1), 55–69.doi:10.1007/s10459–005–2328–y

ABOUT THE AUTHOR

A native of North Carolina, James Osler was born and raised in the City of Medicine. An accomplished artist, Osler enjoys using art as a tool to empower others. He completed his B.A. at NCCU with a concentration in Studio. Osler adores teaching. He has always been interested in how information is delivered and continues to explore the many different methods, models, and modes of instruction. After completing a M.A. in Educational Technology he completed a doctorate in Technology Education at North Carolina State University (NCSU). He has authored a series of books and e-books on the creation of empowering entrepreneurial educational experiences. His research focuses on Fundamental Christian Education from the holistic perspective of Qualitative and Quantitative Instructional Design (Osler, 2010). He has authored the Online Graduate Program in Online Instructional Design that is currently a part of the Online Educational Technology Program in the NCCU School of Education. His interests include: a life filled with a love of Almighty GOD and ministry to his fellow man through: teaching, the research, and service. He has been awarded two of the highest honors at NCCU as an employee and as faculty: The Employee Recognition Award for Outstanding Service in 2001 and The University Award for Teaching Excellence in 2008.