Canadian Journal of Educational and Social Studies
Vol. 1(2), 2021, pp.13-28

Equivalence of Parallel Tests in a Basic Statistics Course in Higher Education Using Classical Measurement Theory

Frank Quansah[1]& Andrews Cobbinah[2]

[1]Department of Educational Foundations, University of Education, Winneba, Ghana
[2]Department of Education and Psychology, University of Cape Coast, Cape Coast, Ghana
Correspondence: Frank Quansah, University of Education, Winneba, Ghana.
Email: fquansah@uew.edu.gh

DOI: 10.53103/cjess.v1i2.11

Abstract

Developing and administering parallel test forms to students in higher education offsets the cost of having assessment scores that have low validity. This research demonstrated the validity and equivalence of parallel tests in a Basic Statistics course. Among other things, the study: (1) established and compared the item specifications of the items on the different test forms developed, and (2) determined the extent of parallelism of the alternate test forms. Three carefully designed alternate forms of achievement tests (using item specification and test specification table) were administered to 504 second-year students. In addition, academic resilience scale was administered to the same students to help ascertain the criterion validity of the alternate forms. The study revealed some level of similarities in the statistical specifications of the alternate test forms. Further analysis showed that the three alternate test forms developed were congeneric forms of parallelism. The authors concluded that developing classical parallel forms of the test is not feasible, but having congeneric parallel test forms offset the cost of having less valid scores which do not represent students' attainment levels. Faculty members are encouraged to make use of parallel test forms in assessing students in higher education.

Keywords: Parallel Test, Validity, Congeneric Form, Tau-Equivalent, Classically Parallel

Introduction

Higher education institutions, in Ghana and beyond, have placed significant importance on designing policies and guiding principles to govern their assessment processes. Well thought out and documented strategies, processes and policies, blueprinting to enhance adequate sampling, feedback to assessors and learners, and appraisal of the complete process are central to any testing enterprise (Crocker, & Algina, 2008). Yet, less attention is given to the evaluation of assessment (Fowell et al., 1999); this limit how the validity of the results from these assessments are understood. Recent assessment theory stresses the importance of construct validity, which relies heavily on theory and evidence to offer insight into the assessment. Characteristically, validity evidence is drawn from non-mutually exclusive five dimensions to provide accuracy in the

inferences made. They include data management, curriculum content, correlational analyses, statistical analyses of test data and assessment effects (Downing & Haladyna, 2009; Kane, 2006). The explicit combination of pieces of evidence required for validation is contingent on the assumptions made and the inferences drawn (Messick, 1989), and goes beyond the validity of the assessment tools that produce assessment score data. There is, therefore, a greater need to use several sources of data to support the soundness of the interpretation and use of assessment results. This is due to the dynamic and complex nature of assessments, and the increasing stakes of assessments (Kane, 2006).

In contemporary times, the validity of assessments in higher education institutions has been threatened for three major reasons. First, cheating in examinations has become predominant in most higher education institutions (Diego, 2017), especially when multiple-choice tests are used. This issue thwarts the objective of evaluating the understanding and application of course contents taught to students, and as such, affects the consequential validity of the assessment results (Forkuor et al., 2019). In a recent study conducted by Odongo et al. (2021), for example, it was revealed that students have resorted to examination cheating through innovative approaches such as well-rehearsed body (parts) language and sitting arrangements. Secondly, the re-use of written test items by faculty members is also a major concern for stakeholders due to the consequences it has on the validity of such assessment results. This is supported by the findings of previous empirical studies which have found that reusing a large proportion of test items leads to malfunctioning of the test items (see Case & Swanson, 1998; Wagner-Menghin et al., 2013). Lastly, the need to develop new but equivalent items by faculty members for students who failed the examination or could not sit for the examination for various reasons possess a threat to the validity of the results (Norcini et al., 2011).

The antidotes to the issues raised in the preceding paragraph concerning the threat to validity in assessment have been discussed in the literature. Scholars have recommended the use of parallel forms of a test (Crocker & Algina, 2008; Graham, 2006; Tavakol & Dennick, 2011). Statistically, parallel forms of a test have similar true score estimates and thus, the estimates are explained by the number of measurement errors (Danner, 2016). Parallel test forms include the vocational aptitude tests (Schmale, 2001) and the intelligence structure test (Liepmann et al., 2007). Schuwirth et al. (2011) highlighted that three key inferences are needed to understand the link between different parallel forms, and to establish the extent of validity of the scores: (a) would the same score be obtained by the same student on all forms of the tests? (b) would the same rank ordering be attained by the students on all test forms from the lower achievers to higher achievers. (c) would the same pass or fail decisions be achieved by the students on all forms of the test. The extent to which these inferences are satisfied explains the various degrees of parallelism. According to Feldt (1980), there are degrees of measurement parallelism. The extent to which a psychological measurement is parallel depends on the equality of several

parameters: content similarity, true score constancy, mean, variance, covariance and validity. These parameters determine which type of parallelism exist. The types include classical parallel forms, essentially classical parallel form, tau equivalent form, essentially tau equivalent form and congeneric form. The classical parallel forms operate under a more restrictive measurement model whereas the congeneric forms operate under the least restrictive measurement model (Graham, 2006). The characteristics for each are shown below:

Classical Parallel Forms or Part

A. Content similarity
B. τ_i constancy
C. $\mu x_1 = \mu x_2$
D. $\sigma^2 x_1 = \sigma^2 x_2 = \ldots\ldots$
E. $\sigma x_1 x_2 = \sigma x_1 x_3 = \sigma x_2 x_3 = \ldots\ldots..$
F. $\sigma x_1 y = \sigma x_2 y = \sigma x_3 y = \ldots\ldots\ldots\ldots$

} Observable relationship

Essentially Classical Parallel Forms or Parts

A. Content similarity
B. $\tau_i g = \tau_i g + C_g h$ (not all $C_g h = 0$)
C. $\mu x_1 \neq \mu x_2 \neq \mu x_3 \neq \ldots\ldots$
D. $\sigma^2 x_1 = \sigma^2 x_2 = \sigma^2 x_3 = \ldots\ldots$
E. $\sigma x_1 x_2 = \sigma x_1 x_2 = \sigma x_1 x_3 = \ldots\ldots\ldots\ldots..$
F. $\sigma x_1 y = \sigma x_2 y = \sigma x_3 y = \ldots\ldots\ldots\ldots$

} Observable relationship

Tau – Equivalent Forms or Parts.

A. Content similarity
B. τ_i constancy
C. $\mu x_1 = \mu x_2 = \mu x_3$
D. $\sigma^2 x_1 \neq \sigma^2 x_2 \neq \sigma^2 x_3 \neq \ldots\ldots..$ because $\sigma^2 \epsilon_g \neq \sigma^2 \epsilon_h$
E. $\sigma x_1 x_2 = \sigma x_1 x_2 = \sigma x_1 x_3 = \ldots\ldots..$because $\tau_i g = \tau_i h$
F. $\sigma x_1 y = \sigma x_2 y = \sigma x_3 y = \ldots\ldots$ because $\tau_i g = \tau_i h$

} Observable relationship

Essentially Tau – Equivalent Forms or Parts.

A. Content similarity
B. $\tau_{i\,g} = \tau_{i\,g} + C_{g\,h}$ (not all $C_{g\,h} = 0$)
C. $\mu x_1 \neq \mu x_2 \neq \mu x_3 \neq \ldots\ldots$
D. $\sigma^2 x_1 \neq \sigma^2 x_2 \neq \sigma^2 x_3 \neq \ldots\ldots..$because $\sigma^2 \epsilon_g \neq \sigma^2 \epsilon_x$
E. $\sigma x_1 x_2 = \sigma x_1 x_2 = \sigma x_1 x_3 = \ldots\ldots..$because $C_{g\,h}$ does
F. $\sigma x_1 y = \sigma x_2 y = \sigma x_3 y = \ldots\ldots$ because $\sigma x_g x_h$

} Observable relationship

Congeneric Parts or Forms

A. Content similarity
B. $\tau_{i\,g} = \beta_{g\,h}\,\tau_{i\,g} + C_{g\,h}$ (not all $\beta_{g\,h} = 1.0$, not all $C_{g\,h} = 0$)
C. $\mu x_1 \neq \mu x_2 \neq \mu x_3 \neq \ldots\ldots$
D. $\sigma^2 x_1 \neq \sigma^2 x_2 \neq \sigma^2 x_3 \neq \ldots\ldots\ldots \sigma^2 \epsilon_g \neq \sigma^2 \epsilon_h \neq$
E. $\sigma x_1 x_2 \neq \sigma x_1 x_3 \neq \ldots\ldots\ldots$ and $\sigma^2 \tau_g \neq \sigma^2 \tau_h \neq$
F. $\sigma x_1 y \neq \sigma x_2 y \neq \ldots\ldots\ldots\ldots$

} Observable relationships

This research mainly demonstrated the equivalence of parallel test forms in a Basic Statistics course at the University of Cape Coast, Ghana. This research was grounded on two objectives: (1) to establish and compare the item specifications of the items on the different test forms developed, and (2) to determine the extent of parallelism of the alternate test forms. The need for this study aroused due to three key needs of the university and faculty members to ensure a valid measure of students' achievement. The needs include (a) the need to reduce cheating in examination by designing different but equivalent test forms which would provide a measure of student achievement, (b) the need to develop supplementary test forms which are equivalent to the actual test form, for students who were not able to sit for the actual examination, and (c) overcoming the dangers of re-using test items by faculty members for subsequent examinations conducted for different batches of students. Developing and demonstrating the equivalence of different forms of tests provide meaningful direction and enlightenment for university management on how to address the key needs highlighted, especially for courses with high stakes. In this study, for example, Basic Statistics course is one of those core courses in the university which is registered by a lot of students, and this coupled with the student fears associated with taking the course, the issues raised are inevitable unless practical steps are taken to deal with that.

This area of investigation is one of the fields where less attention has been paid to by scholars in assessment, in terms of empirical research. The majority of documented information only focused on theoretical concepts on parallelism and item analysis (see Crocker & Algina, 2008; Feldt, 1980; Feldt & Brennan, 1989; Graham, 2006; Nitko, 2001; Tavakol & Dennick, 2011). It was only in 2012 that a group of scholars (Malau-Aduli et al., 2012) from Australia demonstrated the reliability, validity and similarity of parallel examination in a medical school course. The authors, however, only compared the statistical specifications of the items on the various forms. Actual parallel testing was not done by comparing other indicators like the covariances, variances and criterion validity among the various forms. This research, aside exploring and comparing the items analysis indices for the test forms, Feldt's (1980) criteria for testing the degree of parallelism was used. This is an add-up to literature and provides a foundation for future scholars to build on. This paper presents a systematic process of how the test forms were designed and administered as well as the data analysis procedure. Therefore, faculty members can use

this material as a guide to the design and administration of parallel test forms. This approach can be used to reduce cheating, item writing workload and accurately measure student achievement in their respective discipline. Also, this paper serves as an instructional material for teaching student courses in assessment and measurement.

Method

Participants

The study comprised second-year students reading Bachelor of Education (B.Ed.) courses in the University of Cape Coast, Cape Coast, Ghana. During the first semester of the second year, these students take "Educational Statistics" which sought to introduce the fundamental understanding of statistics, and how to properly select appropriate statistical procedures based on the data at hand. The study sampled 504 students comprising 240 males (47.6%) and 264 females (52.4%). The participants were sampled from the following programmes: B.Ed. Accounting, B.Ed. Home Economics, B.Ed. Arts, B.Ed. Social Science, B.Ed. Physical Education, B.Ed. Basic Education, B.Ed. Early Childhood and B.Ed. Management.

Data Collection Tools and Their Development Processes

Three alternate forms of achievement tests and a standardized questionnaire were used for the data collection. The purpose of the achievement tests was to know how much knowledge or mastery students have gained after they have gone through a period of classroom instruction in the course. The development of the achievement tests was guided by two major mechanisms: (a) test specification table and (b) item specification.

Test Specification Table

Test specification table was adopted to ensure that the test measured the thinking skills as well as the content that the tests purported to measure (Nitko, 2001). Through a test specification table, a test plan was formulated by deciding on the relative emphasis that each component of cognitive operation should receive with respect to the content areas being assessed. The test specification table was developed based on the course outline (see Table 1).

Table 1: Table of test specification

Contents	Level of cognitive operation				Total
	Recall	Comprehension	Application	Analysis	
Measures of Central Tendency and Dispersion	1	1	1	0	4
Scales of Measurement	0	1	0	0	1
Correlation and Prediction	0	1	1	0	2
Validity and Reliability in Measurement	2	3	0	0	5
Hypotheses Testing	2	1	0	0	3
Parametric and Non-parametric Statistical Procedure	1	1	1	3	6
Total	6	8	3	3	20

As shown in the test specification table, the items were crafted across a number of content areas with varying levels of cognitive operation. Most of the items were sampled from "parametric and non-parametric statistical procedure" and "validity and reliability in measurement". The majority of the items were crafted to operate at the comprehension level (n=8) whereas few operated at the application (n=3) and analysis level (n=3).

This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part. This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part.

Item Specification

Item specifications define for each standard the indication learners are required to demonstrate to exhibit their content mastery and the content limits of tasks including the items with stem and options (Nitko, 2001). Spaan's (2013) scope of items specification was adapted for use in this study. Five dimensions were identified: general description, content limit, format, sample task, and distractor. The format for all the three alternate forms was multiple choice. All distractors were crafted based on misconceptions of the subject matter.

Designing the Three-Alternate Test Forms

According to Lindquist (as cited in Crocker & Algina, 2008), the test developer's task is characterised by two major decisions – what to measure and how to measure it. In this case, the question of what to measure has been addressed using the test specification table. The latter decision on "how to measure" is addressed in the item specification table. For each of the item specifications, three items were developed. These items were similar in content and difficulty (based on expert judgement). In the process of crafting items, the test specification table and item specification were continually referred to and effort was made to ensure that the items matched the specification details. Items crafted were based on what students already knew. In all, three different sets of alternate form tests were developed with 20-items each.

After a week, the items were re-read and critically reviewed to check for faulty items. Items that were unclear, ambiguous and had clues were removed from the test and new ones replaced. Afterwards, the items were given to a colleague to be reviewed. The instructions regarding the time allowed as well as how the questions should be answered were given. The items were arranged in a way such that less difficult items were on top. Even though less difficult items were brought first, it was ensured that the key to the items did not form any identifiable pattern. After the items were assembled, clear directions were given. The authors personally attempted the first alternate test and 15-20 minutes were exhausted by the authors. As a result, 25 minutes was allowed for each alternate form. The scoring rubric was then prepared for each of the forms.

Academic Resilience Questionnaire

Resilience is a psychological concept perceived in some persons that contribute to success notwithstanding the adversity. Resilience demonstrates the capability to bounce back, to beat the odds and is deemed as a human asset. Academic resilience demonstrates a greater probability of being successful in education despite adversity (Cassidy, 2016). The academic resilience scale comprised 19-items on a 5-point semantic differential scale [Unlikely (1) to Likely (5)]. The factor structure, construct validity, discriminant validity and reliability estimates are well established by Cassidy (2016).

Data Collection Procedures

Prior to data collection, the ethical issues were considered and duly followed. First, a letter of permission and clearance was sought from the IRB, University of Cape Coast, Ghana, for approval for the study to commence. The participants were assured of confidentiality, anonymity, and volition. The consent of the participants was sought verbally but they were free to stop at any point in time. The three alternate test forms were

administered on the same day with 5 minutes' time interval for relaxation. That is, the first paper took 25 minutes, then the examinees were allowed 5 minutes to relax but were not permitted to interact with each other. On a whole, 85 minutes were used for the administration; 75 minutes to respond to the three tests and 10 minutes to rest for the whole period. As part of the instruction, examinees provided index numbers on the sheet. Five invigilators were used for the test administration. Clear instructions were given and efforts were made to create a conducive environment for the test takers. The academic resilience questionnaire was also administered to the examinee a day after the administration of the three alternate tests. They were required to provide their index numbers on the instrument so that it can be matched with their scripts. Enough time was given to the students to respond to the questionnaire items.

Data Analysis Strategy

The data were screened for possible data entry errors. Afterwards, series of analyses were conducted. First, item analysis for each of the items of the test forms was done and the optimum analysis was presented since the major focus of the research was the test level. Statistical indices presented include optimum difficulty, discrimination, and distractor functioning. Secondly, four statistical indicators were compared across all the test forms and the questionnaire. These indicators include mean, variance, covariance and issue of validity. Also, repeated-measures ANOVA, Pearson Product Moment Correlation, and Mauchly's Test of Sphericity analyses were conducted.

All the analyses conducted were done in the framework of classical measurement theory (Graham, 2006). Consequently, item difficulty was defined as the proportion or percentage of students who answered the item correctly, which ranges from 0-1. an item with difficulty indices closer to 0 or 1 should be altered or discarded because it is not giving any information about differences among examinees' trait levels or abilities (Allen & Yen, 2002). Thus, item difficulties of about .30 to .70, generally, maximize the information the test provides about the differences among examinees. In distractor functioning, the general rule, according to Nitko (2001), is that "every distractor should have at least one lower group student choosing it, and lower group students than upper group students should choose it" (p. 326). The discrimination index describes the extent to which a particular test item can differentiate the higher scoring students from lower-scoring students (Nitko, 2001).

Results

Item Analysis for the Test Forms

The items analysis of the items on the test forms were carried out. Indicators that were considered include difficulty, discrimination, distractor functioning and reliability

analyses. Table 1 shows the summary of the item analysis for the various test forms.

Table 2: Item analysis for Test Forms A, B, and C

Criteria	FORM A	FORM B	FORM C
Minimum	3	3	2
Maximum	16	14	14
Optimum difficulty	.397	.454	.397
Optimum Discrimination	.468	.410	.643
Reliability (KR20)	.649	.668	.744
Distractor Functioning (each item)			
None	1	2	0
One	2	6	3
Two	9	8	7
Three	8	4	10

As presented in Table 2, the optimum difficulty of the test forms was somewhat similar. For Forms A and C, the same level of difficulty was attained; these two test forms were more difficult than Form B. With regards to discrimination, the index for Form C showed high levels of discrimination of the items as compared to the other two forms. Although Forms A and B had similar overall discrimination indices, Form A had discriminated more than Form B. The Kuder-Richardson 21 reliability estimate showed that the test Form C had more reliable items than test Forms A and B.

The Extent of Parallelism of the Test Forms

The study also determined the degree of parallelism for the test forms. This was done by comparing the indicators as proposed by Feldt (1980).

Mean and Variance

The mean values and variances of each alternate test form are presented in Table 3. The main focus here was to compare the means and variances of each of the tests to find out whether they are equal or not.

Table 3: Mean and variance

Alternate Forms	Mean	Variance	Value	F	p-value
A	7.89	--			
B	8.75	--	.116	32.833	.000
C	7.88	--			
			Mauchly's W	Chi-square	p-value
A	--	6.77			.
B	--	5.31	.992	3.986	136
C	--	8.41			

Source: Feld Survey (2021)

Results in Table 3 showed that the mean value for test form A (M=7.89) was approximately equal to that of test form C (M=7.88). Test form B (M=8.75) was, however, had a mean score that was unequal to the mean scores of test form A and C. It is expected that all the means values to be equal but only two of the test forms had mean values approximately equal. The results from ANOVA repeated measures showed significant differences among the means of the participants on the three test forms, $F(2, 502) = 32.833$, $p < .001$.

With regards to the variance, all the three test forms had different variances and thus, were unequal. This was shown in the Mauchly's test results which revealed significant differences in the variances for the scores from the three test forms, W=.992, p=.136. Test form A had a variance of 6.77, form B had 5.31 and form C yielded 8.41 variance. As expected, the variances were to be approximately the same but it happened otherwise.

Covariance among the Test Forms and the Psychological Test (Validity)

Table 4 presents the results on the covariance among the test forms and the psychological test. The emphasis of this section is to compare specific covariance to find out their equality.

Table 4: Covariance

Alternate Form	A	B	C	Academic Resilience (Y)
Form A		2.540	4.067	1.712
Form B	2.540		2.858	2.206
Form C	4.067	2.858		1.795
Academic Resilience	1.712	2.206	1.795	

Source: Feld Survey (2021)

The covariance among the test forms was unequal. Covariance for test forms AB was 2.540, AC was 4.067 and BC was 2.858. For the validity issues which describes the covariance of a test form with the psychological test, the covariance was found as unequal. The covariance for AY was 1.712, BY was 2.206 whereas CY was 1.795 (Table 4).

To support the understanding of the relationships existing among the test forms and, for that matter, the extent of parallelism, a scatter plot was presented to understand the links between the test scores from the three tests and the resilience variable (see Figure 1). Before this, a normality test was also conducted using Q-Q plots and histograms (see Appendix). It was observed that the distributions of errors were similar across the three test forms.

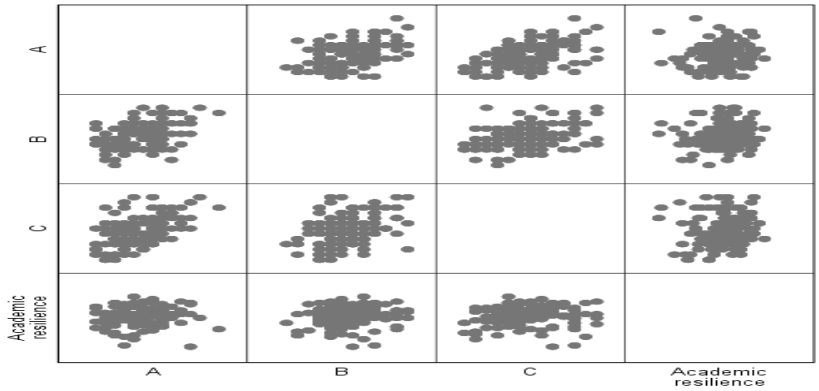


Figure 1: Association among scores from the three alternate test forms and the psychological variable (resilience)

The scatter plots from Figure 1, showed some level of association among the test forms, with the correlation distribution looking similar to each other. The relationship between A & B, A & C, and B & C, all showed a moderate significant association level with a correlation coefficient between .42 and .55. A similar distribution was found for the criterion validity for the test forms and the criterion variable (i.e., resilience). This provided some sense of parallelism existing among the test forms.

Summing all the characteristics, it is clear that all the parameters were unequal. That is to say that the means, variances, covariance among the test forms, and the covariance of the test forms with the psychological test, were all not the same. This suggests that the alternate forms we developed are a congeneric form of parallel test, which is a less restrictive form of parallel test that needs the true score and test content to be equivalent. The evidence of congeneric parallel test is also observed in the similarities of the results from the correlation coefficient, Q-Q plots, item difficulty and reliability coefficients among others.

Discussion and Conclusion

The study revealed that the item specifications for the items on the different test forms were similar though not the same. This was also supported by the correlation matrix scatter plot which showed a similar distribution of scores across the test forms. Nevertheless, the extent of parallelism was found to be the congeneric type. It must be indicated that developing classically parallel forms of the test seems practically impossible due to several factors. Danner (2016) supported this debate and argued that constructing strictly parallel test forms is very difficult since different test items normally measure different parts of the construct. However, consistently developing and validating tests or items using robust approaches are likely to achieve an adequate level of equivalence among alternate test forms (Crocker & Algina, 2008; Graham, 2006; Nitko, 2001; Tavakol & Dennick, 2011). This explains why institutions like Education Testing Services (ETS) have operated and survived for several years by developing and implementing parallel forms of tests for people around the globe. The validity of these tests has been developed and tested over time.

The findings of this study contradicted the findings of Malau-Aduli et al. (2012) who demonstrated that indicators such as the mean values were the same for the test forms they used. Malau-Aduli et al. further concluded that the test forms developed and administered were parallel, although the exact type of parallelism was not indicated. This was because indicators like covariance, variances, and criterion validity were not examined. Once the mean values were the same, the type of parallelism could never be congeneric. The discrepancies in the results of Malau-Aduli et al. against that of this study can be attributed to differences in sample size. Whereas this study used a sample of 504, Malau-Aduli and colleagues used 76 examinees. Hence, attaining no significant difference in mean scores could be due to chance because of the small sample size. Scholars such as Pallant (2010) and Field (2009) have indicated that large sample sizes usually generate significant results.

The validity question within the framework of parallel test development and administration requires that parallel test forms should measure similar or same person abilities and thus, cognitive demands for each item on the test forms should be equivalent. This research showed a minimal level of equivalence in the parallel test forms developed in the Basic Statistics course. This translated into the congeneric form of parallel test. The results of this study start the discussion for the use of parallel forms of tests for assessing students. It is obvious developing classical parallel forms of a test is not feasible, but having a congeneric parallel test form can offset the cost of having less valid scores which do not represent students' attainment levels. At least, this research showed that the statistical parameters of the test forms were quite close when compared. We recommend that faculty members should adopt the use of parallel test forms in assessing students in higher

education. It must be pointed out that test development is a complex process and thus, expertise is required in designing parallel test forms. It is also recommended that there should be adequate training for faculty members on the processes for developing highly valid test forms. Studies of this nature should be conducted in other disciplines to help support the existing findings.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Illinois: Waveland Press.

Case, S., & Swanson, D. B. (1998). Constructing written test questions for the basic and clinical sciences. Philadelphia, PA: National Board of Medical Examiners.

Cassidy, S. (2016). The academic resilience scale (ARS-30): A new multidimensional construct measure. *Educational Psychology*, 7(1), 1-11. https://doi.org/10.3389/fpsyg.2016.01787

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning Press.

Danner, D. (2016). Reliability – The precision of a measurement. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences. https://doi: 10.15465/gesis-sg_en_011

Diego, A. (2017). Friends with benefits: causes and effects of learners' cheating practices during examination. *IAFOR Journal of Education*, 5(2), 121–138. https://files.eric.ed.gov/fulltext/EJ1156266.pdf

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37, 830-837. doi:10.1046/j.1365-2923.2003. 01594.x

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika*, 45, 99-105. https://doi.org/10.1007/BF02293600

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). Phoenix, AZ: Ornyx.

Field, A. (2009). Discovering statistics using SPSS (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Forkuor, J. B. Amarteifio, J., & Attoh D. O. (2019). Students' perception of cheating and the best time to cheat during examinations. *The Urban Review*, 51(3), 424–443. https://doi.org/10.1007/s11256-018-0491-8

Fowell, S. L., Southgate, L. J., & Bligh, J. G. (1999). Evaluating assessment: The missing link? *Medical Education*, 33, 276-281. https://doi.org/10.1046/j.1365-2923.1999.00405.x

Graham, J. M. (2006). Congeneric and (essentially) Tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, 66(6), 930-944. https://doi.org/10.1177/0013164406288165

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131-153). Mahwah, NJ: Lawrence Erlbaum Associates.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). I-S-T 2000 R -

Intelligenz-Struktur-Test 2000 R (2nd ed.). Göttingen: Hogrefe.

Malau-Aduli, B. S., Walls, J., & Zimitat, C. (2012). Validity, reliability and equivalence of parallel examinations in a university setting. *Creative Education*, 3, 923-930. http://dx.doi.org/10.4236/ce.2012.326140

McCabe, D. L. Butterfield, K. D., & Trevi˜no, L. K. (2006). Academic dishonesty in graduate business programs: prevalence, causes, and proposed action. *Academy of Management Learning and Education*, 5(3), 294–305. https://doi.org/10.5465/amle.2006.22697018

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). New York: American Council on Education and Macmillan.

Nitko, J. A. (2001). *Educational assessment of students*. New Jersey: Prentice Hall.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V., & Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33, 206-214. https://doi.org/10.3109/0142159X.2011.551559

Odongo, D. A., Agyemang, E., & Forkuor, J. (2021). Innovative approaches to cheating: An exploration of examination cheating techniques among tertiary students. *Hindawi Education Research International*, 1, 1-7. https://doi.org/10.1155/2021/6639429

Pallant, J. (2010). *SPSS survival manual. A step by step guide to data analysis using SPSS* (4th ed.). Crow's Nest: Allen & Unwin.

Schmale, H. (2001). Berufseignungstest (BET). Tabellenband (4th revised and enlarged ed.). Bern: Hans Huber.

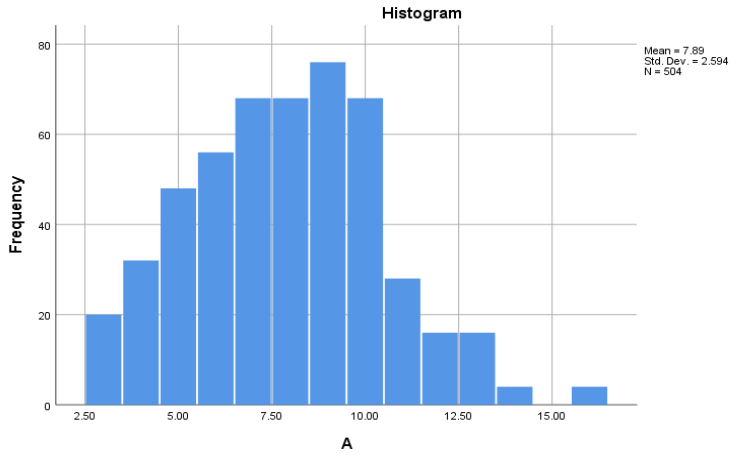
Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., Pangaro, L., Ringsted, C., Swanson, D., Van der Vleuten, C. P. M., & Wagner-Menghin, M. (2011). Research in assessment: Consensus statement and recommendations from Ottawa 2010 Conference. *Medical Teacher*, 33, 224-233. https://doi.org/10.3109/0142159X.2011.551558

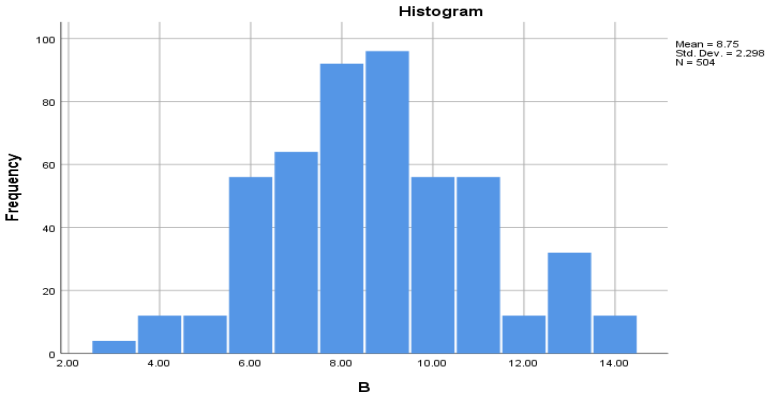
Spaan, M. (2013). Test and item specification development. *Language Assessment Quarterly*, 3(1), 71-79. https://doi.org/10.1207/s15434311laq0301_5

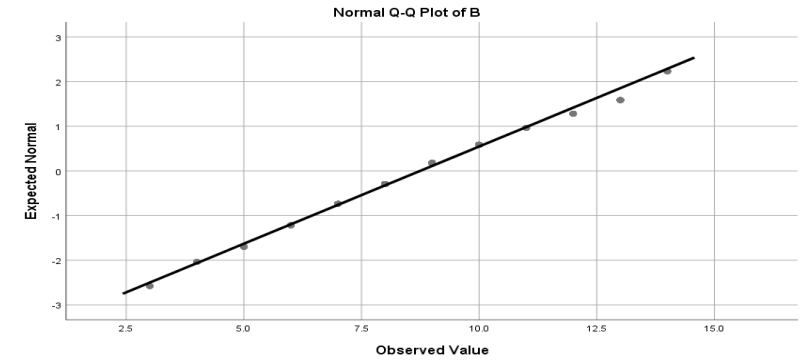
Teixeira, A. A. C., & Rocha, M. F. (2010). Cheating by economics and business undergraduate students: an exploratory international assessment. *Higher Education*, 59(6), 663–701. https://doi.org/10.1007/s10734-009-9274-1.

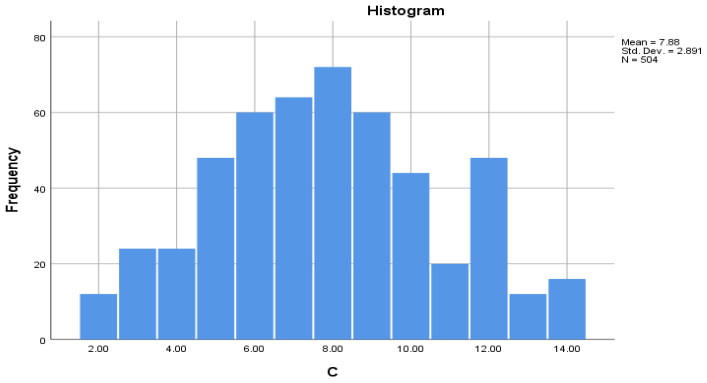
Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: A study using the Rasch model. *ISRN Education*, 1, 1-7. http://dx.doi.org/10.1155/2013/585420

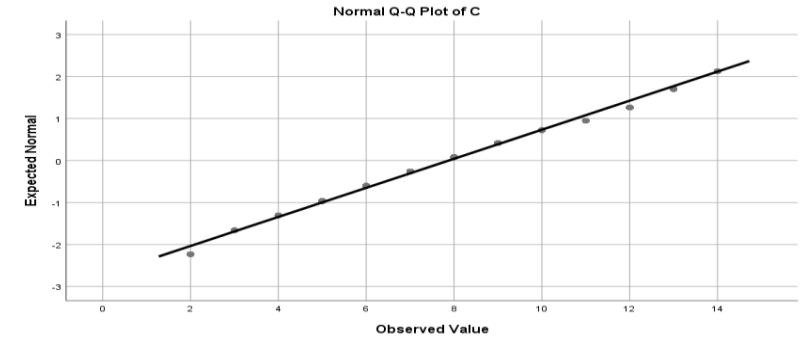
Appendix


Histogram
Mean = 7.89
Std. Dev. = 2.594
N = 504
Frequency
80
60
40
20
0
2.50
5.00
7.50
10.00
12.50
15.00
A


Histogram
Mean = 8.75
Std. Dev. = 2.298
N = 504
Frequency
100
80
60
40
20
0
2.00
4.00
6.00
8.00
10.00
12.00
14.00
B


Normal Q-Q Plot of B
Expected Normal
Observed Value
3
2
1
0
-1
-2
-3
2.5
5.0
7.5
10.0
12.5
15.0


Histogram
Mean = 7.88
Std. Dev. = 2.891
N = 504
Frequency
C
80
60
40
20
0
2.00
4.00
6.00
8.00
10.00
12.00
14.00


Normal Q-Q Plot of C
Expected Normal
Observed Value
3
2
1
0
-1
-2
-3
0
2
4
6
8
10
12
14