European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available online at: www.oapub.org/edu

DOI: 10.46827/ejes.v12i5.5947 Volume 12 | Issue 5 | 2025

PARALLELISM AMONG THREE MATHEMATICS TEST FORMS: ANALYSIS OF STATISTICAL MEASURES

Robert Osman Iddrisu[1i],
Umar Safianu Abukari[2],
Simon Alhassan Iddrisu[3]
[1,3]Department of Educational Foundations,
University for Development Studies,
Tamale, Ghana
[2]Department of Education,
Bagabaga College of Education,
Tamale, Ghana

Abstract:
This study sought to determine the degree of parallelism using three forms of objective-type tests constructed in mathematics and a criterion test in Religious and Moral Education. A total of ninety-three (93) examinees participated in the study. The psychometric measures (characteristics) of these three forms of tests were estimated, compared and used to determine the degree of parallelism of these three forms of tests. Means and equality of variances of the three forms of the test were determined using analysis of variances (ANOVA). We also estimated and compared the covariance of the three test forms, AB, AC, and BC. Finally, we estimated and compared the inter-correlation of each of the three test forms with the criterion test. The results showed that while the variances displayed some degree of equality, the means, covariance, and inter-correlations were not equal. The three test forms are congeneric types of parallel tests. Even though the content and variances of the three test forms were equal, other statistics could not qualify them for any other degree of parallelism.

Keywords: congeneric, parallelism, statistical measures, test forms

1. Introduction

Parallel test forms are the best method of estimating reliability. Even though constructing two parallel tests is the most common approach, there is a way to make sure that the forms are parallel, not just in terms of variances and means, but also in terms of correlations using three parallel tests (Gullicksen, 2013). Gullicksen elaborated further by saying that all the metrics—means, variances, intercorrelations, and reliabilities—are the

[i] Correspondence: email robert.iddrisu@uds.edu.gh

same in parallel tests. In addition, the test forms have approximately equal validity for any given criterion. They also include items that originate from the same content domain, have the same item format, and meet the statistical criteria for parallel testing. In parallel testing, each set of test cases is executed independently of the others. The results of each test run are then compared to identify any differences in the behaviour of the application under test.

Psychological and statistical criteria broadly categorise the criteria for examining the parallelism of tests (Chadha, 2012). The first category uses the sameness of the validity, subject matter, and format, while the latter describes equality in the means, variances, and covariances. Concerning the different parallel forms of tests, Chadha (2012) identified five types of parallel forms. Thus, Chadha (2012) identified five types of parallel forms: classical parallel forms, essentially classical parallel forms, tau-equivalent forms, essentially tau-equivalent forms, and congeneric forms.

1.1 Classical Parallel Forms

The process of constructing classical parallel forms of tests involves creating two or more versions of a test that are designed to be equivalent in terms of their content and difficulty level (Chadha, 2012). This procedure ensures that test takers are not unfairly advantaged or disadvantaged by the test's content or format. To create these parallel forms of the test, Gierl *et al.* (2017) explained that the test items are carefully developed and selected to have similar content and difficulty levels across the different versions. They added that the test takers are then randomly assigned to different versions of the test, with each version containing an equal number of items from the content area. This theory implies that in administering this parallel form of the test, test takers are randomly assigned to one of the versions to ensure that they cannot predict which version they will receive. This technique helps to reduce the likelihood of cheating or bias in the test results.

After the tests are administered and scored, the scores from the different versions are compared to determine if they are statistically equivalent. The statistical equivalent focuses on the equality or approximate equality of the true scores, means, standard deviations, and variances of the three forms of the test; the covariance of the pairs of the three forms of the test; and the criterion, as well as the equality of the inter-correlation of each of the forms and a criterion test. If these statistical properties are equivalent or equal, it is concluded that these tests provide evidence that the different versions of the test measure the same construct and can be used interchangeably to assess individuals or groups and therefore are said to be classical parallel tests.

1.2 Essentially Classical Parallel Forms

Chajewski (2023) explained that essentially, classical parallel forms are a variation of the classically parallel forms in that these versions of the test are not created entirely from scratch but are instead derived from a common pool of test items. It has similar statistical characteristics to the classically parallel forms, except that the means of these tests are not equal. However, the difference between this form and the classically parallel forms is that the true score of one form is equal to the sum of the true score of the form and a constant

(Ti = Tg + Cgh) where Cgh ≠ 0 (Chajewski, 2023). In this approach, a large pool of test items is developed and pretested to determine their difficulty levels and their ability to discriminate between high- and low-performing individuals. Two or more parallel test forms are created from this pool by selecting items with similar content and difficulty. The essentially classical parallel forms method can be a useful approach to test development, provided that appropriate steps are taken to ensure the validity and reliability of the resulting tests.

1.3 Tau-Equivalent Forms

The tau-equivalent parallel test is another parallel test form whose construction involves creating two or more versions of a test that are designed to measure the same construct or skill but with different sets of test items (Sun *et al.*, 2008). They added that these forms are also designed to be statistically equivalent in terms of all measurement properties except inequality of error variances ($\hat{\sigma}^2_1 \neq \hat{\sigma}^2_2 \neq \hat{\sigma}^2_3$). In this approach, the different test versions are not necessarily parallel in terms of item content or difficulty level. To create *tau*-equivalent forms of a test, a large pool of test items is developed and pretested to determine their psychometric properties, such as difficulty and item discrimination levels. From this pool, two or more test versions are constructed by selecting different sets of items that have been found to have similar measurement properties.

Tau-equivalent forms of tests can be useful in situations where it is difficult or impractical to create parallel forms of a test, such as in the case of performance-based assessments or open-ended essay questions. However, tau-equivalent forms may not be fully interchangeable, as the test versions may differ in item content or difficulty.

1.4 Essentially Tau-Equivalent Forms

Essentially tau-equivalent forms are a method of test construction that involves creating two or more versions of a test that are designed to measure the same construct but with different sets of test items that are not necessarily identical or parallel in content or difficulty level (Chajewski, 2023). In essentially tau-equivalent forms, a large pool of test items is developed and pretested to determine their psychometric properties, such as difficulty level and item discrimination. The major differences between the Essentially Tau-equivalent and the Tau-equivalent are that, apart from the inequalities of errors of variances in Tau-equivalent forms, there are also inequalities of means and the true score of one form is equal to the sum of the true score of the form and a constant ($T_i = T_g + C_{gh}$) where $C_{gh} \neq 0$ of the forms of the test are also not equal (Sun *et al.*, 2008). From this pool, two or more test versions are constructed by selecting different sets of items that have been found to have similar measurement properties.

The Essentially Tau-equivalent parallel forms allow for more flexibility in test construction, as it does not require strict item equivalence or parallelism. This can be particularly useful when testing different domains or skills that cannot be easily assessed with identical items. However, it is important to ensure that the different versions of the

test are designed to measure the same construct and that their psychometric properties are equivalent, as failure to do so can result in biased or invalid test scores.

1.5 Congeneric Forms

Congeneric forms or parts of a test refer to two or more versions of a test that measure the same construct but with different sets of test items that are not equal or not approximately equal in their means, standard deviations, and variances of the three forms of the test, the covariance of the pairs of the three forms of the test ($\sigma_{1,2} = \sigma_{1,3} = \sigma_{2,3}$), the covariance of each of the three pairs of tests and the criterion, as well as the inequality of the intercorrelations of each of the forms and a criterion test (Kartowagiran *et al.*, 2019). The description suggests that the similar versions of the test might not be the same in content or difficulty because of the differences in their statistical properties. It is said that, among the different forms of parallel test forms, congeneric forms or parts seem to be the test forms that are easier to achieve or meet their conditions.

This study aimed to see if the ideas about parallel test forms apply to three types of objective mathematics test items created by teachers and to determine how similar these test forms are to each other.

2. Purpose of This Study

This study sought to:

1) Estimate the content relationships, means, variances, covariance, and intercorrelations of the three mathematics test forms.
2) Examine the degree of parallelism among the three forms of multiple-choice items in the objective-type mathematics test.

2.1 Research Questions

The following research questions guided the study:

1) What are the estimates of the means, variances, covariances, and intercorrelations of the three mathematics test forms?
2) To what extent do the three forms of multiple-choice items in the objective-type mathematics test exhibit parallelism?

2.2 Significance of the Study

This study is important because it provides teachers and other test developers with a deeper understanding of test parallelism in educational assessments, especially in the context of objective-type mathematics tests. The research further sheds light on how consistent and comparable these tests are. The findings emphasise the importance of having equal content and variance across test forms while also highlighting that other statistical factors, such as means, covariance, and intercorrelations, are key to truly understanding parallelism. For teachers and test developers, this research stresses the

need for careful psychometric evaluation when designing and comparing test versions, ensuring that assessments are both fair and reliable.

3. Material and Methods

3.1 Construction of Test Items

Multiple choice items were constructed and used for this study. The construction process began with the determination of the purpose of the assessment; thus, the study's purpose was to examine the degree of parallelism of the three forms of the test. As a result, relevant content areas in Mathematics, and Religious and Moral Education were selected based on the grade level (class) and age of the students. Junior High School Students (JHS): three (3) students were selected for the assessment. The content areas that were selected for this study were carefully selected among those topics that students at that level had already been taught. Items were crafted on twelve (12) topics confirmed as areas that students were taught. These content areas formed the basis for which the table of specifications was constructed, as presented in Table 1.

Table 1: Table of Specification

	Instructional Objectives				
Content	**Knowledge**	**Comprehension**	**Application**	**Evaluation**	**Total**
Equations		1			1
Sets	1	1		1	3
Rational Numbers	1				1
Transformation	1	1	1	1	4
Algebraic		1			1
Inequalities		1			1
Profit and Loss	1		2		3
Ratio and Proportion					1
Indices	1				1
Central Tendency	1				1
Vectors	1				1
Construction	2				2
Total	9	5	4	2	**20**

Source: Researchers' construct, 2023

Table 1, which is the table of specifications, is a two-dimensional table that relates the levels of instructional objectives to the course content. The course content topics are arranged vertically to the left (in rows), and the instructional objectives are arranged on the top (in columns). We used it to establish content validity, a measure of how representative a test's scores are across all learning domains. Table 1 illustrates how we constructed a total of twenty items from selected topics, each representing a different thinking process.

3.2 Content Similarities of Test Forms (a, b and c)

The three test forms' content similarity is a key factor in any parallel test forms. The test's nature and purpose determine the content similarities of parallel test forms. In a mathematics test, for example, the content of parallel forms should be similar in terms of the types of mathematical problems tested, the difficulty level of the problems, and the mathematical concepts covered. Kumar *et al.* (2021) explained that to ensure content similarity, test developers typically use a variety of strategies, such as item mapping, expert review, and statistical analysis, to compare the content of the parallel forms. For example, item mapping involves identifying the specific content covered in each item on each form and then comparing the other sets. Item mapping was one way to find content similarities between the three test types. Each item on form A was compared to their corresponding items in forms B and C. For instance, the first question (item 1) in each of the forms A, B, and C was based on equations, and they all shared the same item objective (refer to Table of Specifications, Table 1).

3.3 Test Administration

The tests were administered to a total of ninety-three (93) Junior High School students. Alternative test forms (A, B & C) were administered to the same group of students on the same occasion and time in spiral form. In all, 31 students responded to each test form (A, B and C). However, all the students (93) responded to the criterion test (Form Y).

3.4 Scoring the Test

The various test forms (A, B, C & Y) were hand-scored. The items were dichotomously scored (0, 1) since the items were objective-type tests, namely, multiple choice. Correct answers were marked 1, while wrong answers were marked 0. In all, the total score (performance) of a student was the sum or accumulation of the ones (correct answers) on each test.

4. Results and Discussion

4.1 Item Difficulties

The percentage or proportion of students who answered each question correctly on the test determined the item difficulty level. We determine the item difficulty (p-value) level by dividing the total number of examinees by the number of students who correctly answer the item. R is the number of students who got the question right, and T is the total number of examinees, so mathematically speaking, P = R/T. Between 0.0 (no student got the question right) and 1.0 (when all students answered the item correctly) is the p-index. Perhaps item difficulty should have been named “item easiness”. This evidence indicates that the smaller the p index, the more difficult the item, and the greater the p index, the less difficult the item. It should be noted that item difficulty is calculated for each item. The mean item difficulty of a test can be calculated as the difficulty of the entire test. In this study, the item difficulties of the three forms of the test were determined using the Test Analysis Program (TAP). To be able to compare the item difficulties of the three

forms, the mean item difficulties of the forms were the basis of the comparison, as shown in Table 2.

Table 2: Item Difficulties of Forms A, B and C

Test Forms	Mean Item Difficulties
Form A	0.458
Form B	0.476
Form C	0.595

Source: Test Analysis Program version 19.1.4, 2018.

As shown in Table 2, test form A appeared to be more difficult, followed by test form B, with test form C being less difficult among the three forms. The mean item difficulties of the three forms of the test are, however, considered moderate and are closed. This indicates the similarity of item difficulties among the three forms of the test.

4.2 Means

The mean is a measure of central tendency, which represents the average of a distribution. Although, technically, we call this statistic the arithmetic mean, it is what most people call the average (Cowles, 1986). To compute a mean, add up all the scores and then divide by the number of scores you added. The mean is best used with interval or ratio scales, and it usually predicts an individual's score. Scores on a test act as if all the scores were the mean score and can, therefore, be predicted scores every time. In this study, the mean scores of the three forms of mathematical tests were considered because mean scores of test forms are key determinants of test parallelism. Chadha (2012) suggests that we check the t-values of MA and MB, MA and MC, and MB and MC to determine the equality of means. If all three t-values are non-significant statistically, then we can conclude the equality of the three means of the forms. Table 3 presents the results.

Table 3: Summary Statistics

Groups	Count	Sum	Average	Variance
Form A	31	308	9.935484	8.262366
Form B	31	290	9.354839	9.703226
Form C	31	371	11.96774	10.16559

Source: Field data, 2023.

In Table 3, the results indicated that there were marginal differences in the mean performance of the various forms of tests taken. We conducted further analysis to determine the significance of the observed differences. ANOVA was conducted to check the equality of means in the three forms of the test. It was considered appropriate for this because the mean scores of three independent groups on the mathematics test (dependent variable) were compared. The scores of pupils are measured on the ratio scale of measurement. Table 4 presents the output.

Table 4: ANOVA

Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	116.7097	2	58.35484	6.223148	0.002942	3.097698
Within Groups	843.9355	90	9.377061			
Total	960.6452	92				

Source: Field data, 2023.

As shown in the ANOVA Table 4, the p-value of 0.002942 ($p < 0.05$) is significant, indicating a difference in the means of the three forms of the test. In addition, the F statistic of 6.223 is greater than the F critical value of 3.098, as shown in ANOVA Table 4. This value indicates a difference in the mean scores of the three forms of the test. For this study, no further test (post hoc) was conducted to determine the exact location of the difference in means among the test forms since this study was only interested in establishing differences in means or otherwise. The test forms, therefore, produced means that are not equal; thus, MA ≠ MB ≠ MC.

4.3 Variance

In statistics and hypothesis testing, variance refers to the measure of the spread or dispersion of a set of data points around their mean or average value (Asiret & Sunbul 2016). In other words, variance is a numerical value that describes how much the individual values in a dataset differ from the overall mean. Variance is key in testing because it shows how much variability there is in a test's results. High variance means the results are very variable, making it harder to draw conclusions or find patterns. However, Walter and Rose (2013) explained that if the variance is low, it indicates that the data points are more tightly clustered around the mean, which can make it easier to identify patterns and draw conclusions from the results of the test.

Mathematically, the sample variance is the average of the squared deviations of scores around the sample mean. Variance can be computed by finding the average squared deviation. Using Microsoft Excel, the study's ANOVA equality of means procedure calculated the variance of the three test forms (A, B, and C) to see if they were equal or nearly equal. This was necessary in the determination of the degree of parallelism. The variance of the test forms is shown in the summary statistics in Table 5.

Table 5: Summary Statistics

Groups	Count	Sum	Average	Variance
Form A	31	308	9.935484	8.262366
Form B	31	290	9.354839	9.703226
Form C	31	371	11.96774	10.16559

Source: Field data, 2023.

From the summary statistics in Table 5, there are different variances for the three forms of the test as a form. Form A produced a variance of 8.26; Form B produced a variance of 9.70; and Form C produced a variance of 10.17. Even though the values are close to each other, the variances of the three forms are not physically equal.

However, the difference in the variances, as shown in Table 5, was not enough to conclude on differences between variances without statistical significance. Therefore, we conducted Levine's test of homogeneity of variances using SPSS software. Table 6 displays the output.

Table 6: Test of Homogeneity of Variances

Levine' Statistic	df1	df2	Sig.
.236	2	90	.791

The sig value of 0.791 in Table 6 confirms the assumption of homogeneity of variance. This means that there are equal variances among the three forms of tests. Levine's test of homogeneity of variances concludes that the variances of the three test forms are approximately equal.

4.4 Covariance of Statistical Measures

In statistics, covariance is a measure of the relationship between two variables. In testing, the covariance of two test score sets is the extent to which the scores are related or vary together. More specifically, the covariance of two sets of test scores is a measure of how much the scores from one set of tests tend to increase or decrease as the scores from the other set of tests increase or decrease (Lazarev & Khaybullin, 2017). If the covariance is positive, it means that the scores tend to move together in the same direction (i.e., as one set of scores increases, the other set tends to increase as well). They added that a negative covariance means the scores tend to move in opposite directions (i.e., as one set of scores increases, the other set tends to decrease). A high covariance indicates that the sets of test scores are closely related, while a low covariance indicates that they are not closely related.

In testing, the covariance of two sets of test scores can be used to assess the degree to which the two tests measure the same construct or to evaluate the reliability and validity of the test scores. In this study, the covariance of each of the test forms (A, B and C) on the Mathematics test and a criterion (Form Y) on Religious and Moral Education were calculated (using Microsoft Excel). The following are the outcomes and interpretations of these covariances: AY, BY and CY.

4.5 Covariance of Forms A and Y (AY)

Form A of the Mathematical test and form Y (criterion) produce a covariance of 0.706. This value indicates a positive relationship between the two sets of scores. A positive covariance refers to a situation in which two variables tend to vary together in the same direction. Specifically, a positive covariance means that as one variable increases, the other variable tends to increase as well. Conversely, as one variable decreases, the other variable tends to decrease as well. For example, this study shows that when test scores on Form A increase, test scores on Form B also increase. The positive covariance between the two forms again means that most students who score high on test form A also scored

high on the criterion test Y, and most students who score low on test form A also scored low on the criterion test Y.

It is important to note, however, that a positive covariance does not necessarily indicate a causal relationship between the two variables (Kamat & Nandi 2017). In other words, just because two variables have a positive covariance does not mean that one variable causes the other. Instead, the positive covariance simply indicates that there is a tendency for the two variables to vary together in a certain way. The covariance's positivity indicates that the two score sets tend to vary together.

To fully understand the relationship between the two sets of scores, the interpretation of the covariance should be considered with other statistical measures, like correlation coefficients. The correlation coefficient of test scores A and Y in this study was also estimated at 0.1069, indicating a positive but weak correlation. This confirms the interpretation of the relationship between the test scores between Forms A and Y.

4.6 Covariance of Forms B and Y (BY)

A covariance of 1.569 indicated a positive linear relationship between the two sets of scores. This result is an indication that as scores on Form B increased, scores on the criterion tended to increase as well. However, the covariance by itself does not indicate the strength of the relationship between the scores on Form B and the scores on the criterion. The correlation coefficient of the scores on form B and the criterion was used to assess the strength of the linear relationship between the two sets of scores. We viewed the relationship between the two sets of test scores as weakly moderate, as indicated by a correlation coefficient of 0.395.

4.7 Covariance of Forms C and Y (CY)

The results from the analysis of form C and the criterion Y indicate a negative covariance of -0.411. This negative covariance between form C and the criterion Y indicates an inverse relationship between them. In other words, the data show that students' scores on the criterion rose, but Form C scores fell. Furthermore, Field (2015) explained that a negative covariance indicated that as one variable deviates from the mean (e.g., increases), the other deviates from the mean in the opposite direction (e.g., decreases). Field's explanation is manifested in this study; thus, as scores on Form Y (criterion) deviate from the mean as an increment, scores on Form C (Mathematic test) deviate from the mean in the opposite direction as a decrease. This negative covariance suggests that there is a trade-off between Religious and Moral Education (the criterion) and Mathematics abilities for these students. Thus, students who performed well on the religious and moral education test struggled on the mathematics test, as shown in the test scores for Form C and Form Y.

It is important to note that negative covariance does not necessarily imply causation. In this study, the negative covariance between Religious and Moral Education and Mathematics scores does not mean that the former causes the latter to drop. It simply means that there is a statistical relationship between the two sets of scores that suggests an inverse association.

4.8 Comparison of the Covariance

In this study, we compared the three covariances to determine if they were equal or approximately equal. This comparison is necessary in the determination of the degree of parallelism. Table 7 displays the corresponding covariances.

Table 7: Covariance of Test Forms and the Criterion

	Form Y	**Form Y**	**Form Y**
Form A	0.706		
Form B		1.569	
Form C			-0.411

Source: Field study 2023.

As shown in Table 7, the covariances of AY, BY, and CY are all not equal. This result strongly indicates the extent of parallelism among these test items. We again used Microsoft Excel to compare the covariance of the three test forms (AB, AC, and BC). Table 8 presents the output.

Table 8: Covariance of Test Forms

Test forms	**Form A**	**Form B**	**Form C**
Form A		1.6237	0.6312
Form B	1.6237		3.9116
Form C	0.6312	3.9116	

Source: Field study 2023.

A statistical measure known as covariance illustrates the relationship between two variables. As shown in Table 8, the covariance of Forms A and B (1.62) suggests that as one set of scores increases, the other set of scores tends to increase as well. Again, the covariance of 0.63 between the two forms (A and C) of scores also indicates a positive linear relationship between the two sets of scores, even though the variance is less than one, and finally, the covariance of 3.91 between the two forms (B and C) of scores also indicates a positive linear relationship between the two sets of scores. This is the highest covariance recorded among the covariances in the test forms. In comparing the three covariances, Table 8 shows that they are simply not equal. We can illustrate the covariance as AC > AB > BC. This means that AC ≠ AB ≠ BC.

4.8 Correlation of Test Forms with the Criterion

The correlation of each of the forms of the test and the criterion was taken into consideration in the determination of the degree of parallelism. Each of the correlations of the test forms and the criterion (r_{AY}, r_{BY} and r_{CY}) were determined using Microsoft Excel. A correlation coefficient measures the strength and direction of the linear relationship between two variables (Aşiret & Sünbül, 2016). The value of 0 indicates no relationship, while a value of 1 indicates a perfect positive relationship, and a value of -1 indicates a perfect negative relationship. A value between 0 and 1 (or 0 and -1) indicates the strength of the relationship, with values closer to 0 indicating a weaker relationship and values

closer to 1 (or -1) indicating a stronger relationship. The correlations of the forms of the test and the criterion are shown in Table 9.

Table 9: Correlation of Test Forms and the Criterion

Test Forms	**Form Y**
Form A	0.1069
Form B	0.3952
Form C	-0.0612

Test forms A and Y produced correlation coefficients of 0.107. This coefficient indicates a weak positive linear relationship between the two sets of scores. This means that as one set of scores increases, the other set of scores tends to increase as well, but the relationship is not very strong. The weak correlation coefficient does not necessarily mean that the relationship between the two variables is irrelevant. Again, test Forms B and Y produced correlation coefficients of 0.395. This represented a moderate positive linear relationship between the two sets of scores. This means that as one set of scores increased, the other set of scores tended to increase as well. We can describe this relationship as moderate; it is neither very strong nor very weak.

The correlation of test Forms C and Y produced a correlation coefficient of -0.0612. This coefficient represented a weak negative linear relationship between the scores on the two test forms. This means that as one set of scores increased, the other set of scores tended to decrease. This relationship is not only a weak one but an inverse one as well. The three test forms' correlation coefficients show they are unequal.

5. Findings

The study aimed to assess the parallelism of three test forms based on the content area similarity, the equality or inequality of the means, variances, and covariance of each test form, the covariance of the test forms, the criterion, and the correlation of the test forms with the criterion test. The content area similarity was determined by checking item mapping, item difficulties and the use of the test blueprint. Following the analysis of the data obtained from the administration of these test forms to students, the following findings were revealed:

1) Item mapping on the three forms of the test, similarities in the mean item difficulty indices of the tests and the expert review of these test forms suggested that they were similar in content.
2) The use of an ANOVA indicated a significant difference in the mean scores of the three test forms, leading to the conclusion that they were not equal.
3) Levene's test showed no significant difference, which means we can assume that the variances of the three test forms are equal (Sig > 0.05). Thus, the variances of the three forms were approximately equal.
4) The study found unequal covariance between the three forms of the test and the criterion test.

5) It also found unequal covariance among the pairs in the three forms of the test.
6) We estimated unequal correlation coefficients for each test form and the criterion test. Thus rAY ≠ rBY ≠ rCY.

The results on the criteria for the three test forms' parallelism matched the Congeneric parallel form's properties. Therefore, we appropriately classified the three test forms (Forms A, B, and C) as cogenetic parallel tests. This classification was backed up by the observation that, besides having similar content, the variances were about the same, while all other properties were different, showing a congeneric level of parallel test forms.

6. Recommendations

The study's findings and conclusions led to the formulation of several recommendations for students, test administrators, and researchers. These recommendations suggest applying parallel tests to help minimise bias and achieve greater fairness in the use of equivalent test forms to select applicants through competitive tests. When test administrators use equivalent test forms or versions of the same test, it won't matter when or which form a candidate takes, as all forms will have the same content and item difficulty. The results should help students grasp parallelism and apply it in their lives.

7. Conclusion

The findings on the criteria used to determine the degree of parallelism of the three test forms conformed with the properties of the Congeneric parallel form. Thus, the three test forms (Forms A, B, and C) were appropriately classified as Congeneric parallel tests. This classification was backed up by the observation that, besides having similar content, their variances were about the same, and all other characteristics were different, showing a congeneric level of parallel test forms. In conclusion, while the three test forms have some similarities, they differ in how effective they are and how they relate to the criterion. These findings indicate the need to improve how tests are constructed and evaluated, ensuring that different forms not only cover the same content but also produce consistent and reliable results. More research is needed to understand the causes of these differences and to make test forms more comparable.

Conflict of Interest Statement
We declare no conflicts of interest.

About the Author(s)

Robert Osman Iddrisu is a lecturer in the Department of Educational Foundations Studies at the University for Development Studies. He has worked at the university for three years, published articles, and presented at national and international academic conferences. His research interests include educational assessment, measurement, educational statistics, and education-related research.

Email: robert.iddrisu@uds.edu.gh

Umar Safianu Abukari is a third-year PhD candidate studying Measurement and Evaluation in the Department of Education and Psychology at the University of Cape Coast, Ghana. His research interests include educational assessment and its impact on teaching and learning, with a special focus on educational measurement, instrument validation (test psychometrics), and the development of instruments to effectively measure 21st-century skills, including critical thinking, collaboration, and problem-solving.

Email: abukarisafianu@gmail.com

Dr. Simon Alhassan Iddrisu is a senior lecturer in the Department of Educational Foundations Studies at the University for Development Studies. He has been working at the university for 15 years. Dr. Iddrisu has presented at both national and international research conferences and has published more than ten research articles in peer-reviewed journals. His research interests include educational assessment, statistics, measurements, instrument validation, and education-related research.

Email: aisiimon@uds.edu.gh

References

Asiret, S., & Sunbul, S. O. (2016). Investigating test equating methods in small samples through various factors. *Educational Sciences: Theory and Practice, 16*(2), 647-668. Retrieved from https://files.eric.ed.gov/fulltext/EJ1101189.pdf

Chadha, N. K. (2012). *Applied psychometry in applied psychometry*. SAGE Publications India Pvt Ltd. https://doi.org/10.4135/9788132108221

Chajewski, M. (2023). Classical test theory. *International Encyclopedia of Education* (fourth edition), Elsevier, page 51-58 https://doi.org/10.1016/b978-0-12-818630-5.10008-9

Cowles, M. P. (1986). Review of basic statistics for behavioural sciences. In Canadian Psychology /Psychologie canadienne (Vol. 27). https://doi.org/10.1037/h0084344

Field, A. (2015). SPSS 5th. Dk, 53(9), 1689–1699.

Gierl, M., Daniels, L., & Zhang, X. (2017). Creating parallel forms to support on-demand testing for undergraduate students in psychology. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 288-302. https://doi.org/10.21031/epod.305350

Gulliksen, H. (2013). *Theory of mental tests*. Routledge. https://doi.org/10.4324/9780203052150

Kamat, N., & Nandi, A. (2017). A closer look at variance implementations in modern database systems. *ACM SIGMOD Record, 45*(4), 28-33. Retrieved from https://arxiv.org/abs/1509.04349

Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. Research and Evaluation in Education. https://doi.org/10.21831/REID.V5I2.23721

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple-choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India, 77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Lazarev, V. A., & Khaybullin, R. Y. (2017). The study of difficulty and differential ability of competitive items. *Journal Issues Business.* 253-267 https://doi.org/10.17122/OGBUS

Sun, K. T., Chen, Y. J., Tsai, S. Y., & Cheng, C. F. (2008). Creating IRT-based parallel test forms using the genetic algorithm method. *Applied Measurement in Education, 21*(2), 141-161. http://dx.doi.org/10.1080/08957340801926151

Walter, O. B., & Rose, M. (2013). Effect of item order on item calibration and item bank construction for computer adaptive tests. *Psychological Test and Assessment Modeling, 55*(1), 81. Retrieved from https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2013_20130326/05_Walter.pdf