

Reliability Evidence in Mathematics School-Based Assessment (SBA): The Case of Teacher-Made Test in Techiman High School of Ghana

Isaac Owusu-Darko

Department of Mathematics, Science and Computer Studies, SDA College of Education, Asokore-Koforidua, Ghana

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1406000113>

Received: 30 June 2025; Accepted: 04 July 2025; Published: 24 July 2025

Abstract: Examinations are central to student evaluation in schools, and the reliability of teacher-made tests plays a crucial role in ensuring fair assessment. This study investigates the internal consistency and reliability evidence associated with mathematics teacher-made tests in selected senior high schools in the Techiman Municipal of Ghana. Focusing on Cronbach's alpha, inter-item correlation, corrected item-total correlation, and split-half reliability coefficients, the study assessed terminal examination results over three academic terms. Findings revealed moderate to high levels of internal consistency across terms, with Cronbach's alpha ranging from 0.815 to 0.953. These results indicate a generally acceptable level of reliability in school-based assessments (SBA), though variations suggest areas for standardization and improvement. Implications for improving assessment practices and educational policy are discussed.

Key Words: Assessment, school-based assessment, academic performance, Reliability evidence, Cronbach's alpha, terminal examinations, Teacher-made test.

I. Introduction

Reliability refers to the consistency with which an assessment measures student knowledge, skills, and understanding. When a test yields similar results under consistent conditions, it is considered reliable (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). In the educational context, particularly in mathematics, the reliability of assessments plays a significant role in accurately evaluating student performance.

Mathematics is a core subject in Ghanaian secondary schools, with its mastery being essential for national development goals. However, persistent underperformance in mathematics among students in Techiman Senior High Schools has raised concerns. Students often score poorly in classroom tests and demonstrate a lack of enthusiasm towards the subject. This study explores whether the internal consistency of teacher-made assessments contributes to this challenge. The quality of educational assessment has long been a subject of intense discussion in both academic and policy-making circles. Central to this discourse is the concept of reliability, which refers to the consistency of a test or assessment in measuring what it purports to measure. A test is considered reliable if it produces stable and consistent results over repeated applications or across different sets of items measuring the same construct (AERA, APA, & NCME, 2014). In school settings, particularly in subjects such as mathematics, the reliability of assessments plays a critical role in accurately capturing learners' proficiency and guiding instructional decisions (Brookhart, 2013).

Mathematics is universally acknowledged as a foundational subject that supports logical reasoning, problem-solving, and quantitative thinking—skills that are essential for various academic and life pursuits (Kilpatrick, Swafford, & Findell, 2001). In Ghana, mathematics is a core subject at the basic and senior high school levels and plays a pivotal role in the transition to tertiary education. However, persistent challenges in mathematics education, including students' low performance, limited motivation, and test anxiety, have drawn attention to the quality of assessments used in evaluating students (Anamuah-Mensah, 2020).

In Ghanaian senior high schools, school-based assessments (SBAs) constitute an integral part of the overall evaluation framework. Teacher-made tests are frequently used to evaluate student performance during terminal assessments. Yet, questions often arise regarding the validity and reliability of these tests, especially when there is no standardized protocol for test construction, item analysis, or moderation. Without sufficient evidence of reliability, teacher-made assessments may not accurately reflect students' actual competencies, leading to flawed instructional decisions and unfair grading (Kane, 2016).

The issue of reliability in school-based assessment becomes even more pronounced in mathematics due to the hierarchical nature of the subject. Concepts build progressively upon each other, and a failure to accurately assess foundational knowledge can hinder students' grasp of more advanced material (Cai & Lester, 2010). Thus, if teacher-made assessments are not consistently reliable, educators may miss critical gaps in student understanding or provide misleading feedback. Techiman Senior High Schools, like many others across Ghana, implement teacher-developed examinations as part of their continuous assessment strategy. However, anecdotal evidence and preliminary observations indicate inconsistencies in test quality and student scores across different academic terms. This variation raises concerns about whether the assessments possess sufficient internal consistency, especially in a high-stakes subject like mathematics. This study is, therefore, designed to investigate the internal consistency reliability of mathematics assessments used in these schools.

In educational measurement theory, internal consistency is typically evaluated using statistics such as Cronbach's alpha, inter-item correlation, and split-half reliability (Tavakol & Dennick, 2011). Cronbach's alpha, in particular, provides an estimate of the extent to which test items measure the same construct, with values above 0.7 generally considered acceptable for low-stakes assessments (George & Mallery, 2019). Inter-item correlation provides insight into the degree of relationship among items, while the Spearman-Brown and Guttman split-half coefficients further assess the stability of composite scores derived from test halves (Feldt & Brennan, 1989).

Although Ghana's National Council for Curriculum and Assessment (NaCCA) provides a standard-based curriculum that emphasizes assessment for learning, the implementation of effective and reliable assessments largely depends on the teacher's skill and training in test development (NaCCA, 2019). Previous studies have highlighted that many teachers lack adequate training in psychometric principles, leading to assessments that may be misaligned with curriculum standards or lack internal coherence (Mensah & Frimpong, 2021).

The present study addresses these concerns by systematically examining the reliability evidence from teacher-made mathematics assessments administered during the 2020/2021 academic year in selected schools within Techiman Municipality. By analyzing Cronbach's alpha coefficients and other reliability indices across different academic terms, the study aims to provide empirical insight into the quality and consistency of internal school assessments. This, in turn, can inform future practices in test development and teacher training within the Ghanaian education system.

II. Literature Review

Reliability is generally described in terms of score 'stability.' The Standards for Educational and Psychological Testing (1999) defines reliability as "the consistency of [such] measurements when the testing procedure is repeated on a population of individuals or groups" (p. 25). Reliability typically refers to the measurement error that is introduced into the "entire measurement process" (p. 27) and both limits the degree to which generalizations can be made beyond the specific testing event and quantifies the confidence that can be held in the value assigned to any performance. "Reliability data ultimately bear on the repeatability of the behavior elicited by the test and the consistency of the resultant scores" (p. 31).

Internal consistency is a measure of reliability. Reliability refers to the extent to which a measure yields the same number or score each time it is administered, all other things being equal (Hays & Revicki, 2005). Classical test theory views an observed response as consisting of the sum of true score and error (Nunnally & Bernstein, 1994). True score for an individual is assumed to be invariant on repeated measurements. However, two parallel measurements will yield nonidentical observed scores as a result of random error of measurement. Because random errors are normally distributed, the true score will be located at the mean of the distribution of parallel measures.

Conceptualizing Reliability in Educational Assessment

Reliability in educational assessment is a cornerstone of effective measurement, ensuring that results derived from tests are repeatable, dependable, and minimally affected by measurement errors (Crocker & Algina, 2008). A reliable test allows educators and stakeholders to make informed decisions based on evidence that genuinely reflects a learner's ability or knowledge. In the context of classroom-based assessments, reliability ensures that scores are stable across time and conditions, and that any observed differences in student performance are attributable to actual learning differences, not flaws in the instrument (AERA, APA, & NCME, 2014).

There are multiple types of reliability in assessment, including test-retest reliability, inter-rater reliability, and internal consistency. Test-retest reliability examines stability over time; inter-rater reliability assesses consistency between different scorers; and internal consistency focuses on the coherence of items within a single test. Each form plays a unique role, but internal consistency is particularly critical for teacher-made tests, which are often not externally validated (Nitko & Brookhart, 2014).

Reliability also has implications for fairness and equity in education. Inconsistent or unreliable assessments can perpetuate biases or create misleading narratives about student capabilities. Consequently, reliability is not merely a technical issue—it is a matter of educational justice. Ensuring that assessments reliably reflect what students know and can do is central to ethical and effective teaching (Stiggins, 2005).

Internal Consistency as a Measure of Reliability.

Internal consistency measures how well the items on a test work together to assess the same underlying construct. It is a particularly important measure when the goal is to combine multiple item scores into a single composite score, such as a mathematics test score derived from various word problems, equations, and algebraic expressions (Nunnally & Bernstein, 1994).

Cronbach's alpha (α), introduced by Cronbach (1951), remains the most commonly used statistic for estimating internal consistency. It is calculated based on the average inter-item correlation and the number of items in the test. An alpha coefficient above 0.7 is generally considered acceptable for classroom-level assessments, with 0.8 or higher preferred in high-stakes or summative evaluations (George & Mallery, 2019). However, very high alpha values (above 0.95) may suggest item redundancy, which can compromise test efficiency (Tavakol & Dennick, 2011).

Despite its widespread use, Cronbach's alpha makes several assumptions, including tau-equivalence (equal true score variances for all items), which may not hold in practice. In such cases, additional indicators like McDonald's omega or the split-half reliability coefficient may offer more nuanced perspectives (Revelle & Zinbarg, 2009). Nevertheless, alpha remains a valuable initial metric, particularly when assessing classroom-based tests that may lack the psychometric rigor of standardized instruments.

Internal consistency also reveals much about test quality during post-administration analysis. Low inter-item correlations may indicate that some items are not aligned with the central construct, potentially confusing students or diluting the overall score's interpretive value. For mathematics assessments, internal consistency can be especially telling, given that conceptual coherence is essential for accurate measurement of skills such as problem-solving and logical reasoning (Cai & Lester, 2010).

In this study, the internal consistency of teacher-made mathematics assessments is explored using Cronbach's alpha across multiple academic terms. This will provide insight into the stability and quality of the assessments used in Ghanaian Senior High Schools (SHS) and inform future directions for test development and teacher training.

Complementary Indices of Internal Consistency

While Cronbach's alpha remains a foundational tool for reliability estimation, it is not without limitations. Alpha assumes tau-equivalence, meaning all test items are equally reliable and measure the same latent trait to the same extent. When this assumption is violated, alpha can misrepresent the actual internal consistency of the instrument (Sijtsma, 2009). To address this, alternative indices have been developed to provide a more accurate picture of internal coherence.

The split-half reliability method, for instance, divides the test into two equivalent halves and measures the correlation between the two sets of scores. The Spearman-Brown prophecy formula is then applied to estimate the reliability of the full-length test (Feldt & Brennan, 1989). This method is particularly useful when test length is a concern or when estimating how reliability might change with longer or shorter tests.

Another alternative is McDonald's omega coefficient, which offers advantages over alpha in contexts involving multidimensional constructs or items with varying factor loadings (McDonald, 1999). Omega estimates reliability based on factor analysis, providing insights into how well the items collectively measure a general factor. This makes it especially suitable for educational contexts where multiple interrelated subskills, such as reasoning and computation in mathematics, are assessed.

In addition, inter-item correlation matrices serve as useful diagnostic tools for detecting items that poorly align with others. Ideally, items should have moderate positive correlations (typically between 0.2 and 0.5), suggesting they measure the same construct without being redundant (Pallant, 2020). Extremely low correlations might indicate ambiguity or misalignment, while very high correlations can signal redundancy.

Taken together, these complementary measures enrich the understanding of test reliability and offer practical strategies for refining assessments. Teachers and test developers in Ghana and similar contexts should be encouraged to triangulate reliability evidence using multiple indices to ensure accurate and meaningful assessment outcomes.

Unidimensionality and Factor Saturation in Assessment

Internal consistency metrics like alpha and omega are grounded in the assumption of unidimensionality—the notion that all test items reflect a single underlying construct. When this assumption is violated, reliability estimates may be misleading. For instance, a mathematics test that blends algebraic reasoning with geometry and data interpretation without structural clarity may yield inflated alpha values while masking multidimensionality (Revelle & Zinbarg, 2009).

To evaluate unidimensionality, researchers often employ exploratory and confirmatory factor analysis to assess the degree of common variance explained by a general factor. One such approach is Explained Common Variance (ECV), which quantifies the proportion of variance attributable to a general factor relative to all extracted factors (Ten Berge & Sočan, 2004). A high ECV (typically >0.70) suggests the test is sufficiently unidimensional to justify using composite scores (Sijtsma, 2009). Another useful technique is the use of hierarchical omega (ω_h), which quantifies the proportion of total score variance attributable to a general factor while accounting for the presence of group factors (Revelle, 1979). When ω_h is high, it indicates that most of the variance in test scores is driven by a single trait, reinforcing the validity of composite score interpretation.

In educational testing contexts, especially in mathematics, unidimensionality is desirable but not always feasible. Different subdomains, such as arithmetic, algebra, and statistics, may each require separate reliability estimation if the goal is diagnostic rather than summative evaluation. Therefore, unidimensionality checks are not only psychometrically prudent but also pedagogically strategic. Overall, assessing unidimensionality and factor saturation strengthens the interpretive validity of reliability coefficients and guides appropriate scoring and reporting practices in classroom-based assessments.

Teacher-Made Tests and Psychometric Concerns

Teacher-made tests remain a critical component of school-based assessment practices, particularly in settings where standardized testing infrastructure is limited or absent. These assessments offer contextual relevance, alignment with day-to-day instructional objectives, and flexibility in test administration (Brookhart, 2004). However, they often present psychometric challenges that can

undermine their reliability and validity.

In Ghana, where continuous assessment contributes significantly to students' final grades, the reliance on teacher-made tests is even more pronounced. Despite their ubiquity, studies have indicated that many teacher-developed tests lack systematic construction processes, such as blueprinting, pilot testing, and item analysis (Anamuah-Mensah, 2020). As a result, the reliability of such assessments—particularly in technical subjects like mathematics—can be compromised, affecting fair grading and instructional feedback.

Psychometric concerns typically arise from item-level flaws, including ambiguous wording, inconsistent difficulty levels, poor alignment with curriculum standards, and overlapping content domains (Nitko & Brookhart, 2014). These flaws may result in low internal consistency, thereby weakening the assessment's capacity to reflect students' true performance.

Additionally, many teachers may not be adequately trained in principles of educational measurement. For example, concepts such as item-total correlation, discrimination index, and test reliability may be unfamiliar or underutilized in daily practice (Mensah & Frimpong, 2021). Consequently, test construction often becomes a routine task rather than a reflective, evidence-based process.

Furthermore, external moderation and peer review of teacher-made assessments are often absent in many Ghanaian schools, making it difficult to identify and correct flaws prior to test administration. This institutional gap contributes to the variability in test quality both within and across schools. Addressing these psychometric concerns calls for intentional capacity building among teachers, including targeted professional development in assessment literacy. Training on item writing, reliability estimation, and test data interpretation can enhance the quality of assessments. Moreover, embedding assessment design principles into teacher education curricula can foster a more sustainable culture of high-quality testing practices.

Therefore, in contexts like Ghana, improving the psychometric soundness of teacher-made tests is not just a technical necessity but a pedagogical imperative. Ensuring reliability in school-based assessment supports accurate diagnosis of learning gaps, equitable grading, and informed instructional planning—all of which are essential for enhancing student achievement in mathematics and beyond.

Specific Objectives

The study has the following as its specific objectives.

- To investigate the nature of reliability evidence in the performance of students selected for cross-sectional terminal examinations in the 2023/2024 academic year..
- To determine the nature of correlation and predictive evidence in the performance of students in selected cross-sectional terms in the 2023/2024 academic year.
- To investigate whether mean differences exist between internal end-of-term assessments for the split-half teacher-made test at the selected cross-sectional terms in the 2023/2024 academic year.

III. Method

A quantitative design was adopted. The sample comprised 65 students from two Arts classes in Techiman Senior High School. Internal assessment scores from three successive academic terms were collected. Teacher-made tests for the selected terms were collected with marking rubrics for study and face validation. Marked scripts and students' recorded scores were then coded into SPSS version 21 for the data analyses. Reliability was assessed using Cronbach's alpha, inter-item correlation matrices, and split-half methods. Descriptive and inferential statistics, including ANOVA and Intraclass Correlation Coefficient (ICC), were used to analyze differences and relationships among test items (Field, 2013).

The population consisted of all senior high school students enrolled in the Arts classes of Techiman Senior High School during the 2020/2021 academic year. A purposive sampling technique was employed to select 65 students from two respective intact classes (General Arts 1 and 2). The selection was based on accessibility, availability of complete termly scores, and the uniformity in the structure of mathematics tests administered by their subject teachers.

The data were derived from teacher-made mathematics tests administered across three consecutive terms (Term 1, Term 2, and Term 3) within the 2023/2024 academic year. These tests formed part of the school's regular end-of-term assessment framework. Scores were recorded for two sections of each term's test, designated as Section A and Section B, reflecting varied content areas and question formats (e.g., structured questions vs. word problems) based on split-half test administration.

After obtaining ethical clearance and institutional permission, raw score data were extracted from school assessment records with the assistance of subject teachers. Anonymity and confidentiality of student identities were ensured. The scores were entered into SPSS version 21 for cleaning, coding, and statistical analysis.

To address the objectives of the study, Data Analysis Techniques (DAT) considered multiple reliability estimation procedures under Cronbach's Alpha assumptions for each term's test to assess internal consistency. Split-Half Reliability was analysed using the

Guttman and Spearman-Brown coefficients; teacher-made tests selected were split into halves (Section A and Section B) to evaluate score agreement. An Inter-Item Correlation Matrix was then generated to observe the correlation between each pair of test items, providing insight into homogeneity. Intraclass Correlation Coefficient (ICC) was used to help determine the consistency of test scores across the terms.

All statistical procedures followed established reliability analysis guidelines (Tavakol & Dennick, 2011; Field, 2013), and interpretation of reliability coefficients was based on conventional cut-off values: ≥ 0.9 (excellent), $0.8-0.89$ (good), $0.7-0.79$ (acceptable), < 0.7 (questionable).

This methodology ensured a comprehensive evaluation of the consistency and coherence of teacher-made assessments in the mathematics domain, offering valuable insights into classroom-level test development practices in Ghana, subject to the requirement of standard-based curriculum (NaCCA, 2019), School-Based Assessment domains prescribed in the mathematics curriculum.

IV. Results and Discussion

The nature of reliability evidence in the performance of students of Techiman SHS Arts A in selected cross-sectional terms in the 2023/2024 academic year is analysed descriptively in Table 1 below

Table 1: Descriptive statistics for selected terms in 2023/2024 academic year

	N	Min	Max	Mean	Std. Dev	Skewness	
						Statistic	Std. Error
Internal Score for Term 1 section A	65	6.00	38.00	26.32	7.04917	-.841	.297
Internal Score for Term 1 Section B	65	12.00	58.00	40.33	9.61489	-.665	.297
Internal Score for Term 2 Section A	65	13.00	38.00	26.58	6.54526	-.045	.297
Internal Score for Term 2 Section B	65	20.00	58.00	37.55	8.28785	.009	.297
Internal Score for Term 3 Section A	65	6.00	39.00	27.64	7.41078	-.576	.297
Internal Score for Term 3 Section B	65	12.00	58.00	40.33	9.61489	-.665	.297

Table 1 presents the descriptive statistics of student scores across three terms, each divided into two sections (A and B). The mean scores ranged from 26.32 to 40.34, with standard deviations between 6.55 and 9.61, indicating moderate variability in student performance. Skewness values showed slight deviations from normality but remained within acceptable limits (± 1), suggesting that the distributions were not heavily skewed.

Cronbach's alpha coefficients were computed to assess the internal consistency of the test items in each term. For Term 1, an alpha of 0.854 indicated good reliability. Term 2 showed the highest alpha of 0.953, suggesting excellent internal consistency, while Term 3 had an alpha of 0.815, again falling within the acceptable-to-good range. These results confirm that the teacher-made assessments, although not standardized, possessed a considerable degree of internal coherence (George & Mallery, 2019).

Split-Half Reliability and Predictive Consistency

Split-half reliability analysis further corroborated the internal consistency findings as seen from Table 2.

Table 2: Reliability Statistics for all Terms

Reliability Statistics				Term 1 Cronbach's Alpha	
Cronbach's Alpha	Part 1	Value	.594	Term1: 0.831	
		N of Items	3 ^a	Term2: 0.825	
	Part 2	Value	.590	Term3: 0.953	
		N of Items	3 ^b	6	
	Total N of Items			2	
Correlation Between Forms			.746	<i>Cronbach's Alpha Based on Standardized Items</i>	
Spearman-Brown Coefficient	Equal Length		.854	Term1: 0.854	
	Unequal Length		.854	Term2: 0.953	

Guttman Split-Half Coefficient		.831	Term3: 0.815
Split half Reliability Statistics for all termly assessment			
Cronbach's Alpha	Part 1	Value	.594
		N of Items	3 ^a
	Part 2	Value	.590
		N of Items	3 ^b
	Total N of Items		6
Correlation Between Forms			.941
Spearman-Brown Coefficient	Equal Length		.969
	Unequal Length		.969
Guttman Split-Half Coefficient			.968

The Spearman-Brown coefficient ranged from 0.854 to 0.969 across terms, and the Guttman split-half coefficient also yielded strong reliability values above 0.8. These findings suggest that the division of test items into halves (Sections A and B) did not significantly affect the reliability of the composite test scores. The inter-item correlation matrices revealed strong positive correlations between items in the same term, especially between sections within the same test. For instance, correlations between Term 1A and 1B, and Term 3A and 3B, exceeded 0.7, indicating that items were measuring similar constructs consistently.

As can be seen on Table 2, the reliability statistics show a correlation between forms of 0.746, Spearman Brown coefficient, equal length of 0.854, unequal length of 0.854, Guttman Split-Half coefficient of 0.831 and a predictive evidence of 0.854 from the Alpha based on standardized items.

From term two, the reliability statistics show a correlation between forms of 0.941, Spearman Brown coefficient, equal length of 0.969 and unequal length of 0.969, Guttman Split Half coefficient of 0.968 and predictive evidence of 0.953 from the Cronbach's alpha based on standardised items. From term three, the reliability statistics show a correlation between forms of 0.941, Spearman Brown coefficient, equal length of 0.969 and unequal length of 0.969, Guttman Split Half coefficient of 0.968 and predictive evidence of 0.815 from the Cronbach's alpha based on standardised items.

Intraclass Correlation and Variance Analysis

The intraclass correlation coefficient (ICC) was calculated to assess the reliability of scores across the three terms. An ICC of 0.831 ($p < 0.01$) with a 95% confidence interval of [0.723, 0.897] indicated a high level of agreement in the repeated measures across terms. This suggests that the student performance data were stable over time, reinforcing the consistency of the teacher-made assessments as seen from Table 6

Inter-Item Correlation Matrix

	Internal Score for Term 1 section A	Internal Score for Term 1 Section B	Internal Score for Term 2 Section A	Internal Score for Term 2 Section B	Internal Score for Term 3 Section A	Internal Score for Term 3 Section B
Internal Score for Term 1 section A	1.000	.746	.010	-.059	.708	.746
Internal Score for Term 1 Section B	.746	1.000	.154	.056	.941	1.000
Internal Score for Term 2 Section A	.010	.154	1.000	.895	.092	.154
Internal Score for Term 2 Section B	-.059	.056	.895	1.000	-.024	.056
Internal Score for Term 3 Section A	.708	.941	.092	-.024	1.000	.941
Internal Score for Term 3 Section B	.746	1.000	.154	.056	.941	1.000

Inter-Item Covariance Matrix

	Internal Score for Term 1 section A	Internal Score for Term 1 Section B	Internal Score for Term 2 Section A	Internal Score for Term 2 Section B	Internal Score for Term 3 Section A	Internal Score for Term 3 Section B
Internal Score for Term 1 section A	49.691	50.530	.480	-3.447	36.960	50.530
Internal Score for Term 1 Section B	50.530	92.446	9.690	4.435	67.075	92.446
Internal Score for Term 2 Section A	.480	9.690	42.840	48.546	4.444	9.690
Internal Score for Term 2 Section B	-3.447	4.435	48.546	68.688	-1.504	4.435
Internal Score for Term 3 Section A	36.960	67.075	4.444	-1.504	54.920	67.075
Internal Score for Term 3 Section B	50.530	92.446	9.690	4.435	67.075	92.446

To further examine the internal structure of the teacher-made mathematics assessments, an inter-item covariance matrix was constructed (Table 8). This matrix highlights the degree to which scores on one section of the test covary with scores on another section across the three academic terms.

The diagonal values represent the variances of each test section, with the highest observed in Term 1 Section B (92.446) and Term 3 Section B (92.446), suggesting these sections contributed most to total score variability. The strongest covariances appeared between: Term 1 Section A and Term 1 Section B (50.530), Term 1 Section B and Term 3 Section B (92.446), and Term 3 Section A and Term 3 Section B (67.075).

These values indicate that performance on these sections was highly consistent, likely due to their shared cognitive demands or content domains. The lowest (and negative) covariance occurred between Term 1 Section A and Term 2 Section B (-3.447), and between Term 2 Section A and Term 3 Section A (-1.504). These negative values may signal item misalignment, inconsistent difficulty levels, or instructional gaps between those sections.

Covariance values provide a foundational layer for understanding inter-item consistency. High, positive covariances suggest that students who performed well on one section generally performed similarly on related sections—strengthening evidence of internal coherence in the tests (Field, 2013). However, negative or near-zero covariances, such as those observed between Term 1A and Term 2B, and Term 2A and Term 3A, may warrant further review. These inconsistencies could stem from abrupt shifts in topic areas, uneven instructional delivery, or insufficient scaffolding between terms. Such gaps may disrupt students' ability to transfer learning from one test context to another (Nitko & Brookhart, 2014). This is summarized in the matrix Scatterplot below.

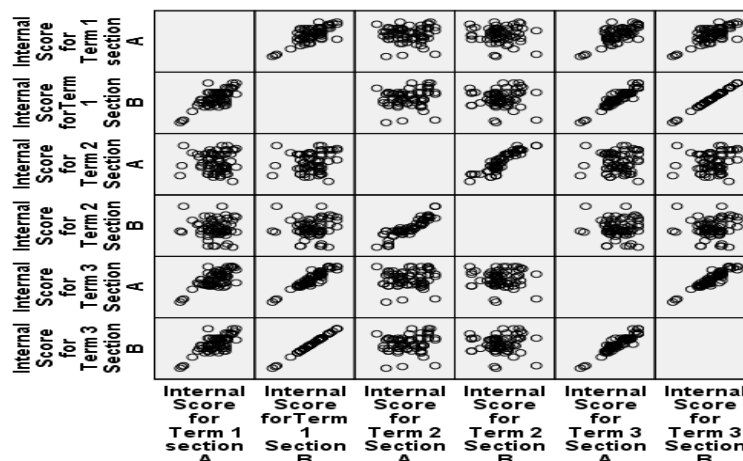


Figure 1: Matrix scatter-plot of students observed scores

Test Additivity and Variance Structure

Analysis of variance (ANOVA) was conducted to determine if statistically significant differences existed between mean scores across test sections and terms. The results of the one-way ANOVA yielded non-significant F-values ($F = 0.65$, $p = 0.524$ for Section A; $F = 1.99$, $p = 0.140$ for Section B), indicating no significant difference in scores across terms. This aligns with the reliability estimates, suggesting that students performed consistently across all administered assessments.

The results from the ANOVA with Friedman's Test and Tukey's Test for Non-additivity (Table 4.11) offer further insights into the structural integrity of the assessment scores. The Friedman's chi-square value of 84.346 ($p < 0.001$) suggests significant differences in item means across the six components (three terms $\times$ two sections). This indicates that not all test sections contributed equally to overall student performance, possibly due to variations in content difficulty or item structure.

Table 3 presents the ANOVA structure of the test analyses

Table 3: Variance and Additivity Analyses

		Sum of Squares	df	Mean Square	Cochran's Q	Sig
Between People		7782.277	64	121.598		
Within People	Between Items	6384.008	1	6384.008	53.901	.000
	Residual	1314.492	64	20.539		
	Total	7698.500	65	118.438		
Total		15480.777	129	120.006		
Grand Mean = 33.3308						
Intraclass Correlation Coefficient						
		95% Confidence Interval		F Test with True Value 0		
		Intraclass Correlation ^b		Lower Bound	Upper Bound	
				Value	df1	df2
Single Measures		.711 ^a		.566	.813	5.920 64 64
Average Measures		.831 ^c		.723	.897	5.920 64 64
ANOVA with Friedman's Test and Tukey's Test for Non-additivity						
		Sum of Squares	df	Mean Square	Friedman's Chi-Square	Sig
Between People		13693.831	64	213.966		
	Between Items	15778.300	5	3155.660	84.346	.000
	Nonadditivity	687.216 ^a	1	687.216	19.426	.000
Within People	Residual	11284.984	319	35.376		
	Total	11972.200	320	37.413		
Total		27750.500	325	85.386		
Total		41444.331	389	106.541		
Grand Mean = 33.1308						
a. Tukey's estimate of power to which observations must be raised to achieve additivity = -.167.						

Additionally, the Tukey test for non-additivity in Table 3 reported a significant chi-square value of 19.426 ($p < 0.001$), highlighting potential interaction effects or lack of additivity in the data. Tukey's estimate for the transformation power (-0.167) suggests a moderate departure from additivity, implying that the relationship between items and student ability may not be strictly linear. This has implications for score interpretation, especially when using total scores as indicators of student proficiency.

Nevertheless, the relatively low non-additivity component (687.216 out of 41,444.331 total variance) indicates that while the deviation from additivity is statistically significant, it is practically limited in scale. Educators and test developers should be aware of such effects but need not overcorrect unless they intend to apply more advanced measurement models. These results underscore

the importance of test dimensionality diagnostics as emphasized by McDonald (1999), who advocates for careful evaluation of linearity and unidimensionality in classical test theory applications. Ignoring such diagnostics can lead to misinterpretation of composite scores and misalignment between test intent and observed outcomes.

Overall, these findings support the need for balanced test design and periodic variance structure analysis in school-based assessments (SBA) as suggested by the National Council for Curriculum and Assessment (NaCCA). Identifying non-additivity ensures that test scores accurately represent student performance without distortion due to interaction effects between items and student abilities.

Interpretation and Implications

The high reliability coefficients observed in this study demonstrate that the mathematics assessments administered in Techiman SHS exhibited a commendable level of internal consistency. This contradicts common perceptions that teacher-made tests lack psychometric rigor. While variability existed across terms, it was not substantial enough to question the dependability of scores. However, the presence of high inter-item correlations, particularly in Term 2, may suggest some degree of item redundancy. This offers an opportunity for professional development in item writing, ensuring that questions are diverse yet aligned with a common construct. Additionally, the consistent performance across sections implies that teachers maintained a balanced test design, which is commendable.

These findings align with Kane's (2016) framework for validation, which emphasizes the need for both consistency and construct alignment in score interpretation. The consistent results also reinforce the potential of school-based assessments to generate valid inferences when guided by data-driven quality control measures.

V. Conclusion

This study set out to investigate the internal consistency and overall reliability of teacher-made mathematics assessments administered to senior high school students in Techiman, Ghana. Through the application of various statistical measures—including Cronbach's alpha, split-half reliability, inter-item correlations, intraclass correlation coefficients, and variance structure analysis—the study revealed that the assessments were generally reliable, consistent, and stable over time.

The results suggest that despite the non-standardised nature of school-based assessments, teachers in the selected schools demonstrated commendable coherence in test construction. High reliability coefficients across three academic terms affirm that teacher-made assessments can function effectively as valid tools for measuring student performance, particularly in mathematics where conceptual depth and procedural accuracy are critical.

Nonetheless, the study also uncovered specific areas requiring attention. High inter-item correlations in some terms indicate potential item redundancy, while the presence of non-additivity in variance structure calls for improved balance in item design. These findings align with existing literature (e.g., Sijtsma, 2009; Revelle & Zinbarg, 2009), which stresses the importance of both internal consistency and structural validity in educational assessment.

Ultimately, the findings reinforce the argument that reliability should not be an afterthought in school-based assessment but a central concern. With appropriate training, support, and reflective practice, teacher-made assessments can rival formal standardized tests in both accuracy and pedagogical relevance.

VI. Recommendations

Based on the findings and their implications, the following recommendations are proposed for Education ministries and teachers to consider:

There should be a Capacity Building in Test Construction. Education ministries should organize regular in-service workshops for mathematics teachers focused on psychometric principles, especially test blueprinting, item analysis, and reliability estimation.

Colleges of Education and teacher training programs should embed advanced assessment literacy, ensuring pre-service teachers are well-equipped in constructing and evaluating valid and reliable assessments.

Schools should incorporate periodic reliability analysis as part of assessment quality assurance, particularly for high-stakes tests that contribute significantly to academic progression.

Teachers should review and revise items with extremely high inter-item correlations to avoid redundancy. Piloting test items before formal administration can help detect such issues early.

Establish peer review and moderation frameworks where teachers collaboratively design and review assessments to ensure balance, fairness, and psychometric soundness.

By acting on these recommendations, educational stakeholders in Ghana and similar contexts can strengthen the integrity and instructional value of school-based assessments, ultimately supporting more accurate and equitable learning outcomes.

References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. American Educational Research Association.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014).
3. Anamuah-Mensah, J. (2020). Teacher assessment practices and the development of learners' competencies in Ghanaian schools. *Ghana Journal of Education*, 17(2), 45–61.
4. Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429–458. <https://doi.org/10.1111/j.1467-9620.2004.00348.x>
5. Cai, J., & Lester, F. K. (2010). Why is teaching with problem-solving important to student learning? National Council of Teachers of Mathematics Research Brief. <https://www.nctm.org>
6. Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory (2nd ed.). Cengage Learning.
7. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
8. Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). Macmillan.
9. Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE.
10. George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference* (4th ed.). Allyn & Bacon.
11. George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference* (16th ed.). Routledge.
12. Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership*, 60(5), 6–11.
13. Kane, M. T. (2016). Validating score interpretations and uses. *Language Testing*, 33(3), 293–303. <https://doi.org/10.1177/0265532215607443>
14. McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.
15. Mensah, F. A., & Frimpong, J. (2021). Teachers' knowledge and practices of assessment in Ghanaian basic schools. *African Journal of Teacher Education*, 10(1), 23–39. <https://doi.org/10.21083/ajote.v10i1.6204>
16. Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.
17. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
18. Pallant, J. (2020). *SPSS survival manual* (7th ed.). Open University Press.
19. Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, 14(1), 57–74.
20. Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145–154. <https://doi.org/10.1007/s11336-008-9102-z>
21. Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
22. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
23. Stiggins, R. J. (2005). From formative assessment to assessment FOR learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324–328. <https://doi.org/10.1177/003172170508700414>
24. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53–55. <https://doi.org/10.5116/ijme.4dfb.8dfd>
25. Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69(4), 613–625. <https://doi.org/10.1007/BF02289858>