International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Pupil Performance in Multiple Choice Test Items and their Corresponding Structured Items in Mathematics

Francis Chirume[1], Prof. M. Kurebwa[2]

[1]Research Manager, Research and Development Unit, Zimbabwe School Examinations Council
[2]Lecturer/Chairperson: Educational Studies, Zimbabwe Open University

Abstract
This study delves into the intriguing relationship between student performance on Multiple Choice (MC) and structured test items in Grade 7 Mathematics in Zimbabwe's Midlands province. The research aimed to determine if students demonstrate similar competency levels when assessed using these different formats. Data from eight schools classified into boarding, urban day, rural day, and resettlement groups, were analysed using descriptive statistical techniques. Contrary to the hypotheses suggesting that prior exposure to MC items might enhance performance on structured tests, the results consistently showed that students performed better on the MC component. This trend suggests that factors like strategic guessing and test-wiseness may artificially inflate MC scores. Additionally, disparities in performance were observed across school types, with boarding schools excelling and resettlement schools falling behind in both assessment formats, highlighting the impact of socioeconomic factors on educational outcomes. These findings raise concerns about the reliance on MC assessments as true indicators of mathematical competence, given their susceptibility to guessing strategies. The study recommends further research with a larger and more diverse sample to provide a robust evidence for educational policy and assessment practices in Zimbabwe and beyond.

Keywords: Pupil performance, Multiple Choice test items, structured items, Test formats, Educational Assessment, Mathematics.

1.0 Introduction

1.1 Background to the Study

The disparity in student performance across different assessment formats is a persistent concern in educational research. While various test formats exist, including multiple choice, structured, essay, practical, and oral presentations, each is designed to measure specific skills and abilities. Multiple choice questions offer the efficiency of broadly sampling a subject domain; concerns remain about their limitations in assessing higher order cognitive skills such as problem solving and critical thinking (Martinez, 2017). Conversely, structured or open ended formats, while potentially more demanding in terms of response processing, are often lauded for their ability to elicit deeper understanding and analytical skills (Boudah, 2020).

This tension between assessment efficiency and the depth of cognitive engagement is particularly relevant to the case of the Zimbabwean Grade 7 Mathematics examinations. The current system, with its dual

format of a multiple choice paper 1 (60% weighting) and a structured, free response paper 2 (40% weighting), presents a compelling paradox. Despite covering the same syllabus content and ostensibly assessing the same mathematical competencies, the consistent trend of higher student performance on the multiple choice paper, as highlighted in the Zimbabwe School Examinations Council Statistical Report (2022), raises critical questions about the validity of these assessments in accurately gauging students' true mathematical understanding.

This discrepancy becomes even more intriguing when considering the potential for guessing in multiple choice assessments; a factor that could artificially inflate scores (Haladyna, 1997). While paper 2, with its emphasis on detailed problem solving and demonstration of mathematical processes, aligns more closely with the assessment of higher order thinking skills, the persistently lower student performance on this paper suggests a disconnect between the intended learning outcomes and the assessment format.

This study, focussing on the relationship between learner performance on corresponding multiple choice and structured items on Grade 7 Mathematics, has the potential to make a significant contribution to this ongoing debate. By directly comparing student responses on items assessing identical mathematical concepts in both formats, this research can shed light on the nuanced ways in which assessment format might influence student performance and provide valuable insights into the cognitive processes at play. This study aligns with a growing body of research emphasising the need for a balanced approach to assessment, one that recognises the strengths and limitations of different formats in accurately measuring the complexities of student learning (Bao & Kilic, 2022).

1.2 Statement of the Problem

This study aimed to investigate and compare pupil performance in multiple choice test items and their corresponding structured items in Mathematics to determine if there are significant differences in performance between these two test formats and if the skills measured align closely. Additionally, the research sought to explore how pupil performance varies across these test formats concerning gender and school type, factors that have not been extensively examined in previous studies. By addressing these questions, the research aimed to contribute to a better understanding of the effectiveness and comparability of multiple choice and structured test items in assessing mathematical proficiency among students.

1.3 Purpose of the Study

The purpose of the study was to investigate the performance of pupils in multiple choice test items and their corresponding structured items in Mathematics. It sought to establish the relationships between the two test formats, thus allowing the study to compare learner competencies in the test formats and how the competencies are related. The research sought to answer the primary question thus: Do pupils perform the same on the multiple choice test items and their corresponding structured items in Mathematics?

1.4 Research Objectives

The study sought to:

- Determine the relationship between pupils' correct answers on the multiple choice paper and correct answers on the structured paper.
- Determine the relationship between pupil performance in multiple choice test items and the corresponding structured test items.
- Establish any gender differences in the performance of pupils in the two test formats.
- Establish any differences by school type in the performance of pupils in the two test formats.

1.5 Significance of the Study

The significance of the study is multifaceted and holds the potential to impact on various stakeholders in

the educational landscape. By examining the relationship between student performance on corresponding multiple choice and structured Mathematics items, the study could yield valuable insights into the complexities of assessments and their impact on student learning outcomes.

Firstly, the findings could be instrumental in informing pedagogical practices. By pinpointing specific areas where students demonstrate inconsistencies in their understanding across different assessment formats, teachers can gain a more nuanced understanding of their students' strengths and weaknesses. This, in turn, can enable them to tailor their instructional strategies to address specific areas where students struggle to translate their knowledge into different assessment contexts. For instance, if students excel in multiple choice items related to a specific mathematical concept but falter when asked to apply the same concept in a structured problem solving scenario, it signals a need for more emphasis on application and problem solving during instruction.

Secondly, the study holds implications for assessment development and refinement. By providing empirical evidence on the comparability of multiple choice and structured items intended to assess the same mathematical constructs, the research could inform the work of test developers, including those at the Zimbabwe School Examinations Council. This could lead to the development of more robust and equitable assessments that accurately measure students' true mathematical understanding while minimising the potential biases associated with specific assessment formats.

Ultimately, the insights gleaned from this study have the potential to contribute to a more equitable and effective Mathematics education system in Zimbabwe. By shedding light on the interplay between assessment format and student performance, the study can empower educators, assessment developers, and policymakers to make more informed decisions that support the mathematical success of all learners.

1.6 Hypotheses

The following hypotheses guided the study:

H01: There is no significant difference in the performance of pupils in multiple choice test items and their corresponding structured items.

H02: There is no gender difference in the performance of pupils in multiple choice test items and their corresponding structured items.

H03: There is no significant difference in the performance of pupils in the two test formats by school type.

2.0 Literature Review

2.1 Theoretical Framework

The study investigated the cognitive processes underlying student performance in Mathematics assessments, specifically focussing on how different question types engage working memory and influence learning outcomes. Cognitive Load Theory (CLT), developed by John Sweller, provides the theoretical framework for this investigation. CLT posits that working memory, the cognitive system responsible for processing information during learning, has limited capacity and duration. This limitation necessitates careful consideration of the cognitive demands imposed by learning tasks, particularly within assessment design.

The CLT distinguishes between three types of cognitive load: intrinsic, extraneous, and germane. Intrinsic load refers to the inherent complexity of the information being processed, such as the difficulty of a mathematical concept. Extraneous load, on the other hand, arises from the way information is presented, with poorly designed materials or instructions increasing unnecessary cognitive demands. Germane load,

however, is directly related to the learner's construction of schemas and understanding, reflecting the mental effort invested in making meaningful connections.

Within Mathematics assessments, different question types impose varying levels of cognitive load, impacting both student performance and learning outcomes. Structured questions, while requiring deeper cognitive processing and potentially promoting long-term learning, also come with a higher intrinsic load due to their complexity. Conversely, multiple-choice questions, with their lower load, may rely more on recognition than deep understanding, potentially limiting their ability to assess conceptual mastery.

This study, therefore, aimed to unpack the interplay of these cognitive load types within the context of Mathematics assessments. By analysing student performance and strategies across different question types, the research sought to gain insights into how cognitive load influences learning and how educators can design assessments that optimises both challenge and understanding.

2.2 Multiple Choice Tests

These are tests where a stem is given and alternative answers are provided. Among the set of options is the correct answer (key) and the remaining incorrect answers are the distractors. The distractors should be functional so that they have the power to distract the attention of the candidate from the key. The distractors therefore should be plausible.

Multiple choice tests have several advantages. Because of their structure and format, they have a good coverage of the domain under assessment. Hence, multiple choice tests are good and relevant for revision purposes. The learners' incorrect responses can be used to diagnose their mathematical problems and difficulties.

These tests can be easily marked, because of their structure where candidates simply write letters which represent correct answers. As a result, such tests have the opportunity for timely feedback. Roediger and Marsh (2005) add that apart from being easy to mark, such tests improve student performance on other tests to come as a result of testing effect. Generally, candidates find it less difficult to prepare for multiple choice tests since they have the tendency to reduce test anxiety among learners (Snow, 1993).

Item analysis is easy to carry out for multiple choice tests. This will be easy to improve the tests. Apart from analysing the test items, students' scores are also easy to analyse because of their nature.

There are some disadvantages associated with multiple choice tests. Chief among them is the issue that the tests may not accurately portray learner ability as a result of guessing. In addition learners may not be able to synthesise concepts, according to Popham (2010), and higher order thinking skills might not be portrayed. A lot of time is taken when constructing multiple choice tests. Therefore, it is difficult to set multiple choice tests for items requiring organisation and presentation of ideas as observed by Popham (2010). Therefore, multiple choice tests render themselves to construct under-representation, which affects the validity of assessment, Messick (1995).

The incorrect answers candidates are exposed to affect them when they take other tests later. Roediger (2005) points out that candidates tend to remember the incorrect answers and will take them as correct in later tests. Therefore students will learn the wrong things.

2.3 Structured Tests

Structured tests are constructed so that candidates provide their own answers. In Mathematics the candidates will be showing all the processes, procedures, formula and algorithms used to arrive at the answer. As a result, the tests provide information with regards to learner misconceptions as they solve problems. Structured tests reduce measurement error, since they are not prone to guessing. In structured

tests, learners cannot work backwards to arrive at the answer, unlike in multiple choice tests as observed by Bridgeman (1992).

Free response tests allow learners to show their strengths and weaknesses as they provide answers to problems. They diagnose students' misconceptions, and they test better learner multiple abilities, and capabilities. They assess all the cognitive processes and other related skills.

In Mathematics, mathematical processes can earn high marks even if the answer is not correct. Responses are not affected by guessing. Learner mathematical competencies are best revealed in a structured or free response paper.

Structured tests take time to mark and they are affected by measurement error. Scoring reliability is less in structured tests than multiple choice tests. Therefore, this affects the objectivity in assessment. In addition to these limitations, structured tests are affected by test anxiety. Research by Crocker and Schmitt (1987) found that the negative effects of test anxiety on scores were moderate on multiple choice questions but severe on structured items. The issue of providing solutions to structured tests through explanations and showing mathematical processes induces anxiety among candidates, resulting in candidates failing to proficiently express themselves.

2.4 Testwiseness and Guessing

When taking a test, candidates can gain marks as a result of testwiseness and guessing. In testwiseness, candidates choose a correct answer without knowing that it is correct. Test wise learners study the item, search for mistakes in the construction and make guesses. Testwise candidates are not usually knowledgeable of the subject matter.

Related to testwiseness is the issue of guessing. Candidates can guess in all multiple choice items in a test and pass the test. There are two types of guessing, random guessing and educated guessing. Random guessing has the absence of knowledge, while in an educated guess, there is evidence of some knowledge of content (Cronbach, 1998).

Results of multiple choice tests can be influenced by testwiseness as observed by Simkin and Kuechler (2005) and random guessing may also affect the results, where it is thought the test was scored well when in actual fact no learner abilities were realised. In this study the researcher minimised testwiseness by making sure that the tests were error free and also ensured uniform administration of the tests across and within schools.

2.5 Reliability, Validity and Bias

Reliability refers to the consistency of a measuring tool in yielding the same or similar results after repeated administration. A reliable test is stable over time and within itself. While reliability deals with consistency and stability, validity deals with accuracy in measurement. A test is valid if it accurately measures what it is supposed to measure. Validity also deals with the spread of test items within the domain that is assessed. In this study, the two test formats were pilot tested to ensure validity and reliability of test items.

Assessment specialists and teachers should guard against test bias. A test is biased if it favours one group of examinees than other groups. Bias can be based on religion, racial, cultural, gender, socio-economic status and geographical location of candidates. Items biased on these variables would be said to have item differential functioning on the basis of these variables, thus candidates positively affected by these variables will tend to score highly than the rest of the examinees. The presence of bias renders the test scores invalid as observed by Lam (1995). For example, in Mathematics, the ability to comprehend a

question is a bias in the measurement of mathematical skills, as learners with limited English skills are affected (Stenmark, 1989).

A good assessment is fair, valid and reliable. These principles can be improved by following proper test construction stages, and pilot testing the instruments. In the case of a school, the views of other teachers should be sought in moderating the tests.

2.6 Related Studies

Given the rich landscape of previous research studies on test formats and learner performance, it is evident that there is need for further exploration and analysis in this area. The studies conducted by Bridgeman (1992), Lukhele, Thissen, and Wainer (1994), and Walstad and Becker (1994) have laid a foundation for understanding the relationship between structured tests and multiple choice tests. However, as these studies were conducted around two decades ago, it is imperative to delve into current literature to ascertain whether the findings still hold true in today's educational landscape.

In the study by Fleming (1998), the differential impact of test formats on learners of varying abilities was explored, shedding light on the perception of test difficulty by teachers and the performance of different groups of learners. This study underscores the importance of considering learner abilities and perceptions when designing and administering tests.

Hamilton's (1994) study on pupil preferences for test formats revealed valuable insights into student perceptions of test difficult and personal preferences. The findings of this study highlight the importance of considering student feedback and preferences when designing assessments to enhance engagement and motivation

Mazzeo, Schmitt, and Bleinstein's (1992) research on gender differences in test performance across different formats brings to the forefront the issue of gender disparities in assessment outcomes. The findings of this study emphasise the need for further investigation into gender-specific learning strategies and preferences to ensure equitable assessment practices.

Building on the existing body of literature, Livingstone (2009) introduced a new dimension by exploring the relationship between skill improvement in structured items and performance in multiple choice tests. This study underscores the complexity of test formats and their implications for measuring learner competencies effectively.

Kimball's (1989) study on learning strategies for males and females in Mathematics provides valuable insights into gender-specific approaches to problem-solving and algorithmic tasks. Beller and Gafni's (2000) research further expands on this by examining candidate preferences, test performance, and gender differences across test formats. These studies collectively highlight the importance of considering gender-specific learning preferences and strategies when designing assessments to cater for diverse learner needs.

In conclusion, the existing literature on test formats, learner preferences, and gender differences in test performance provides a robust foundation for further exploration and analysis. By synthesising current research findings with these seminal studies, a comprehensive understanding of the complexities surrounding test formats and assessment practices can be achieved, paving the way for more tailored and effective educational assessments. This holistic approach is paramount in developing assessments that are not only fair, valid, and reliable but also aligned with the evolving goals of education.

3.0 Methodology

3.1 Research Design

The study adopted the descriptive survey method. The choice to employ a descriptive survey design for

the study rests on the assumption that this approach can illuminate the nuanced relationship between test format and student performance within a naturalistic classroom setting. By collecting and analysing data on student performance across both multiple choice and structured test items, alongside relevant teacher and pupil characteristics, the study aimed to identify potential trends and correlations. While this design cannot establish causal relationships, it provides a valuable lens for examining how student performance might differ across test formats and whether specific learner or instructional factors correlate with these differences. Acknowledging the inherent limitations of descriptive research, the study sought to provide a rich, detailed analysis of student performance within a real-world context, informing future research and pedagogical practices related to Mathematics assessment. This design was appropriate in that it enabled the researchers to collect data, analyse and describe the performance of pupils in the two test formats. The survey also enabled the researcher to describe teacher and pupil characteristics together with the conditions under which the instruments were administered.

3.2 Population, Sample and Sampling Procedures

The population under study consisted of Grade 7 primary school pupils in the Midlands province. The province was randomly selected from the 10 provinces for the study. Grade 7 candidates were the ideal candidates for the study because they had completed all the topics in the syllabus and as set in the respective question papers. The accessible population were all the schools in the selected province.

A total of eight schools took part in the study. The schools were categorised into four groups that is, boarding, urban day, rural day and resettlement. Two schools were randomly selected from each school type category resulting in a sample size of eight schools.

3.3 Data Collection Instruments

A questionnaire for teachers and two tests were used to collect data. The tests consisted of a multiple choice paper and a structured paper. Both the multiple choice test and the structured test had 25 items which were scored out of 50. The test items were the same word for word and occupied the same position throughout the two tests. The only difference was that options were provided in the multiple choice paper and pupils were required to work out the problem and select the answer while in the structured test, pupils were to work out the problem and get the answer. The items covered all the four key topics of the Junior School Mathematics syllabus; that is; number, operations, measures and relationships. The items were also set to address all the skill levels of the syllabus, namely; knowledge, routine manipulation, understanding and application, and problem solving.

A questionnaire for teachers who taught the pupils was constructed to collect qualitative data. The questionnaire was designed to solicit demographic data of the respondents, their experience in teaching Grade 7 pupils, and whether they marked or have set public examinations. The questionnaire also gathered information on the teachers' knowledge of the Junior Primary School Mathematics syllabus, with respect to content, assessment objectives and scheme of assessment. This information was required since it had a bearing on the performance of the respective candidates and how the pupils would tackle the test questions.

3.4 Validity and Reliability of Instruments

The tests were moderated by a team of curriculum and assessment experts. The items were aligned to syllabus content and syllabus specification grid. This was done to ensure the content validity of the test. The multiple choice paper was scrutinised and distractors analysed for their power to distract the attention of the candidate from the key. All the papers were pilot tested to Grade 7 pupils at a school not in the sample. Pilot study results led to further modification of the instruments. Pilot study results showed that the test items were valid and highly reliable as reflected in the table below.

Table 1: Reliability Statistics for Multiple Choice and Structured Test Items

Cronbach's Alpha	Part 1	Value	1.000
		N of Items	1[a]
	Part 2	Value	1.000
		N of Items	1[b]
	Total N of Items		2
Correlation Between Forms			.929
Spearman-Brown Coefficient	Equal Length		.963
	Unequal Length		.963
Guttman Split-Half Coefficient			.959
a. The items are: Multiple Choice Scores			
b. The items are: Structured Test Scores			

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.959	.963	2

The tables above reflect that the tests were highly reliable as they yielded high reliability coefficients, Cronbach's Alpha of 0.959. This meant that the tests would yield the same or similar results if administered to the same or similar pupils at time intervals, therefore the test items were very stable. In addition to the stability of the tests, the tests were standardised. The tests yielded a coefficient of 0.963 according to Cronbach's Alpha Based on Standardised Items.

The two tests were correlated and the inter-item correlation coefficient was 0.929. The inter-item correlation matrix is shown below.

Inter-Item Correlation Matrix		
	Multiple Choice Scores	Structured Test Scores
Multiple Choice Scores	1.000	.929
Structured Test Scores	.929	1.000

The inter-item correlation coefficient of 0.929 was above the acceptable standard (usually 0.7 or above). The inter-item correlation is a measure of the correlation of each item in a test with each and every other item in a test. This meant that the items in the two test formats were in agreement.

3.5 Data Collection Procedures

Sample schools were informed in advance that their Grade 7 pupils would sit for the tests in Mathematics. They were informed of the nature of the tests and how the tests would be administered. This was done to enable teachers and pupils prepare for the tests. The schools were also notified about the actual dates their pupils would write the tests.

Upon arrival at a school, the researcher explained the purpose of the visit to the head. One Grade 7 class was randomly selected to take part in the study in the case of a school with two or more Grade 7 classes. Once a class was selected, all the pupils in that class would take the tests. Pupils were given candidate numbers (identification numbers) which they wrote on answer sheets on both tests. This enabled the researcher to marry paper 1 scripts to paper 2 scripts and easy data analysis.

The multiple choice test was administered first. Although in the multiple choice test, the alternatives were provided and pupils were required to choose the correct answer, the pupils were provided with plain paper where they would work out the problem first and then select the key from the options and indicate it on the answer sheet provided. This instruction was also sent as part of the advance information, in addition to further clarification by the researcher.

Pupils were given 30 minutes break before they took the structured test. In the structured test, pupils worked out the problem on the spaces provided on the question paper. Although the tests were to be written in 1 hour each, slow pupils were allowed to complete the tests so as to minimise data lose through non-response of items. The researcher led the data collection process at each sampled school through issuing out instructions and monitoring pupils writing. This was done to enable consistency in test administration across sampled schools. The grade 7 teacher was present to assist by distributing stationery and other materials to pupils.

Class teachers completed questionnaires while their pupils were writing tests.

Administration of the tests took eight days that is one day per school. After administration of the tests, candidate scripts were scored. To allow for consistency in scoring and minimising the error score, all the scripts were marked by the researcher.

3.6 Data Presentation and Analysis Procedures

Data was displayed using tables, graphs, and textual presentations. Bar graphs were utilised to illustrate a comparative analysis of pupil performance in multiple choice and structured tests, aiming to evaluate the consistency of performance between the two test formats. The analysis of the data was conducted using Microsoft Excel and the Statistical Package for the Social Sciences (SPSS). SPSS was employed to conduct T-tests and Analysis of Variance to compare pupil performance in multiple choice and structured test items.

4.0 Data Presentation, Analysis and Interpretation

4.1 Sample Characteristics

A total of 321 pupils took part in the study. Of these, 183 were females constituting 57% while 138 were males forming 43% of the respondents.

The distribution of pupils by school type was as reflected in the table below.

Table 2: Distribution of respondents by school type

School Type	Number of Pupils	Percentage
Boarding	101	31,46

Resettlement	63	19,63
Rural Day	62	19,31
Urban Day	95	29,60
Total	**321**	**100%**

The table shows that there were more pupils from boarding and urban day schools than resettlement and rural day schools. The reason could be attributed to large class sizes that characterise boarding and urban day schools. The majority of resettlement and rural day schools did not have large class sizes. The average class size in these schools is 35 according to the Ministry of Primary and Secondary Education Annual Report of 2020.

4.2 Relationship between the Percent of Pupils' Correct Answers on Multiple Choice and Structured Items

An analysis of pupils' correct responses for each item was done by gender. The results were presented in the table below.

Table 3: Pupils' correct responses on multiple choice and structured items by gender

Item	**Number and % of Correct Responses in Multiple Choice Items**		**Number and % of Correct Responses in Structured Items**	
	Girls	**Boys**	**Girls**	**Boys**
1	**138 (75, 4%)**	98 (71, 1%)	105 (57, 4%)	**82 (59, 4%)**
2	102 (55, 71%)	**86 (62, 3%)**	89 (48, 6%)	**69 (50%)**
3	**55 (30%)**	40 (28, 9%)	30 (16, 4%)	**28 (20, 3%)**
4	**76 (41, 5%)**	53 (38, 4%)	**41 (22, 4%)**	26 (18, 8%)
5	106 (57, 9%)	**87 (63%)**	70 (38, 3%)	**57 (41, 3%)**
6	31 (16, 9%)	**38 (27, 5%)**	31 (16, 9%)	**27 (19, 6%)**
7	32 (17, 5%)	**39 (28, 3%)**	21 (11, 5%)	**21 (15, 2%)**
8	47 (25, 7%)	**39 (28, 3%)**	30 (16, 4%)	**23 (16, 7%)**
9	58 (31, 7%)	**50 (36, 2%)**	31 (16, 9%)	**31 (22, 5%)**
10	112 (61, 2%)	**72 (52, 2%)**	34 (18, 6%)	**33 (23, 9%)**
11	77 (42, 1%)	**80 (58, 0%)**	62 (33, 9%)	**65 (47, 1%)**
12	**133 (72, 7%)**	94 (68, 1%)	111 (60, 7%)	**86 (62, 3%)**
13	**83 (45, 4%)**	61 (44, 2%)	39 (21, 3%)	**39 (28, 3%)**
14	**91 (49, 7%)**	61 (44, 2%)	**58 (31, 7%)**	40 (29%)
15	**132 (72, 1%)**	90 (65, 2%)	**83 (45, 4%)**	55 (39, 9%)
16	53 (29, 0%)	**41 (29, 7%)**	29 (15, 8%)	**27 (19, 6%)**
17	53 (29, 0%)	**51 (37, 0%)**	40 (21, 9%)	**39 (28, 3%)**
18	**82 (44, 8%)**	60 (43, 4%)	34 (18, 6%)	**38 (27, 5%)**
19	**91 (49, 7%)**	64 (46, 4%)	**70 (38, 2%)**	45 (32, 6%)
20	**81 (44, 3%)**	60 (43, 4%)	**63 (34, 4%)**	42 (30, 4%)
21	**45 (24, 6%)**	29 (21%)	**18 (9, 8%)**	12 (8, 7%)
22	47 (25, 6%)	**40 (28, 9%)**	43 (23, 5%)	**33 (23, 9%)**
23	**66 (36, 1%)**	46 (33, 3%)	18 (9, 8%)	**17 (12, 3%)**

24	**93 (50, 8%)**	66 (47, 8%)	**42 (22, 9%)**	30 (21, 7%)
25	**77 (42%)**	47 (34%)	**46 (25, 1%)**	30 (21, 7%)

NB*:* Bolded figures represent the highest percentage of correct responses in the two test formats by gender.

The table data indicates that girls achieved the highest percentage of correct responses in 14 out of 25 multiple choice items. This ratio corresponds to 56%, signifying that girls provided more correct answers than boys in 56% of the multiple choice questions. Conversely, in 68% of the structured items, boys outperformed girls in terms of correct responses. This suggests that girls demonstrated a relatively higher percentage of correct answers than boys in multiple choice items, while the opposite trend was observed for structured or free response items. However, these differences in correct responses were found to be statistically insignificant.

4.3 Performance of Pupils in Multiple Choice and Structured Tests

Table 4: Pupils' mean scores in multiple choice and structured tests

Test Type	N	Mean	Std. Deviation	Std. Error of Mean
Multiple Choice	321	19,57	10,77	0,60
Structured Test	321	12,93	12,22	0,68

Figure 1: Bar graph showing mean marks of pupils in multiple choice and structured tests.

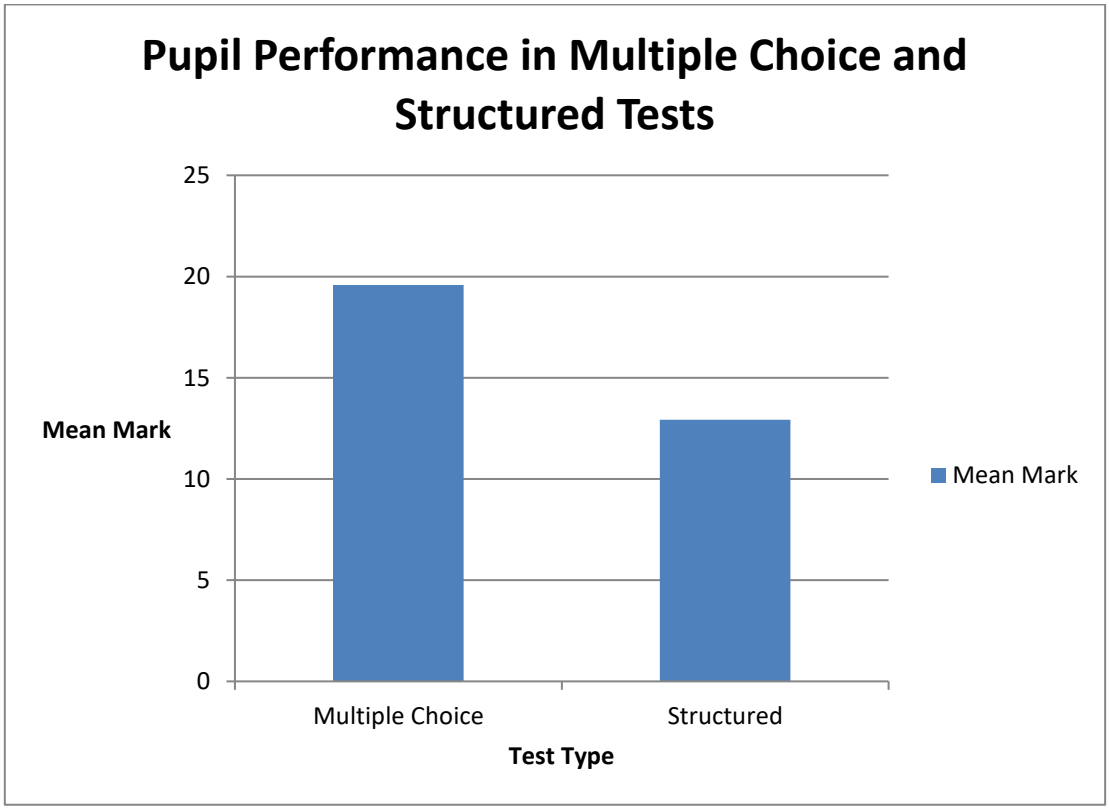


Table 4 shows the mean marks of pupils in multiple choice and structured tests. A total of 321 pupils took both tests. The mean mark for the multiple choice test (19, 57) was higher than that for the structured test which was 12, 93. The standard deviation for the multiple choice test was 10, 77 while that of the structured test was 12, 22. This showed that the marks for the structured test were more spread than scores for

multiple choice test. The results reflect that pupils did better in the multiple choice test despite the fact that the structured test was administered later. Under normal circumstances pupils were supposed to score high in the structured test paper as a result of practice effect. The multiple choice test was administered first and after a 30 minute break pupils wrote the structured paper and because of practice effect, test wise pupils were supposed to score highly in the free response paper since pupils had been exposed to the same items in the first paper (multiple choice). The fact that pupils did better in the multiple choice paper is reflective of the guess factor which influenced pupils as they took the multiple choice paper.

A one-sample t-test was performed to determine if there was a significant difference between the two means. The results are shown in the table below:

Table 5: T-test for the significant difference between multiple choice and structured test mean scores.

One-Sample Test						
	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Multiple Choice Scores	32.571	320	.000	19.57944	18.3968	20.7621
Structured Test Scores	18.951	320	.000	12.93146	11.5890	14.2740

The hypothesis test results show that the means were different at the 5% level of significance. This shows that the performance of pupils in the multiple choice paper was significantly higher than their performance in the structured test as reflected by the two-tailed test at the 5% level of significance. These results were confirmed by a one-way analysis of variance as reflected in the table.

Table 6: **One way analysis of variance for the significant difference in the means of multiple choice and structured tests.**

ANOVA					
Multiple Choice Scores					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	32929.179	45	731.760	48.038	.000
Within Groups	4189.045	275	15.233		
Total	37118.224	320			

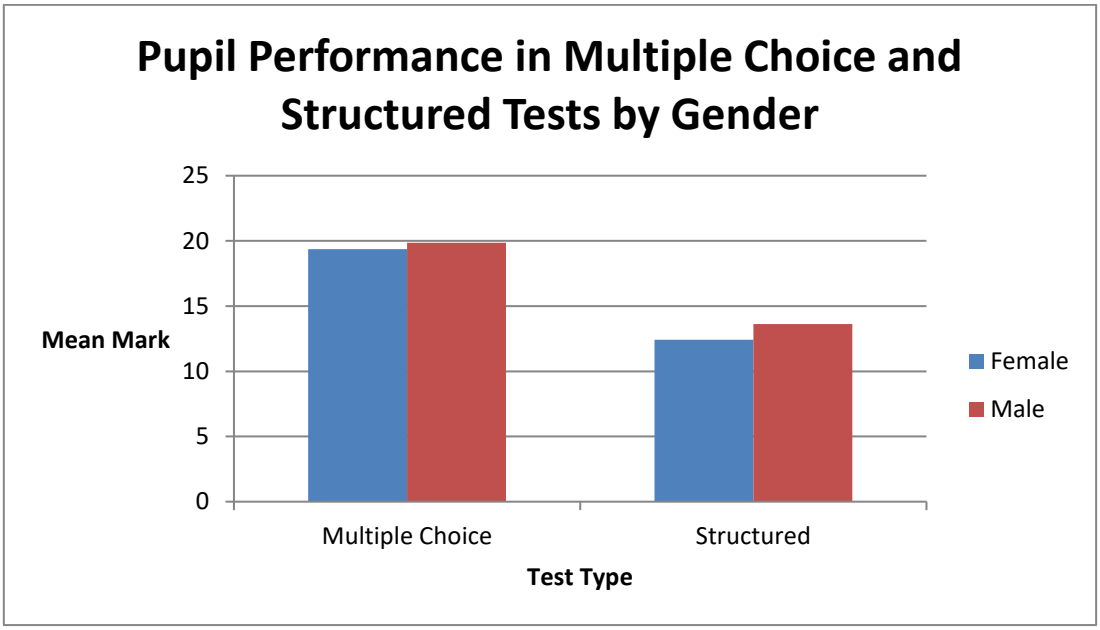
ANOVA					
Structured Test Scores					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	42467.565	45	943.724	48.392	.000
Within Groups	5362.928	275	19.502		
Total	47830.492	320			

4.4 Performance of Pupils in Multiple Choice and Structured Tests by Gender

Table 7: Mean Scores of Pupils by Gender

Multiple Choice Scores				
Gender	**Mean**	**N**	**Std. Deviation**	**Std. Error of Mean**
Female	19.37	183	10.30	.76
Male	19.86	138	11.39	.97
Total	19.58	321	10.77	.60

Structured Test Scores				
Gender	**Mean**	**N**	**Std. Deviation**	**Std. Error of Mean**
Female	12.42	183	11.57	.86
Male	13.60	138	13.05	1.11
Total	12.93	321	12.23	.68



The tables and the corresponding graph show that males performed better than females in both tests. The mean for females in the multiple choice paper was 19, 37 while males registered a mean performance of 19, 86. For the structured paper, the mean for females was 12, 42 and males had a mean of 13, 61. When the means for both papers are compared, it is clear that the means for the multiple choice paper were much higher than the corresponding structured paper. In both tests pupils did well in the multiple choice paper than the corresponding structured paper. A one way analysis of variance test, showed no evidence at the 5% level of significance in the difference between the means by gender.

Table 8: **Analysis of variance for mean differences by gender**.

ANOVA Table							
			Sum of Squares	df	Mean Square	F	Sig.
Multiple Choice Scores * Gender	Between Groups	(Combined)	18.391	1	18.391	.158	.691
	Within Groups		37099.834	319	116.300		
	Total		37118.224	320			

ANOVA Table							
			Sum of Squares	df	Mean Square	F	Sig.
Structured Test Scores * Gender	Between Groups	(Combined)	111.022	1	111.022	.742	.390
	Within Groups		47719.471	319	149.591		
	Total		47830.492	320			

The tables show that although males performed better than females in all test types, there was however no evidence to suggest that the difference in performance was significant. When means for multiple choice and structured test papers were compared, results showed a significant difference at the 5% level.

A comparative analysis of pupil performance by gender in the two tests was done. The performance of females in multiple choice and structured test items was done together with the performance of males in the two tests. This was done in order to determine the prevalence of the guess factor in pupil performance.

Table 9: Mean performance of pupils by gender

Gender	**Mean: Multiple Choice Test**	**Mean: Structured Test**	**Mean Difference**
Females	19,37	12,42	6,95
Males	19,86	13,61	6,25
Total	**19,58**	**12,93**	**6,65**

The mean for females in the multiple choice paper was 19, 37 and that for structured test was 12, 42. Males recorded a mean mark of 19, 86 in multiple choice and 13, 61 in the structured paper respectively. The mean for boys was greater than that for girls in the two tests. Boys performed better than girls in the two test formats.

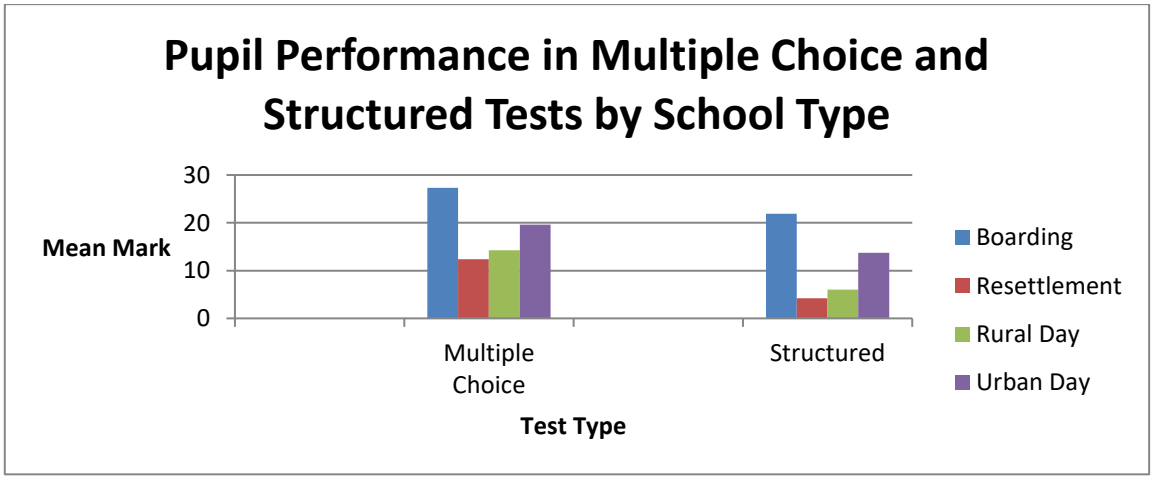
The mean difference for females was 6, 95 while that for males was 6, 25. Everything else being equal, the performance of pupils in both tests was supposed to be the same since pupils who are good in Mathematics should score high irrespective of the test type. Pupils who are also poor in Mathematics are expected to score low marks in multiple choice and structured tests. This was not the case in this study. Therefore the mean difference could be attributed to the guess factor which influenced the rise in marks for the multiple choice paper. Taking this factor into consideration, it showed that the guess factor for females was higher than males. Therefore the study could conclude that the prevalence of random guessing in multiple choice items is high in females than male students.

4.5 Performance of Pupils in Multiple Choice and Structured Tests by School Type

Table 10: Mean scores of pupils by school type

Multiple Choice Scores				
School Type	**Mean**	**N**	**Std. Deviation**	**Std. Error of Mean**
Boarding	27.35	101	10.68	1.06
Resettlement	12.38	63	5.01	.63
Rural Day	14.23	62	7.19	.91
Urban Day	19.59	95	10.24	1.05
Total	19.58	321	10.77	.60

Structured Test Scores				
School Type	**Mean**	**N**	**Std. Deviation**	**Std. Error of Mean**
Boarding	21.90	101	11.49	1.14
Resettlement	4.21	63	4.55	.57
Rural Day	6.02	62	7.05	.89
Urban Day	13.69	95	12.36	1.27
Total	12.93	321	12.23	.68



Information portrayed by the table and the graph shows that pupils performed better in the multiple choice paper than the structured paper. This was the same scenario across all school type categories. For boarding schools, the mean mark for the multiple choice paper was 27, 35 compared to 21, 90 for the structured paper. Resettlement schools recorded a mean score of 12, 38 for the multiple choice component and 4, 21 in the structured component. Rural day schools had a mean mark of 14, 23 in the multiple choice paper while the structured paper had a mean of 6, 02 for the same. For urban day schools, the mean for the multiple choice paper was 19, 59 compared to 13, 69 for the structured component. Over-rally, the means for the multiple choice test were significantly higher than the sister structured paper.

Table 11: Analysis of variance for the difference among means by school type

ANOVA Table							
			Sum of Squares	df	Mean Square	F	Sig.
Multiple Choice Scores * School Type	Between Groups	(Combined)	11134.668	3	3711.556	45.281	.000
	Within Groups		25983.557	317	81.967		
	Total		37118.224	320			

ANOVA Table							
			Sum of Squares	df	Mean Square	F	Sig.
Structured Test Scores * School Type	Between Groups	(Combined)	15942.034	3	5314.011	52.826	.000
	Within Groups		31888.459	317	100.595		
	Total		47830.492	320			

An analysis of variance for the difference in means for multiple choice and structured test scores revealed a significant difference at the 5% level in the means across all school types. There was a significant difference in the performance of pupils in boarding, resettlement, rural day and urban day schools. The performance was also different by test type, with pupils performing better in multiple choice test items than the structured paper.

An analysis of pupil performance by school type was done for all the tests. The results were shown in the table below.

Table 12: Mean performance of pupils by school type

School Type	Mean: Multiple Choice Test	Mean: Structured Test	Mean Difference
Boarding	27,35	21,90	5,45
Resettlement	12,38	4,21	8,17
Rural Day	14,23	6,02	8,21
Urban Day	19,59	13,69	5,9

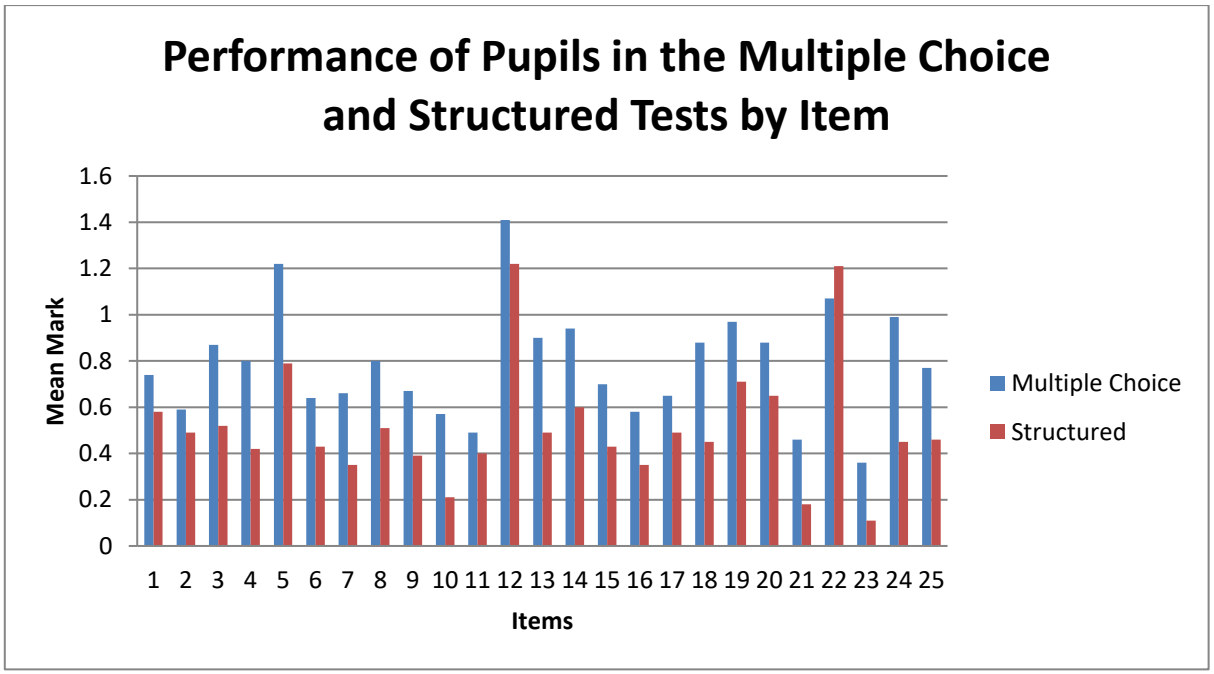
Total	**19,58**	**12,93**	**6,65**

The table shows that boarding schools performed better than urban day schools in both tests. Resettlement schools were the worst in performance in both tests. The mean difference for each school type is reflected in the table. Taking the mean difference as the contributory factor to pupils' random guessing in the multiple choice test, it means that boarding and urban day schools had a low random guessing factor in the multiple choice paper than rural day and resettlement schools. Pupils in rural day and resettlement schools guessed answers in multiple choice tests than pupils in boarding and urban day schools.

4.6 Pupil Performance in Multiple Choice and Structured Tests by Item

Table 13: Comparison of pupil performance in multiple choice and structured tests by item.

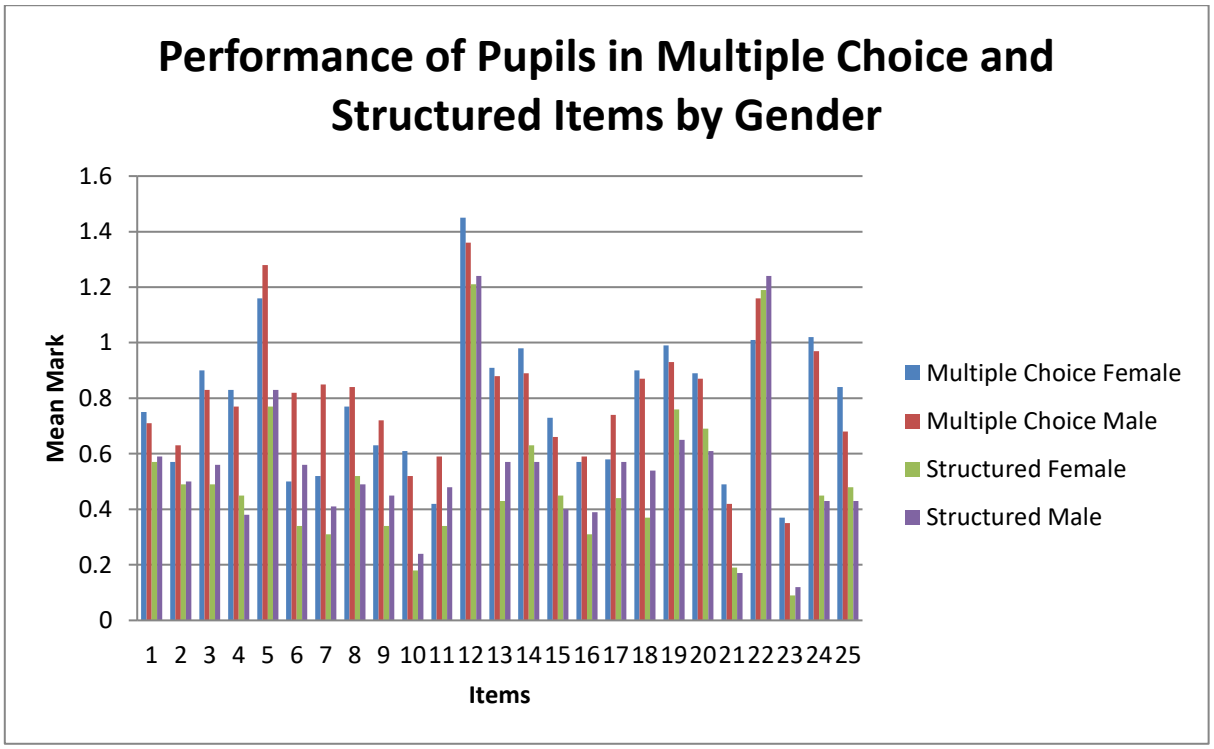
Item	Mean: Multiple Choice Test	Mean: Structured Test	F- ratio	Significance
1	0,74	0,58	161,8	0,000
2	0,60	0,49	252,2	0,000
3	0,87	0,52	60,88	0,000
4	0,80	0,42	126,4	0,000
5	1,22	0,79	204,7	0,000
6	0,64	0,43	129,3	0,000
7	0,66	0,35	59,7	0,000
8	0,80	0,51	76,3	0,000
9	0,67	0,39	92,8	0,000
10	0,57	0,21	65,3	0,000
11	0,49	0,40	141,3	0,000
12	1,41	1,22	104,9	0,000
13	0,90	0,49	75,8	0,000
14	0,94	0,60	95,4	0,000
15	0,70	0,43	80,4	0,000
16	0,58	0,35	79,6	0,000
17	0,65	0,49	405,3	0,000
18	0,88	0,45	48,5	0,000
19	0,97	0,71	138,4	0,000
20	0,88	0,65	319,7	0,000
21	0,46	0,18	52,7	0,000
22	1,07	1,21	49,4	0,000
23	0,36	0,11	72,4	0,000
24	0,99	0,45	37,8	0,000
25	0,77	0,46	65,7	0,000



The table and the corresponding graph illustrate the mean scores of pupils in multiple choice and structured items. An analysis of the table and the graph shows that the means for all multiple choice items are greater than those of structured items with the exception of item 22. The difference in the means was tested by a one way analysis of variance (ANOVA). The results were also reflected in the table. The ANOVA results showed that the difference between the means was significant at 5% level for all the items. It could be concluded that pupils performed better in multiple choice items than the respective free response items.

4.7 Performance of Pupils in Multiple Choice and Structured Test Items by Gender and School Type

The graph shows the performance of pupils in each item by gender.



Mean marks for multiple choice items were greater than those for structured items for both female and male pupils. However, an analysis of variance performed showed no significant differences in the means by gender.

4.8 Discussion of Results

The observed performance differences between boys and girls across the two test formats present compelling avenues for further exploration. While boys, on average, achieved higher scores on both the multiple choice and structured items, the finding that girls demonstrated a relative advantage on the multiple choice format, while boys excelled on the structured items, aligns with existing research on gender and test-taking.

This pattern may be linked to the nature of the formats themselves. As Traub and McRury (1990) found, students often perceive multiple choice tests as easier to prepare for and achieve higher scores on the tests, potentially due to the possibility of guessing. This perception, coupled with potential differences in test-taking strategies employed by boys and girls, could contribute to the observed performance variations. However, the lack of statistical significance for these differences underscores the need for caution in drawing definitive conclusions based solely on this dataset.

The disparity in performance between the multiple choice and structured items, particularly the higher scores on the former, raises concerns about the potential influence of guessing, especially for students in rural day and resettlement schools. This aligns with previous research highlighting the susceptibility of multiple choice assessments to test-wiseness strategies (Simkin and Kuechler, 2005) and random guessing (Cronbach, 1998). The significantly lower scores on the structured items, coupled with a high proportion of unanswered questions, suggest potential challenges in reading comprehension and problem-solving skills, particularly among students in these school settings.

These findings underscore the importance of emphasising mathematical processes and problem-solving strategies in instruction, regardless of the assessment format. As Marzano, et al (2001) emphasise, encouraging students to articulate their thinking processes not only enhances their understanding of mathematical concepts but also provides valuable insights into their learning progression. This is particularly crucial for structured assessments, where the demonstration of problem-solving steps is essential for accurate evaluation of learner performance.

The observed performance discrepancies across school types further highlight the potential influence of instructional practices and teacher preparedness on student outcomes. The lower performance of rural day and resettlement schools on both test formats, particularly the structured items, suggests a need for targeted interventions and professional development opportunities focused on enhancing teachers' capacity to effectively teach and assess mathematical problem- solving skills.

In conclusion, while the study's findings offer valuable insights into the relationship between test formats, student characteristics, and performance, they underscore the need for a multi-faceted approach to Mathematics assessment. Balancing the use of multiple-choice and structured items, coupled with a focus on developing students' problem-solving skills and ensuring equitable access to quality instruction across all school types, are crucial steps towards creating a more equitable and meaningful assessment system in Mathematics education.

5.0 Conclusions and Recommendations

5.1 Conclusions

On the basis of the research findings, the study concluded that:

1. Pupils performed better in the multiple choice test than the free response test. Therefore, pupil competencies in multiple choice test items were not related to their competencies in free response items.
2. The percentage of pupils' correct answers was higher on the multiple choice format than the free response format.
3. The structured paper had several items not responded to (left blank) while the multiple choice paper had no unanswered items. Therefore the free response paper was prone to non-response of items than the multiple choice paper.
4. Boys performed better than girls in both the multiple choice and structured items. Although this was the case, girls tend to favour multiple choice tests than structured tests as revealed by the mean differences.
5. Pupils performed better in multiple choice test items than free response items. However, since the items in both tests were the same with respect to item position, content and skills tested, and phraseology (word for word), the high marks obtained by pupils in multiple choice test items could be attributed to random guessing factor and pupils' testwiseness.
6. The multiple choice test was administered first and the structured test later, and under normal circumstances pupils could have scored better in the structured test because of practice effect and testwiseness. The fact that this was not the case meant that pupils did well in the multiple choice test as a result of random guessing. This meant that in multiple choice tests, sometimes pupils are rewarded for content not mastered and skills not acquired.
7. There was a significant difference in the performance of pupils by school type. Boarding schools were the best in performance in the two test formats, followed by urban day schools, with resettlement schools being the least in performance. When mean differences were considered, that is multiple choice mean minus free response mean, it was found that rural day schools and resettlement schools had higher mean differences, suggesting that pupils in these schools were greatly influenced by the guess factor in the multiple choice test.
8. Pupils in boarding and urban day schools tend to work out problems in multiple choice tests while pupils in rural day and resettlement schools would simply guess the answers. The guess factor in multiple choice tests was high in rural day and resettlement schools. Therefore candidates who might have been considered to have passed Mathematics at Grade 7 level might have passed as a result of random guessing and thus did not pass in the true sense.

5.2 Recommendations

Based on the results of the study, the following recommendations were made:

1. Teachers should encourage and empower pupils to show mathematical processes when solving a problem in both multiple choice and structured items. This would eliminate random guessing in multiple choice items.
2. It was found that multiple choice tests are highly prone to guessing the correct answer and specifically rural day and resettlement schools were found to excel much in multiple choice tests as a result of random guessing and for Zimbabwe, the bulk of grade 7 candidates come from such schools. It is therefore essential to empower teachers in these schools on how to construct valid and reliable multiple choice items and administer them in a manner where pupils show processes of arriving at the correct answer.

3. Teachers must strive to teach reading with understanding in Mathematics. Literacy is key to success in solving mathematical story problems. Assessment of literacy in mathematics would enable pupils understand word problems.
4. Assessment of candidates at grade 7 level should be reversed in terms of paper weightings. Paper 1 which is a multiple choice component should be given a lesser weight than paper 2 which is a free response component. It is in paper 2 where there is evidence of mastery of mathematical content and skills, hence the need to have a higher weighting.

6.0 References

1. Bao, I. & Kilie, U. (2022). An investigation of students' cognitive processes when responding to multiple-choice versus open-ended mathematics assessment items. ***International Journal of Science and Mathematics Education***, 20 (4), 811-830.
2. Boudah, D. J. (2020). The influence of item format on student performance on higher-order thinking skills. ***Assessment and Evaluation in Higher Education***, 45(1), 1-17.
3. Bridgeman, B. (1992). A Comparison of quantitative questions in open-ended and multiple-choice formats. ***Journal of Educational Measurement,*** 29(3), 253-271.
4. Croker, L. & Schmitt, A. (1987). Improving multiple choice test performance for examinees with different levels of test anxiety. ***The Journal of Experimental Education***, 55(4), 201-205.
5. Fleming, K., Ross, M., Tollefson, N., & Green S. B. (1998). Teachers' choices of test item formats for classes with diverse achievement levels. ***The Journal of Educational Research***, 91(4), 222-228.
6. Haladyna, T. M. (1997). ***Writing Test Items to Evaluate Higher Order Thinking Skills***. Allyn Bacon, Boston, MA.
7. Hamilton, L. (1994). *An Investigation of Students' Affective Responses to Alternative Assessments Formats.* Paper presented at the Annual Meeting of National Council on Measurement in Education.
8. New Orleans, LA. Kimball, M. (1989). A new perspective on women's math achievement. ***Psychological Bulletin***, 105, 198-214.
9. Lam,T.C. (1995). Fairness in performance assessment. ***Eric Digest***, Ed391982, 1-6.
10. Livingston, S.A. (2009). Constructed-response questions: Why we use them, how we score them. ***ETS: R&D Connections*** 11, 1-8.
11. Martinez, M. E. (1991). A comparison of multiple-choice and constructed figural response items. ***Journal of Educational Measurement***, 28, 131-145.
12. Martinez, M. E. (2017). A comparison of multiple-choice and constructed-response assessments of common core Mathematics. ***Journal of Educational Measurement,*** 54 (2), 236-253.
13. Marzano, R. (2003). ***What works in schools: Translating research into action***. ASCD.
14. Mazzeo, J. Schmitt, A.P., & Bleistein, C.G. (1991). *Do women perform better, relative to men, on constructed-response tests or multiple-choice tested? Evidence from the Advanced Placement Examination*. Paper presented at the annual meeting of National Council of Measurement in Education, Chicago. IL.
15. Messick, S. (1995). Validity of psychological assessment. ***American Psychologist***, 50, 741-749.
16. Ministry of Primary and Secondary Education (2020). *Annual Report*. Harare, Zimbabwe.
17. Popham, W. J. (2010). **Classroom Assessment: What Teachers Need to Know**? Pearson Allyn & Bacon.
18. Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed response items: A

random effects synthesis of correlations. ***Journal of Educational Measurement***, 40 (2), 163-184.

19. Stenmark, J.K. (1989). **Assessment Alternatives in Mathematics: An Overview of Assessment Techniques that Promote Learning**. Regents: University of California.
20. Snow, R. E. (1993). *Construct validity and constructed-response tests*, in E. B. Randy, and C. W. William (Eds), **Construction versus choice in cognitive measurement**. Hillsdale, New Jersey: Lwarence Erlbaum, 45-60.
21. Traub, R. E. & MacRury, K. (1990). Multiple-choice vs. Free response in the testing of scholastic achievement**.** ***Tests and Trends***, Vol 8, 128-159.