

**QUALITY OF KISWAHILI LANGUAGE TEACHER-MADE TESTS: A CASE
STUDY OF SECONDARY SCHOOLS IN BAHATI DIVISION OF NAKURU
DISTRICT IN KENYA**

BY

DAVID MACHARIA

EGERTON UNIVERSITY LIBRARY

**A Thesis Submitted to the Graduate School in the Partial Fulfilment of the
Requirements for the Award of the Degree of Master of Education (Curriculum
and Instruction) of Egerton University**


**Egerton University
September, 2006**

X

DECLARATION AND RECOMMENDATION

DECLARATION

I declare that this is my original work and it has not been submitted in this or any other form for the award of a degree in this and any other university.

Sign.....
Date..... 04.09.06

David Macharia

RECOMMENDATION

This thesis has been submitted for examination with our approval as university supervisors

Sign.....

Date..... 04.09.06

PROF. M. Ndirangu

Sign.....

Date..... 04.09.06

DR. J. O. Onyango

COPYRIGHT

No part of this thesis may be reproduced in any retrieval system, or transmitted in any form or means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the author or Egerton University on that behalf.

©Macharia Mwangi

DEDICATION

To my father who struggled relentlessly to keep me in secondary school when dropping out was imminent; whose words of encouragement echoed in my heart and inspired me, even when hope seemed to be gone, and gave my life every good reason to live on.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my two supervisors Prof. M. Ndirangu and Dr. J. O Onyango for their unfailing encouragement and support during this strenuous exercise. It is their personal commitment, timely and valuable feedback and professional advice that made this thesis a success.

A special tribute goes to Dr. Fr. S. N Mbugua for his preliminary advice laying the foundation for this research and allowing me to use his personal library. I am also grateful to heads of departments and teachers of the schools visited for their invaluable co-operation. I am also forever grateful to my dear mother who agreed to tender and care for my little daughter throughout the period of my studies. My gratitude also goes to all those who in one way or another contributed to the completion of this thesis.

ABSTRACT

Teacher-made tests perform a fundamental role in the educational process. They motivate students, diagnose their difficulties, select students, place them for special programmes or assignments, evaluate instructional programmes and predict probability of success on a future learning task. Many teacher-made tests have been accused of having very low reliability and validity, testing only the lower cognitive skills and often failing to measure the intellectual capability of the learners. In view of this problem, the study sought to investigate the quality of teacher-made tests and establish some of the factors that hindered teachers from embracing the basic requirements of test construction, resulting in inaccurate tests. The study adopted a descriptive survey research design and it involved all secondary school Kiswahili teachers in Bahati Division. The Division has 39 secondary schools with 76 trained Kiswahili teachers. All the 76 teachers were involved in this study. In addition, 20 heads of department were selected for the interview using simple random sampling. A questionnaire, an interview schedule and a checklist formed the main instruments for data collection. The instruments were moderated after pilot study before collection of actual data. Content validity of the instruments was verified by the study supervisors and other educational experts in the faculty of Education and Human Resources. The Cronbach Alpha formula was used to calculate reliability coefficient. The questionnaire had a reliability coefficient of 0.76. The Statistical Package for Social Sciences (SPSS version 11.5) was used to analyse the collected data to provide descriptive statistics. The major findings of this study were that teachers did not use established psychometric procedures in tests construction. The tests' content validity was not verified, reliability estimates were not established and item quality was not analysed. Tests were found to have low reliability and low discrimination index. The study is likely to make a significant input in changing the language assessment practise of the secondary teachers in Bahati division and will serve as a reference material for teachers, teacher educators, curriculum planners, the Kenya National Examination Council (KNEC), educational administrators and student teachers.

TABLE OF CONTENTS

DECLARATION AND RECOMMENDATION.....	ii
COPYRIGHT.....	iii
DEDICATION.....	iv
ACKNOWLEDGEMENT.....	v
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	xi
LIST OF FIGURES.....	xii
LIST OF ACRONYMS AND ABBREVIATIONS.....	xiii

CHAPTER ONE

INTRODUCTION.....	1
1.1 Background of the Study.....	1
1.2 Statement of the Problem.....	3
1.3 Purpose of the Study.....	4
1.4 Objectives of the Study.....	4
1.5 Research Questions.....	4
1.6 Significance of the Study.....	4
1.7 Assumption of the Study.....	5
1.8 The Scope of the Study.....	5
1.9 Limitations of the Study.....	6
1.10 Definition of Operational Terms.....	6

CHAPTER TWO

REVIEW OF RELATED LITERATURE.....	8
2.1 Introduction.....	8
2.2 Continuous Assessment and National Examinations.....	8
2.3 Uses of Teacher-made Tests.....	12
2.4 Kinds of Tests.....	15
2.5 Testing Strategies.....	17
2.6 Characteristics of Quality Tests.....	19
2.6.1 Test Validity.....	19
2.6.2 Instructional Objectives.....	22
2.6.3 Blooms Taxonomy of Educational objectives.....	23

2.6.4	Table of Specification.....	25
2.6.5	Subject Specialists.....	27
2.6.6	Item Bank (Pool).....	27
2.7	Reliability.....	29
2.7.1	Estimating the Reliability Co-efficient of Teacher-made Tests.....	30
2.7.2	Increasing Test Reliability by Lengthening the Test.....	33
2.7.3	Scorer Reliability.....	34
2.8	Factors Affecting Test Reliability and Test Validity.....	35
2.9	Item Analysis.....	39
2.9.1	Item Difficulty.....	39
2.9.2	Item Discrimination.....	40
2.10	Theoretical Framework.....	42

CHAPTER THREE

RESEARCH METHODOLOGY.....	45	
3.1	Introduction.....	45
3.2	The Research Design.....	45
3.3	Population.....	45
3.4	Sample and Sampling Procedures.....	46
3.5	Instrumentation.....	46
3.5.1	The Questionnaire.....	46
3.5.2	Checklist.....	47
3.5.3	Interview Schedule.....	47
3.6	Reliability and Validation of Research Instruments.....	48
3.7	Data collection.....	48
3.8	Data Analysis.....	49

CHAPTER FOUR

RESULTS AND DISCUSSION	50	
4.1	Introduction	50
4.2.	Validity of Teacher-made Tests	50
4.2.1	Basing Test on Instructional Objectives	51
4.2.2	Using Bloom's Taxonomy of Educational Objectives	52
4.2.3	Using the Table of Specification	54

4.2.4	Use of Subject Specialist	56
4.2.5	Using Item Bank	57
4.3	Reliability of Teacher-made Tests	58
4.4	Item Analysis for Teacher-made Tests	60
4.4.1	Item Difficulty	61
4.4.2	Item Discrimination	62
4.5	Factors Hindering Teachers from Constructing Quality Tests	63

CHAPTER FIVE

CONCLUSIONS, IMPLICATIONS AND RECOMMEDATIONS		69
5.1	Introduction	69
5.2	Conclusion and Implementations	69
5.2.1	Validity of Teacher-made Tests	69
5.2.1.1	Basing Tests on Instructional Objectives	69
5.2.1.2	Using Bloom’s Taxonomy of Educational Objectives	70
5.2.1.3	Using the Table of Specification	70
5.2.1.4	Using other Subjects Specialists	70
5.2.1.5	Using Item Bank	71
5.2.2	Reliability of Teacher-made Tests	71
5.2.3	Item Analysis for Teacher-made Tests	72
5.2.4	Factors Hindering Teachers from Constructing Quality Tests	72
5.3	Recommendations	73
5.4	Suggestions for Further Research	74
REFERENCES		76

APPENDICES

APPENDIX A: TEACHERS’ QUESTIONNAIRE	81
APPENDIX B: INTERVIEW SCHEDULE	87
APPENDIX C: CHECKLIST	89
APPENDIX D: SAMPLE TEST 1	91
APPENDIX E: COGNITIVE LEVELS FOR SAMPLE TEST 1	93
APPENDIX F: SAMPLE TEST 2	94
APPENDIX G: RELIABILITY ANALYSIS	98

APPENDIX H: SAMPLE TEST 399
APPENDIX I: ITEM ANALYSIS103
APPENDIX J: RESEARCH PERMIT 106

LIST OF TABLES

Table 1:	Taxonomy of educational objectives	24
Table 2:	A Typical table of specification	26
Table 3:	Relationship between tests length and test reliability	35
Table 4:	Levels of item discrimination.	42
Table 5:	Variables and their analysis	49
Table 6:	Sources of Kiswahili test items	51
Table 7:	Teachers' responses on the use of Bloom Taxonomy of Education Objectives	52
Table 8:	Teachers' responses on the use of Table of Specifications	54
Table 9:	H.O.D responses on test construction procedures	55
Table 10:	Kiswahili teachers' responses on the use of item bank	57
Table 11:	Teachers' responses on whether they estimated reliability of their tests	58
Table 12:	Reliability Index of Kiswahili TMTs	58
Table 13:	Kiswahili Teachers responses on estimating scorer reliability	59
Table 14:	Kiswahili Teachers responses on whether they conduct item analysis	60
Table 15:	Percentage of <i>D</i> –value for the Kiswahili TMTs	62
Table 16:	Factors hindering teachers from constructing quality tests	64
Table 17:	H.O.D Responses on factors affecting construction of quality tests	65

LIST OF FIGURES

Figure 1:	Test – Item Card (Adapted from Cangelosi, 1990)	28
Figure 2:	Aspects of Systematic Instruction	43
Figure 3:	Percentage of Items in Bloom’s Taxonomy in Tests Constructed by Teachers	53
Figure 4:	Use of Subject Specialists in Test Construction	56
Figure 5:	Percentage of <i>P</i> –value for the Tests	61
Figure 6:	Number of Lessons for Teachers per Week	66

LIST OF ACRONYMS AND ABBREVIATIONS

AEAA- Association of Education Assessment in Africa

AU- African Union

CAT – Continuous Assessment Test

CKRC- Constitution of Kenya Review Commission

HOD – Head of Department

HTSG – Head Teachers Support Group

KCPE- Kenya Certificate of Primary Education

KCSE- Kenya Certificate of Secondary Education

KNEC- Kenya National Examination Council

MOEST- Ministry of Education Science and Technology

SBA – School Based Assessment

TAC- Teacher Advisory Centre

TMT-Teacher-made Test

TSC- Teacher Service Commission

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

Kiswahili language has taken incredibly big strides in the world of languages. For instance, in February 2003, it was formally adopted as one of the official languages of the African Union (AU). At the same time, Yale University in the United States has established a website to popularise Kiswahili globally. Currently, the website, www.yale.edu/swa dictionary has over 60,000 words and is used 30,000 times daily worldwide (Atieno, 2001). Besides being spoken in East Africa, Kiswahili is also spoken in Rwanda, Burundi, Eastern Congo, parts of Zambia and Malawi and even in the Comoros (Mbaabu, 1991).

In Kenya, despite being a national language, the new constitution draft bill recommends that Kiswahili be promoted to an official language (CKRC, 2002). Furthermore, it is one of the compulsory subjects in the primary and secondary schools and is examined at both the Kenya Certificate of Primary Education (KCPE) and the Kenya Certificate of Secondary Education (KCSE). The biggest role played by Kiswahili in Kenya is to enhance national integration and is also used as a vehicle through which national policies reach the common man (Koech, 2000; Mackay, 1985).

The need for an accurate testing of such a language cannot be over emphasized. Testing is an essential part of teaching and learning. The extent to which tests contribute to improved learning and instruction is determined by the principles underlying their construction and use (Gronlund, 1982; Popham, 1990). To a learner, depending on the quality of tests, they can direct students' attention either towards or away from the objectives of instruction, encourage them focus on a limited aspect of the course content (harmful backwash) or direct their attention to all important areas (beneficial backwash) (Hughes, 1997). On the part of the teacher, tests can lead to rewarding superficial learning or in-depth understanding. They can provide

dependable information for instructional decisions, or they can provide biased and distorted information (Hughes, 1997; Popham, 1990; Wood, 1995).

The characteristics of a quality test are reliability, validity and item analysis (Popham, 1990; Thorndike, 1997). Reliability refers to consistency of test judgments and results. On the other hand a test is said to be valid when it is specific in aim, relevant in content and reflects the skill being tested. In addition, item analysis is conducted to establish the individual quality of test items (Davies, 1991; Dreckmeyr & Fracer, 1991; Ebel & Frisbie, 1991; Hughes, 1997; Wood, 1995). Thus, the teacher is able to select, substitute, revise and even discard some test items. Eventually the validity and reliability of a whole test is enhanced. The gravity of the decisions that can be made by inaccurate tests cannot be underestimated. For instance, in biblical times Jephthah used the term *shibboleth* as a test word to distinguish the fleeing Ephraimites (who could not pronounce the *sh*) from his own Gileadites. Those who failed Jephthah's test had their heads chopped off (The Holy Bible, King James Version; 2001). It is unimaginable what great injustice was done to Gileadites who became victims of this invalid and unreliable test (one shot test).

In the Kenyan secondary schools, the largest numbers of tests the learners are subjected to, apart from the KCSE given by the Kenya National Examination Council (KNEC), are prepared by their teachers (Ayot, 1984; Ng'ang'a, 1996). These teacher-made tests (TMTs) perform fundamental roles in the educational process. They motivate students, monitor their progress, diagnose their difficulties, grade them, select, predict their performance and are used to evaluate teaching methods and learning materials (Airasian, 1991; Ebel & Frisbie, 1991; Hughes, 1997; Kimemia, 2002; Sattler, 1992; Thorndike, 1997; Underhill, 1995)

Although the Totally Integrated Quality Education & Training (Republic of Kenya, 1998), KNEC and the Association of Educational Assessment in Africa (AEAA), advocate for the integration of TMTs into the national examinations, the 19th AEAA conference held in Nairobi in 2001 reported that the poor quality of TMTs gave rise to unreliable and invalid assessment (Siringi, 2001). Kellagan and Greaney (2004) observe that although teachers assessment would seem to have the greatest potential to enhance student's achievement, these assessments often are poor quality and are

unlikely to foster the development of higher-order and problem solving competences in the learners. A study by Nga'ng'a (1996) in the Nairobi Province revealed that there was a high correlation between commercially produced mock examination results and the KCSE examination scores. Unfortunately, there was no significant correlation (0.010-0.080) between TMTs results and KCSE examination scores. It was argued that this could have probably resulted from poor testing skills amongst the teachers

Elsewhere in the United States, Anderson (1989) analysed 120 TMTs and found out that in 90% of the cases, no empirical test development procedures, that is, item analysis, table of specification or statement of objectives were given. Reliability estimates were not given in 80% of the cases. In essence, too many significant educational policies are being shaped as a consequence of test results. Too many decisions about the educational effectiveness of teachers and administrators are being based on pupils test scores. Too many citizens are forming their views about public education as a consequence of examination results (Popham, 1991). Teachers therefore, must not only become more knowledgeable regarding test and measurement matters in general, but must also play an active role in improving the testing quality and instruction to profit from their application.

1.2 Statement of the Problem

The major function of classroom tests is to provide evidence on whether instructional objectives have been achieved and to guide students learning. The Kenya National Examination Council guidelines and the Inspectors Manual on evaluating School Based Assessments (SBA) underscore the importance of following psychometric procedures in constructing Kiswahili tests and tests in other subjects as well. However, Kiswahili TMTs have been found to lack a significant relationship with the KCSE results scores. This phenomenon is not unique to Kiswahili as a subject of study. It has been suggested that the lack of correlation could have probably resulted from the way the tests are constructed. In addition, there is a public dissatisfaction in Kenya on the over reliance on one national examination to make a decision about the learner's achievement after many years study. The public concern has been to have TMTs incorporated in the wider concept of Continuous Assessment in all subjects and

be part of the final examination results. In order to achieve this goal, it is important to establish the quality of Kiswahili TMTs in terms of their conformity to psychometric procedures of test construction.

1.3 Purpose of the Study

The purpose of this study was to determine whether Kiswahili language teachers follow procedures that promote quality TMTs and identify the major constraints they may be facing in the construction of quality TMTs

1.4 Objectives of the Study

Specifically the study sought to achieve the following objectives.

- a) To determine whether Kiswahili teachers establish validity of their tests
- b) To determine whether Kiswahili teachers estimate the reliability of their tests
- c) To investigate whether Kiswahili teachers establish the quality of their test items (item analysis).
- d) To establish factors that may hinder Kiswahili teachers from constructing quality tests.

1.5 Research Questions

The study sought to answer the following research questions.

- a) Do Kiswahili teachers establish the validity of their tests?
- b) Do Kiswahili teachers estimate the reliability of their tests?
- c) Do Kiswahili teachers establish the quality of their test items (item analysis)?
- d) What may hinder Kiswahili teachers from constructing quality tests?

1.6 Significance of the Study

This study is significant in a number of ways. First, it comes at a time when KNEC has proposed to incorporate continuous assessment tests into the national

examinations. This will provide KNEC and the general public with information about the quality of Kiswahili TMTs in secondary schools. Secondly, this study is an important reference material for scholars, secondary school teacher educators specifically in Kiswahili language testing and education administrators.

Thirdly, parents, politicians and the community who contribute educational resources and upkeep of the learners in the secondary school will find this study useful. This is because quality TMTs can predict future performance of the learners. Thus, parents and politicians have come to judge schools and teachers on the general performance of the learners in final examination (Hughes, 1997; Rensnick & Klopfer, 1989).

Lastly, curriculum developers in Kiswahili might also find the study important since classroom tests prepared by teachers are developed to assess achievement within a specific curriculum, thereby maximizing curriculum match. They can also detect changes in knowledge arising from the curriculum. (Ng'ang'a, 1996; Salvia & Ysseldyke, 1991; Shapiro, 1989).

1.7 Assumptions of the Study

The following assumptions were made about the study:

- a) All respondents involved in the study were frank and co-operative in providing the required data.
- b) The level of education of the learner would not affect the quality of the tests

1.8 The Scope of the Study

The study involved all trained Kiswahili teachers of Bahati Division of Nakuru District in Kenya. In addition, one end of term test by each teacher was assessed to establish its reliability coefficient, percentage of items at each level of cognitive domain, item difficulty and discriminating power.

1.9 Limitations of the Study

- a) The study only dealt with major factors affecting the construction of Kiswahili teacher- made tests.
- b) Only end of term examinations were used because they were likely to be more comprehensively constructed and thus give more relevant data.

1.10 Definition of Operational Terms

For the purpose of this study, the following terms will have the given operational meanings:

Achievement Test: A test designed to measure effects of specific programme of instruction or how well a learner has mastered instructional content.

Construct Validity: How well a test performance explains possession of certain psychological traits or qualities.

Content Validity: How well a test reflects the content matter or subject taught.

Correlation Coefficient: A pure number, limited by values of -1.00 and +1.00 that expresses the degree of relationship between two sets of scores. (Lyman, 1991).

Criterion Related Validity: Test validity based on a correlation of scores between a test and criterion value.

Criterion-Referenced Test: A test in which the performance of a learner is described by attaining certain set level. (Ebel & Frisbie, 1991; Hughes, 1997).

Diagnostic Test: A test used to identify learners' strengths and weaknesses.

Item Bank: A collection of test items that are all designed to be relevant to the learning objectives.

Item Difficulty: Proportion of learners who scored an item in a test correctly.

Item Discrimination: Ability of a test to separate high achieving learners from low achieving learners.

Norm-Referenced Test: A test in which performance is described relative to a position a learner occupies among the others (Ebel & Frisbie, 1991).

Placement Test: A test used to place learners at a stage or programme most appropriate for their ability.

Proficiency Test: A test used to show whether learners have reached a certain standard with respect to certain specified programme that they are expected to undertake.

Quality Test: A test whose items are reliable, valid, or a representative of the content and elicits the desired outcome.

Reliability: Ability of a test to give consistent results.

Scorer Reliability: The degree of agreement between scores from two or more scorers from a single test.

Taxonomy: A system of classification.

Teacher-made Test: A test prepared by a teacher. Also, classroom test.

Test: A planned process of measuring a sample of behaviour.

Validity: The degree to which a test describes accurately what is being measured.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

2.1 Introduction

This literature review section outlines the fundamental requirements of a quality test. The requirements reviewed include, test validity, instructional objectives, Blooms taxonomy of educational objectives, table of specification, test reliability, scorer reliability, item analysis, use of subject specialists in building of content validity and item bank (sample pool). Also included in the literature review are continuous assessment and national examinations, kind of tests, testing strategies and theoretical framework to guide the study.

2.2 Continuous Assessment and National Examinations.

National examinations are useful in the learning process because they are used to establish and maintain academic standards and encourage a learning society. They have also been used to ease the process of selection and promotion, and sometimes, job placement (Kimemia, 2002; Oirere, 1999). As good as these purposes might look, public examinations can indeed be discriminative and disadvantageous to the learners. Critics of national examinations have argued that the insistence on using one final examination offered at the end of a course as the only sure measure of what the student has learnt in school is faulty and should be re-examined. Musau (2004) argues that the externally imposed examinations do not measure the totality of the learner. While they may measure cognitive achievement sufficiently, they do not measure the affective skills adequately. For instance, such traits as interest, honesty, character and attitudes are often ignored. On the other hand, Continuous Assessment does not confine itself to assessing cognitive behaviour, but also includes attitudinal and motivational characteristics, which permit one to make accurate inferences about the learner (Airasian, 1991).

Further, Oirere (1999) argues that even that which is examined by a final examination is not always the right measure of what a child has learnt throughout the course, for

examinations can be deceptive. Out of the many topics learnt throughout the course, only a few are tested, which implies that all who fail are not backward academically but have only been subjected to a single examination at the end of a course. Many cases have been reported where learners have written their examinations from hospital beds or from police cells. Writing examinations in such situations can adversely affect the performance of the learner and thus give distorted information about the academic achievement of the learner after undergoing the course. In Kenya such a course could be taking 8 years in primary school or 4 years in secondary school.

National examinations for example KCPE and KCSE are used for certification and accreditation. Such examinations designate some pupils as educational rejects. For instance when one scores grade E. Kimemia (2002) notes that even if a child fails an examination after eight years or twelve years of school, the student is not uneducated. In fact, the years of compulsory schooling makes the student learn a lot and gain much more experience than one who never entered the school system.

Students who are designated as 'poor' are coerced to repeat classes. For instance, a student who fails to qualify for entry into public secondary school may be forced to repeat class eight. Kellaghan & Greaney (2004) observe that the low transition between standard 6 and 7 is partly explained by the fact that schools discourage weaker pupils from taking KCPE for fear that their participation would lower the school's mean score. Fear of failure in national examinations is the major reasons why extra teaching and coaching has gained popularity in Kenya. Many of the students whose grades are not good enough resort to private coaching. In the process, anxiety and stress builds up since they have to work over weekends and during school holidays, which denies them the essence of childhood that is to learn, play and have fun at the same time (Oirere, 1999). In addition, many private primary schools relocate the weak students to other schools for fear of lowering their mean grade. (Siringi, 2005). Writing examinations from unfamiliar environment by itself is stressful (Hopkins & Antes, 1985; Popham, 1990),

Closely related to the foregoing, is the observation that external examinations have tended to make teaching examination oriented. Only what is examined is taught. This

increases pressure on students to pass, which leads to a situation where testing dictates what has to be taught rather than the goals, objectives, values and perceived needs of the society as envisaged in the curriculum (Anderson, 1989; Musau, 2004). In the end, those aspects of the curriculum that are not examinable are ignored hence narrowing down the curriculum and encouraging rote learning (Kellaghan & Greaney, 2004)

Due to public dissatisfaction with over reliance on the results of one national examination given at the end of the study, there are concerns to incorporate TMTs as part of broader concept of Continuous Assessment (CA). Teachers have used TMTs as measures of academic achievement and used the results to make inferences about their pupils. The nature of continuous assessment and the inferences which result are wide ranging, encompassing not just pupils academic performance but their motivation, self-concept, interests, attitudes and values as well (Airasian, 1991). All assessments contain some error and imprecision and relying heavily upon a single assessment for decision making can lead to incorrect decisions.

Continuous Assessment has an advantage of being guidance oriented. Since it involves data gathered over along period of time, it will yield more accurate data reaching the teacher early enough to modify instruction. This will play a vital role in the diagnosing and remediating areas of learners weaknesses (Ebel & Frisbie, 1992; Popham, 1990). Continuous Assessment also fosters pupil- teacher relationship based on individual interaction. Pupils learn that the teacher values their achievements and that their assessment outcomes have an impact on the instruction they receive. One-to- one communications between teachers and pupils can motivate pupils to continue attending school and to work hard to achieve higher levels of mastery.

Airasian, (1991) and Kimemia (2002), observe that CA allows teachers to evaluate the effectiveness of their teaching strategies and teaching aids relative to the curriculum and those teaching techniques and teaching aids as dictated by the needs of their pupils. In addition, CA provides information on achievement of certain marks or scores. Thus CA enable teachers to monitor achievement of various cognitive levels before it is too late to achieve them (Bolyard & Hatch, 2003).

Finally, teachers can share assessment results with important stakeholders such as parents, other teachers, community members and learners themselves. Regular reports from the teachers based on CA allows the parents to know about their children's progress. Armed with this knowledge parents can assist and support children with studies and probably buy other supporting educational materials. It can therefore be argued that if examining of students would involve keeping and using of school CA done throughout the students class work and years in school, it would form a more reliable assessment than a single examination at the end of the course (Kirere, 1999).

Some countries in Africa like Namibia, Malawi and Swaziland have involved scores generated from CA in the final assessment of learners in primary schools. In other countries like Zimbabwe, South Africa and Zambia, there are plans to incorporate CA in national examinations (Bolyard & Hatch, 2003; Kellaghan & Greaney, 2004).

In Kenya, a committee to look into the possibility of incorporating CA into the national examinations was set up in the year 2001 (Siringi, 2001). Although this initiative was hatched at this time, the idea of incorporating classroom tests in the final examinations has been very popular among Kenyans as it is reflected in Sessional Paper number 10 of 1988, Mackay Report of 1985 and Totally Integrated Quality Education and Training report of 1998 (Musau, 2004; Njoroge, 1996; Siringi, 2001).

If this dream shared by many Kenyans were to be realised, all the teachers in the country need not only to be equipped with test and measurement skills but also apply them in order to construct quality tests. However, a joint communiqué issued by AEAA conference held in Nairobi in 2001 indicated that the greatest impediment to incorporating CA in national examinations would be the quality of tests developed by teachers (Siringi, 2001).

A study carried out by Ng'ang'a (1996) revealed that there was a high correlation between mock examinations and KCSE results scores but there was no significant relationship between TMTs and KCSE scores. Ng'ang'a argues that this could have resulted from the poor testing skills employed by the teachers. There was need therefore to establish the quality of tests constructed by teachers in Kenya before

considering incorporating them into the national examinations. If teachers are to construct valid and reliable tests, it is imperative that they follow psychometric procedures and practices in test construction (McMillan, 2000). Thus, there was need also to investigate what could be hindering teachers from applying these psychometric procedures in test construction.

2.3 Uses of Teacher-made Tests

Tests constructed by teachers in the classroom can motivate learners. Anticipated tests act as extrinsic motivation to learning while internal desires or needs act as intrinsic motivators. For most learners, motivation provided by tests is indispensable. What learners stand to gain or lose in a given situation is motivating to all. Tests help learners to make many of their decisions about trying to learn, how hard to try and when to stop trying. The experience of almost all students and teachers support the position that students tend to study harder when they expect a test than when they do not (Ebel & Frisbie, 1991; Lien, 1980). Gronlund (1982) supports this view by observing that periodic testing motivates students by providing them with short-term goals ^{which} they strive to achieve. For this reason alone it is essential to give periodic tests. The more times a learner looks over or revises a certain piece of knowledge, the more chances he has that he will recall the material in future. However, it is only feedback from well constructed tests that can motivate students to improve on their performance (Thorndike, 1997). Tests that are poorly constructed or used punitively can just as effectively discourage the learners or misdirect their learning. The way a teacher tests affects the way a learner learns. If a teacher wants to encourage the learner to develop good learning habits, he must make sure that the test items he uses motivate learners to do so.

TMTs are also used to diagnose students learning. To diagnose is to determine the nature of a difficulty whether it be a disease or a misconception (Ashworth, 1982). On marking the students' tests, a teacher can find out that a large percentage of the students had difficulty with a particular question. It becomes possible to find out the cause of the difficulty by studying how the learners answered the question. Airasian (1991) and Gronlund (1982) assert that much of the assessment data teachers gather

from classroom tests is used to identify, understand and remediate pupils problems and learning difficulties.

Teachers often use the tests they construct to give reports about learners achievement to parents, head teachers, the learner himself and sometimes to external examination boards. Most parents are vitally concerned with the progress of their children in school. They may differ in the depth of this concern and differ even more in the resources they provide for coaching in school skills and for supporting their children's efforts to learn. But all parents must be considered partners with the school. And if they are to be effective partners, they must know how their children are progressing. Parents need to know the level at which their children are functioning in each school subject and be warned promptly of any potential difficulties (Thorndike, 1997). All kinds of test results provide a concrete basis for communication from school to parents and for interpretation of the child's progress and difficulties. For most TMTs a norm-referenced interpretation is provided that compares the child's general level of performance in a subject area with that of his/her class, other students in the school or a broader regional group.

Classroom tests can be used to select students. Decisions about permitting students to pursue certain courses. Admitting them to colleges and selecting them for certain occupations depends largely upon judgements recorded by teachers (Davis, 1991; Ashworth, 1982). In addition, teachers help students in making decisions concerning their future undertakings such as a course of study. This would avoid choice of subjects or study courses that one is not competent in. Furthermore, teachers frequently use test scores to divide classes into smaller in-class groups for instruction. The primary purpose for such grouping is to make instruction more effective by forming small clusters of students who have similar instructional needs. The type of tests teachers produce to select students depends on the type of students required. If a teacher wanted to select students who would be entered for scholarships, the teacher would construct a test to cover as a large content as possible and choose difficult questions. On the other hand, if a teacher wished to find out which students required remedial treatment, he would design a test containing many easy questions that only very poor students would fail.

Well constructed TMTs can be used to predict future performance of learners (Hopkins & Antes, 1985; Thorndike, 1997). As learners progress through the school system they gradually take on more responsibility for decisions concerning their future educational decisions. Past achievement is one type of information that could influence their future achievement. The measurement of present achievement is the best predictor teachers have for future achievement. For example, a well constructed Kiswahili TMT can predict learners' performance in Kiswahili at KCSE. The present level of achievement may therefore affect the learners' future educational decisions in that the performance in a particular course at present may influence his or her decisions about educational specialization in future. It is important to note that poor quality tests may predict the future imperfectly.

Tests are often used by teachers to assign grades. Although the nature of grading may vary from school to school, it is true that all schools require classroom teachers to judge and grade the academic performance of their pupils (Airasian, 1991). Grades are deeply embedded in the educational culture. Since they are entered into permanent school records, they have become the basis for a wide range of actions within the school, between schools and with the outside world. For instance, eligibility for admission to certain programmes or departments, for scholarship or bursaries and for continuing in school is often determined by academic standing. Thus, there are many points within the educational system where grades interact with the administrative and instructional process to affect the students' progress (Popham, 1990)

Most theorists from behaviourists to cognitive psychologists have emphasized the need for feedback in the facilitation of learning (Thorndike, 1997). We know that it is difficult for someone to improve unless they know how well they are doing. The process of growing and developing requires the testing of limits. For instance, Children need to know how they stand in a Kiswahili test in relation to peers and some goal. Parents too need to have this information with them. It is therefore critical that the feedback be as accurate as possible and this can only be achieved through quality test instruments. Furthermore parents have a right to the accurate reporting of their children's progress in terms they understand. Avoiding the anguish of assigning grades by only giving high grades is a dereliction of duty. Nothing is more damaging to parents-school rapport than to have them erroneously believe that their children

have no academic problem (Thorndike, 1997). The school also has the responsibility to certify that students have mastered the assigned curriculum in the course they have taken by giving accurate grades. If a student receives a satisfactory grade in a course, it is reasonable for a prospective institution or college to assume that the grade represents a meaningful mastery of subject matter.

Teacher-made tests are used in research in which language testing is used to test hypotheses in relation to our understanding of language and learning (Davies, 1991). For instance, the status and concept of Kiswahili language proficiency, and the structure of language ability have been discussed in recent years by Kiswahili language teachers using language testing techniques to produce data. The results of Kiswahili TMTs have been used to provide such data.

A major function of a classroom test is to evaluate student achievement of intended course content (Ebel & Frisbie, 1991). Tests provide evidence on whether objectives of instruction have been achieved. Ashworth (1982) observes that teachers should not only test the knowledge their students have acquired but should also test the students comprehension of the subject matter, their ability to apply knowledge, analyse new situations using processes he has previously learnt, synthesise information and ability to evaluate. Besides, TMTs help to determine whether methods and materials of instruction were appropriate and how well learning experiences were sequenced (Thorndike, 1997)

2.4 Kinds of tests

Most authorities in educational measurement concur that four types of tests exist. These are, proficiency, diagnostic, placement, and achievement tests (Anastasi, 1982; Davies, 1991; Hughes, 1997; McArthur, 1991).

Proficiency tests in language are designed to measure the learner's general ability (MacArther, 1991; Underhill, 1995). Proficiency means having sufficient command of the language for a particular purpose. Hughes (1997) adds that proficiency tests measure a pupil's ability in a language regardless of any training they may have had

in that language. A candidate's ability is assessed according to how far it matches certain criteria judged to be essential for proficiency in a particular task.

The content of the proficiency test therefore, is not based on the content or objectives of a language course, which people taking the test may have followed. Proficiency tests used for predicting performance in the language being tested on some future activity. An example of this would be a test designed to determine whether a student's English is good enough to follow a course of study at a university. Such a test may even attempt to take into account the level and kind of English needed to follow Arts subjects, Science subjects and so on (Davis, 1991). However, according to Davis (1991) and Hughes (1997), most of these tests are commercially produced in countries where they are used. Such examining bodies are independent of teaching institutions. Teachers in the class are hardly involved in the construction of such tests.

The other kinds of tests are the diagnostic. Diagnostic tests are used to identify students' strengths and weaknesses (Hughes, 1997; Underhill, 1995). According to Davies (1991), diagnostic tests are interested on the failure- what has gone wrong in order to provide remedies. It does not produce information in the form of a score. But as a list of areas in which the learner is strong in and those in which he needs further practice or remedial work.

This type of test typically includes a relatively large number of test items in each specific area with slight variations from one item to the next so that the cause of specific learning errors can be identified. The intention is to probe deeper into the causes of learning deficiencies that are left unresolved by achievement tests. However, very few tests are constructed for diagnostic purposes (Underhill, 1995). In a similar observation, Hughes (1997) says that most classroom teachers rarely construct tests specifically for diagnostic purposes. Lack of good diagnostic tests is unfortunate. Such tests could be extremely useful for individualised instruction or self-instruction. Learners would be shown where gaps exist in their command of language and could be directed to sources of information, exemplification and practice.

A placement test is used to identify the right level of the learner's ability for example in language (Davis, 1991). These kinds of tests are intended to provide information, which would help to place a learner at the stage in a learning programme most appropriate to their abilities. Airasian (1991) underscores the importance of placement tests by observing that teachers who misjudge the levels of their pupils' ability fail to review or point out important concepts to pupils.

Achievement tests form the bulk of teacher-made tests. They are designed to measure the effects of specific program of instruction (Anastasi, 1982; Gronlund, 1982; Murphy, 1996). For instance they can take a sample of the language skills that have been covered on the course and test how well the learner has mastered those elements or achieved the intended objectives. Achievement tests are typically used at the end of a period of learning or school year. The content of the tests are samples of what has been in the syllabus during the time under scrutiny (Davis, 1991). In contrast to proficiency tests, achievement tests are directly related to language courses.

Achievement test are of two kinds:

- Final achievement tests
- Progress achievement tests

Final achievement tests are administered at the end of a course of study. They may be written and administered by the Ministry of Education or official examining boards. In Kenya, such tests are KCPE and KCSE. Progress achievement tests as their name suggests are intended to measure the progress that the students are making. Since progress is towards achievement of course objectives, these tests are related to instructional objectives. Tests objectives provide more accurate information about individual and group achievement and it is likely to promote more beneficial effect or teaching (Hughes, 1997).

2.5 Testing Strategies

There are two most common strategies in educational testing. These are: norm-referenced testing and criterion-referenced testing. Although there are similarities between these two approaches, there are also fundamental differences between them.

These differences hung on the interpretation of the test scores and the way the two tests are constructed (Gronlund, 1982; Popham, 1990)

A norm-referenced test typically measures a more general category of examinees' competence. A criterion-referenced test on the other hand typically focuses on a more specific domain of examinees behaviour (Anastasi, 1982; Ebel & Frisbie, 1991). Because of their breadth, norm-referenced tests can provide an overall estimation of how well the examinee has performed regarding the general field of skill or knowledge to be measured. On the other hand, criterion-referenced test will not attempt to assess comprehensive mastery of the field. Instead, major subskills within the field will be measured (Popham, 1990). These subskills are measured with more precision by more items per attribute. This is because mastery testing is not applicable beyond these specific subskills (Anastasi, 1982; Gronlund, 1982). In essence, several criterion-referenced tests each measuring its own specific domain would be required to tap the skills and knowledge being assessed by a typical norm-referenced test.

The second significant difference is the way the test scores are interpreted. For both criterion and norm-referenced tests, the examinees raw scores are first secured. However, it is what is done thereafter with the scores that make the difference. A norm-referenced test is used to ascertain an individual's status with respect to the performance of the other individuals on that test (Ebel & Frisbie, 1991; Hughes, 1997). It does not directly tell us what the student is capable of doing. On the other hand, in criterion-referenced testing, an examinee test performance is reported in terms of specific skills he has mastered (Davies, 1991; Popham, 1990). It does not matter in principle whether all the candidates or none of them is successful. Those who perform satisfactorily 'pass' and those who don't, 'fail'. (Davies, 1991; Hughes, 1997).

It is however difficult to see how criterion-referenced tests can be constructed in a completely separate way from norm-referenced tests, that is, without the usual canons of item discreteness and discrimination especially for TMTs, which are based on instructional objectives (Anastasi, 1982; Gronlund, 1982). It can therefore conclude that for the TMTs given in the classroom, norm referencing and criterion referencing

are essentially two sides of the same phenomenon. One side looks at what pupils are capable of doing and how best they can do it, the other side looks at what they need to do (Davis, 1991).

2.6 Characteristics of Quality Tests

Teaching and testing are twins, complementing and integral aspects of the teaching process. Therefore, if objective evaluation of learning outcomes is to be achieved, every teacher must be thoroughly aware of the techniques of constructing quality classroom tests. This is only possible by following psychometric procedures that characterise construction of quality tests. These include:

- Establishing test validity
- Estimating tests reliability
- Conducting item analysis

2.6.1 Test Validity

According to Hughes (1997) and Ayot (1984), a test is said to be valid if it measures accurately what it is intended to measure. This agrees with Thorndike and Hagen (1977) who adds that validity of a test refers to the ability of a test to measure, all of what it is intended to measure and nothing besides that. The Kenya National Examination Council in its guidelines on evaluating School Based Assessment (SBA) underscores the importance of establishing test validity for Kiswahili TMTs and other subjects (Mucheru, 2005). It is important to remember three things about validity. First even when a test is valid, it is only valid for a specific skill: a valid test for grammar cannot be used as a test for writing. Secondly, a valid test taps the central issues in the unit of work being tested. Thirdly a valid language test depends on the linguistic content of the test as well as technique used (Cangelosi, 1990). For example, a test of the speaking skill, which uses a perfectly valid conversational situation, but does not test elements of language such as pronunciation, stress or intonation, is not valid test of speaking skill (Ayot, 1984). Several writers, including, Airasian (1991), Ebel & Frisbie (1991), Hughes (1997), Popham and (1990) have

identified four types of validity, namely face validity, criterion-related validity (predictive and concurrent), construct validity and content validity.

A test is said to have face validity if it looks as if it measures what it is supposed to measure (Dreckmeyer & Fracer, 1991; Hughes, 1997). For example, a test, which pretends to measure pronunciation ability but did not require the candidates to speak, might be thought to lack face validity. Another example is a test for computer application packages that is given as a theory paper and does not require learners to manipulate the computer during the test. Gregory (1996) and Dreckmeyer & Fracer (1991) argue that face validity is a matter of social acceptability, and not a technical form of validity in the same category as content, criterion-related and construct validity. This is because a test, which does not have face validity, may not be accepted by education experts or learners themselves. The candidates' reaction to it may mean that they do not perform on it in a way that truly reflects their ability. Thus, from public relations stand point it is crucial that tests possess face validity. Otherwise those who take the tests may be dissatisfied and doubt the value of the test.

Another approach to test validity is criterion-related validity. This approach to validity is used to determine how far the results on the test agree with those provided by some independent and highly dependable assessment of the candidates' ability (Salvia & Ysseldyke, 1991; Thorndike, 1997). That is, some other measure is taken as the criterion of 'success' and the test is judged in terms of its relationship to that criterion measure. A criterion is an accepted standard against which a test is compared to, in order to validate its use as a predictor. For example, scores of a dictation test can be accepted generally as a measure of spelling achievement. If a true-false spelling test were to be constructed and the scores obtained were compared with scores of a dictation test, if it were found that there is a high correlation between the two tests, the true-false test will be an acceptable measure of spelling achievement. There are essentially two kinds of criterion-related validity i.e. concurrent validity and predictive validity.

In concurrent validity, the learner's scores in a test are correlated with an established test, which has already been independently validated. Thus determining the present standing on a criterion measure (Ebel & Frisbie, 1991; Underhill, 1995). This would

produce a correlation coefficient, which suggests the extent to which the tests are measuring the same thing. According to Gregory (1996) and Hughes (1997), concurrent validity is commonly done when a test is being considered as a replacement for a more time-consuming method of obtaining information. For instance, the true-false test referred to above can be used instead of the dictation test because of the increased efficiency in scoring afforded by the true-false test.

On the other hand predictive validity is concerned with the degree to which a test can predict candidates future performance (Davis 1991). The concern here is to show that a positive relationship exists between scores on the test (the predictor) and scores on some acceptable measure of future performance (criterion). In this case a learners scores on a language test are correlated against performance on some other test or task at some future point. When the correlation is high, it can be concluded that the test is a useful predictor of future performance. That is, there is support for using the scores to predict success in a future test.

Construct validity becomes a prime concern in testing when an individual test performance is used as a basis for inferring the possession of certain psychological traits or qualities. Some examples of a construct include anxiety, creativity and reading. No single type of evidence is satisfactory for determining construct validity. What is important is to make predictions that are in harmony with the theory underlying the particular construct. (Ebel & Frisbie, 1991; Davis, 1991). If the data are in harmony with our predictions, they support the validity of the interpretation of the scores as a measure of the construct. For instance, the theory of writing tells us that underlying the writing ability are a number of sub-abilities (constructs) such as control of punctuation, sensitivity to demands of style and so on. Tests can then be construct that are meant to measure these sub-abilities. The scores obtained will indicate the writing ability of a learner.

However, construct validity is more concerned with the validity of a hypothetical construct measured by a particular test like aptitude tests and personality scales rather than achievement tests. In addition, measurement experts usually opt for construct related validation strategies when they are attempting to assess attributes such as those found in affective domain (Ebel & Frisbie, 1991; Hughes, 1997; Popham; 1990).

The final and the most important type of validity for TMTs is content validity. Content validity is the prime concern of achievement tests (Anastasi, 1982; Ayot, 1984; Hughes, 1997). Since as indicated earlier that TMTs are mainly achievement tests, the study was therefore concerned with content validity of Kiswahili tests. When dealing with content validity the concern is on how well the test measures the subject matter and the expected learning outcomes achieved during the instructional process (Davis, 1991; Hopkins & Antes, 1985). Moreover, the key element in content validity is the extent to which a test measures a representative sample of the larger domain it is supposed to represent (Gronlund, 1976; Popham, 1990). It can therefore be argued that a test without a high degree of content validity cannot be expected to provide highly valid data. Thus, judgments based partly or wholly on that information would lack validity (Hughes, 1997).

The establishment of the content validity of a classroom test cannot be done by any straightforward statistical procedure. The KNEC guidelines on the construction of SBA states that teachers can only be sure that their tests are measuring what they are intended to if they followed procedures that promoted content validity (Mucheru, 2005). Establishment of content validity becomes primarily a matter of presenting a logical approach characterized by clearly stated objectives, table of specification, clearly stated levels of cognitive domain (Blooms taxonomy), use of subject specialists and building of item pool (Hopkins & Antes, 1985; Marshall & Hales, 1977; Popham, 1990).

2.6.2 Instructional Objectives

Educational objectives are statements that describe the behaviour that pupils are to perform after instruction (Airasian, 1991). Other names for educational objectives are, instructional objectives, learning objectives, behavioural objectives, and performance objectives (Airasian, 1991; Thorndike & Hagen, 1977). Educational objectives serve important functions in the instructional process namely:

- Identifying intended pupils' outcomes.
- Providing direction for the teacher in selecting instructional materials and activities

- Providing basis for testing.

A teacher must have a set of educational objectives to serve as a sounding board for his tests (Ashworth, 1982; Wood, 1995). This approach is based on the premise that if teaching has been directed by instructional objectives, and the test reflects the teaching, then the test will measure how well the instructional objectives have been achieved by the learners. If a test is to achieve this goal, instructional objectives must be clarified and stated in terms of observable student behaviour (Hughes, 1997; Marshall & Hales, 1977). KNEC emphasises the importance of instructional objectives in building content validity for Kiswahili tests and others subjects. KNEC guideline states that it is not enough for the teacher to merely write objectives in his/her lesson plan. The teacher also needs to thoroughly understand the curriculum and the subject matter to be able to come up with relevant and measurable instructional objectives (Mucheru, 2005). Consequently, it can be argued that if the instruction has been geared towards these objectives, and a test is composed of tasks that measure the degree to which the student has achieved the objectives, then the test can be said to have content validity (Anastasi, 1991; Hopkins & Antes, 1985).

2.6.3 Blooms Taxonomy of Cognitive Domain

Perhaps the best framework for grouping objectives and helping teachers maintain the necessary general perspective is the taxonomy of educational objectives as propounded by Bloom and Krathwol. The taxonomy classifies instructional objectives into cognitive, affective and psychomotor domains (Gronlund, 1982; Popham, 1990; Smith, 1984). The cognitive domain includes all intellectual or thinking function, understanding, analysing interpreting etc. The affective domain includes all the feelings and emotions: believing, valuing, desiring etc. The psychomotor domain includes all physical behaviour, throwing running, jumping etc.

Schools attempt to develop behaviour in all the three domains but examinations concentrate mostly on the cognitive domain (Ebel & Frisbie, 1991; Gronlund, 1982). Because of this tradition, this research will be more concerned with testing of the cognitive domain in Kiswahili. Cognitive skills can be classified into six major areas

from the most simple and concrete to the most complex and abstract (Dreckmeyr & Fracer, 1991; Popham, 1990; Smith, 1984). Table 1 illustrates Blooms taxonomy of educational objectives.

Table 1

Taxonomy of educational objectives

Taxonomy categories	Intended learning outcomes
Knowledge (Recalling previously learnt material)	Identifying, naming, defining, listing, matching, describing, listing, selecting
Comprehension. (Grasping meaning of material)	Classifying, explaining, summarizing, converting, predicting, distinguishing
Application. (Using learnt material in concrete situation)	Demonstrating computing, solving, modifying, arranging, operating, relating
Analysis. (Breaking materials into its component parts)	Differentiating, diagrams, estimating, separating, inferring, ordering, subdividing
Synthesis. (Putting parts together into a whole)	Combing, creating, formulating, designing, composing, constructing, rearranging, revising
Evaluation (Making judgment about the value of material)	Judging, criticizing, comparing, justifying, concluding, discriminating

Source: Adapted from Ebel & Frisbie (1991) and Gronlund (1982)

The six categories generally are cumulative, which means that a skill classified as high level category includes all the categories below it. By classifying each objective according to the taxonomy, a Kiswahili teacher can know clearly if he/she is teaching and testing only for memorization and recall or if he/she is teaching and testing higher mental skills. According to the Inspectors' Handbook, Bloom's taxonomy of

Educational Objectives provides a guide in how test items should be stated. The way the stem of the test item is framed will indicate whether the teacher is testing more of low or high cognitive skills (MOEST, 2000). However, Thorndike and Hagen (1977) argued that several studies (Lawrence, 1963; Pfeifer, 1965; Scannell & Wagen, 1960) had shown that the majority of items on teacher-made tests (as many as 98%) required only rote recall of very specific information. Ashworth (1982) echoes the same opinion when he notes that more often than not, trivial details are tested by teacher-made tests while understanding, reasoning, creativity and practical application and ability are neglected.

2.6.4 Table of Specification

If a classroom test is to be useful for a specific purpose, it must be built from a designed plan, much as physical structure is built from its design plan. Building trade people call their plan blue print. Test constructors call their plan for test, a table of specification or a test blue print (Hopkins & Antes, 1985; Popham, 1990). A table of specification or the two-way test grid simultaneously lists the content and level of cognitive skills required on a test so that they can be consistent with the instructional objectives (Dreckmeyr & Fracer, 1991; Popham, 1990). It relates the outcome to content and indicates the relative weight to be given to each area depending on the emphasis given during instruction (Dreckmeyr & Fracer, 1991; Gronlund, 1982).

Clearly, two-way grid can be helpful in apportioning items on tests so that teachers can avoid inadvertent overemphasis or under emphasis (Popham, 1990; Smith, 1984). In essence, a careful use of the table of specification can help teachers ensure the validity of their tests, both by content and by level of cognitive skills employed. It is the only assurance that a classroom test has content validity (Ebel & Frisbie, 1991; Gronlund, 1982). Table 2 illustrates a typical table of specification

Table 2:

A Typical table of specification

Content \ Objectives	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
	A. Sarufi (Grammar)						
1. Kuafikisha		1				2	3
2. Ngeli	1					1	2
3. Matumizi ya 'vile'	1	1				2	4
4. Vivumishi na viashiria	1					1	2
5. muundo wa Sentensi			1	1	1		3
B. Msamiati (vocabulary)							
1. Maneno yenye maana sawa	2					1	3
2. Maumbo	1		1				2
3. Tashbihi	1	1	1				3
4. Methali	1	1					2
5. Semi	1	1	1				3
Total	9	5	4	1	1	7	25

Source: Adapted from Popham (1990) and Hopkins & Antes (1985)

The Ministry of Education in Kenya expects that Kiswahili teachers draw a table of specification for every test they give. At a glimpse, the table of specification should show relative emphasis given to various sub-topics and the relevant cognitive skills tested (MOEST, 2000). In addition, the KNEC guidelines on construction of SBA states that the test-blue print guides not only the test constructor himself but also provides substantial information to another teacher who would test the same class. It further provides documentation evidence that all long the learners have been subjected to valid tests (Mucheru, 2005).

2.6.5 Subject Specialist

The teacher should make a judgement about the degree of content validity for classroom test. A test should represent a defined universe of content. Because of the breadth of many content fields, it is impossible for a test to measure exhaustively the universe of content involved. Thus, for a test to represent the universe of content properly, it should sample the major aspects of that universe. This is where the teacher's judgement is needed. It is also important for a teacher to make use of a second judgement of a knowledgeable colleague (subject specialist) to supplement his judgement as a check on bias (Hopkins & Antes, 1985; Popham, 1990).

There are two basic strategies that can be applied separately or in concert by a teacher to secure the content validity when he uses a second judgement. First, attempts to incorporate suitable content on the test can be carried out during the test development process. A Kiswahili teacher can call other Kiswahili subject specialist who can exercise their judgement in an effort to build a test that satisfactorily represents a domain of the content intended.

The second strategy is to subject a test to a series of post facto judgement about the representativeness of its content. For example independent Kiswahili subject expert(s) can be asked to review a test, item by item, to see if its items satisfactorily represent the domain of content involved. A personal judgement supported by an 'outside' estimate is more valuable and encourages teachers to construct more valid and meaningful tests (Popham, 1990).

2.6.6 Item Bank (Pool)

In describing the quality of a good test, Davis (1991) stated that, if a test is to reveal how much a pupil knows, it must sample adequately. It must contain enough questions to be truly representative. It is obviously true that the proportion of items in a sample a person gets correct is an estimate of the proportion he would get correct if he answered every item in the item bank. Ayot (1984) suggests that a good test item that has been validated and proved reliable should be saved for later use either with

that class or with another group. Cangelosi (1990) underscores the importance of the item pool by proposing that:

- The act of building item pool focuses attention on one objective at a time, stimulating the teacher to expand his or her ideas on how student achievement in each objective can be tested.
- A test constructed with items drawn from the item pool is more likely to be relevant than a test with items that are designed as the test is being put together.
- Having an item pool makes it easier to construct the test according to the table of specification.
- Having an item pool makes it easier for one to improve tests.

In most schools, item banks are test-item files containing item cards. However, in some institutions computers are used to store test items (Cangelosi, 1990). A typical test-item card is shown in figure 1.

Front side of the Card	Reverse side of the card		
<p>TOPIC: Vivumishi</p> <p>OBJECTIVE: Kufikia mwisho wa kipindi mwanafunzi aweze kueleza na kutumia vivumishi katika sentensi</p> <p>ITEM: Eleza aina za vivumishi vilivyotumiwa katika sentensi hizi:</p> <p style="padding-left: 40px;">(a) Mtu <u>huyu</u> atatumikiwa zawadi.</p> <p style="padding-left: 40px;">(b) Mti <u>mrefu</u> umekatwa</p> <p style="padding-left: 40px;">(c) Kiatu <u>chake</u> kimepotea.</p> <p>SCORING KEY:(1 mark for correct answer)</p> <p style="padding-left: 40px;">(a) Kivumishi Kiashiria</p> <p style="padding-left: 40px;">(b) Kivumishi cha sifa</p> <p style="padding-left: 40px;">(c) Kivumishi kimilikishi</p>	Date	Item <i>p</i> - value	Item D- value

Figure 1: Test –Item Card (Adapted from Cangelosi, 1990)

The front of the card contains information like objectives, test-item and the scoring key. The reverse side contains information like date(s) used, percentage of students answering the items correctly (item difficulty) and how well the item discriminated

between the better and the poorer students (Lyman, 1991; Cangelosi, 1990). Over a period of time, a department of Kiswahili can build a whole set of good items for each class.

2.7 Reliability

It is expected that teachers of Kiswahili and other subjects in Kenya should know how to construct tests with the highest possible reliability. This can only be done in the knowledge that calculating reliability estimates indicates how reliable their tests can be (MOEST, 2000; Mucheru, 2005). According to Wood (1995) reliability refers to the consistency of measurement, that is, how consistent test score or other evaluation results are from one measurement to another. Ebel and Frisbie (1991) propound that the reliability of a test is its ability to give similar results for the same group of students if given at different times or if marked by one or more markers on the one or more occasions. Reliable Kiswahili tests should yield the same score if they were administered to the same pupils in circumstances that remained constant.

Some variation in score can be expected due to temporary fluctuation in memory, attention, effort, fatigue, guessing, and similar factors (Gronlund, 1976; Popham, 1990). This is inevitable and must be accepted. What should be done is to construct tests in such a way that the scores actually obtained on a test on a particular occasion are likely to be very similar to those, which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more reliable the test is said to be.

The reliability is therefore expressed as a correlation coefficient between the two sets of scores. The ideal reliability coefficient is 1 (Hughes, 1997). Reliable tests will have a reliability coefficient tending towards 1 and unreliable tests will have a reliability coefficient tending towards zero. Certain authors have suggested how high a reliability coefficient should be for different types of language tests. For instance, Lado (1964) suggested that good vocabulary, structure and reading tests usually range between 0.90 to 0.99 while auditory comprehension tests are more often in the range of 0.80 to 0.89. Ebel and Frisbie (1991) observe that for a general TMT, the reliability coefficient should be at least 0.65 if the scores are the only information used to make

decisions. However, reliability of 0.5 can be accepted if each score will be combined with other information, for instance, quiz scores and observation scores. However, Davis (1991) and Lado cited in Davis (1988) observe that, TMTs are infamous for their lack of reliability and many have a reliability coefficient approaching zero. Probably most fall in range 0.20 – 0.40.

2.7.1 Estimating the Reliability Co-efficient of Teacher-Made Tests

There are various approaches of estimating reliability of a test. The most commonly used in educational measurement and evaluation are stability estimates, alternate/equivalent forms, and internal consistency.

Stability estimates of reliability are based on the consistency of tests measurement over time. The most common way of determining test stability is to administer the same test twice to the examinees. This would provide two scores for each occasion. If the correlation coefficient is high, it shows that the test is measuring the same ability even though it is administered at different times. Test-retest method is particularly used in situations where the trait being measured is expected to be stable over time (Ebel & Frisbie, 1991; Popham, 1990). In achievement tests the test-retest method has raised a number of objections.

The first objection is using exactly the same test items twice. Since this set of items represents only one sample from what is ordinarily a number of possible test items, the scores on the test provides no evidence on how much the scores might change if a different sample of questions were used. The second objection is that the examinees will generally score somewhat higher the second time because of recall, practice (individual or in group). Thirdly, if the interval between the tests is too long; students' ability may be influenced by maturation. Finally, re-administration of the same test twice to determine how reliable the score are does not appeal to most teachers and learners as a very efficient use of instructional time (Gregory, 1996; Popham; 1990).

Most learners will often score lower on the second testing merely because they view the activity as meaningless. Thus, the test re-test method is not recommended for

estimating the reliability of scores for classroom achievement tests. However, there is no objection of using them with psychological tests so long as the second score is strongly correlated with the first. This is because they will be measuring the stability of behaviour (Anastasi, 1982; Gregory, 1996).

The second approach of estimating test reliability is the use of alternate or equivalent forms. In most cases test developers produce two forms of the same test. The two forms of the same test are administered and then the examinee's scores on the two tests are correlated. The resulting coefficient is referred to as an alternate-form coefficient or equivalence reliability coefficient (Popham 1990; Thorndike, 1997). Gregory, (1996) observes that alternate forms are independently constructed to meet the same specification, often on item by item. Thus, alternative forms of Kiswahili test would incorporate in both tests similar content and cover the same range and level of difficulty in items.

A high reliability estimate shows that the two Kiswahili test forms can be used interchangeably as measure of the same traits. Low reliability is an indication that the two set of test items are not sampling the content equally well. Content sampling is perhaps the most prevalent category of error affecting achievement test scores. That is, some examinee may do better or worse on one form of a test because of particular items sampled. Consequently, estimation methods that are able to detect content sampling errors are most appropriate for achievement test scores. However, alternate forms of test needs a lot of time to construct and also overcome psychometric difficulties of producing truly parallel forms (Ebel & Fresbie, 1991). Although this technique is used extensively by standardized test makers, classroom teachers rarely prepare alternate forms of achievement tests.

The final approach is the internal consistency method. This approach requires only one single administration of a test and it focuses on the consistency of a test's internal elements. Internal consistency of estimates reveals the extent to which the items on the test are internally consistent with one another (Hopkins & Antes, 1985; Popham, 1990). There are three internal consistency techniques of calculating reliability estimates. These are split-half, Kuder-Richardson formulae and coefficient- alpha formulae

Split-half technique consists of dividing a test into two equal halves ordinarily by the odd and even items as though they constitute separate tests. The two sub-scores are then correlated (Anastasi, 1982; Ebel & Frisbie, 1991; Gregory, 1996). The resulting correlation coefficient is considered an estimate of the degree to which the two halves of the test are performing their functions consistently. Splitting a test into two means that the scores on which the reliability is based are from half-length tests. To obtain an estimate of reliability based on the full-length correlation. This is done by the help of Spearman-Brown prophecy formula which using the correlation between two half tests, estimates what the reliability would be on the full-length test. (Gronlund, 1982; Popham, 1990; Salvia & Ysseldyke, 1991). Spearman-Brown formula is given as:-

$$r_{SB} = \frac{2r^{hh}}{1+r^{hh}}$$

Where

r_{SB} = Estimated reliability of the full test

r^{hh} = Reliability of half test

A widely used method of estimating reliability for binary-scored test items is the Kuder-Richardson formula particularly formulae 20(K-R20) and 21 (K-R21). The K-R21 formula is somewhat less accurate than the KR-20 formula but it is also so simple to compute that it is most frequently employed estimate of internal consistency by classroom teachers (Gronlund 1982; Popham, 1990). The KR-20 requires more computation. For instance, it requires information about difficulty of each item. KR-21 formula is given as:-

$$r = \frac{k}{k-1} \left[1 - \frac{\chi(k - \bar{\chi})}{ks^2} \right]$$

Where

k = Number of items

$\bar{\chi}$ = Mean of test scores

s = Standard deviation of test scores

This formula requires just three types of information.

- The number of items
- The mean
- The standard deviation

Coefficient alpha developed by Cronbach (1951) and subsequently elaborated by other scholars like Nerick and Lewis, (1967) and Kaiser and Michael, (1975) can provide a reliability estimate for items scored with values other than binary scored responses (Ebel & Frisbie, 1991, Gregory, 1996, Popham, 1990; Salvia & Ysseldyke, 1991). These other test items include short-answer, essay questions and attitude scales that provide responses such as 'strongly agree to strongly disagree'.

The formula for coefficient alpha is

$$\alpha = \frac{K}{K-1} \left(1 - \frac{\sum S_i^2}{S_x^2} \right)$$

Where

K = Number of items on the test

S_x^2 = Variance of the total test

S_i^2 = The sum of the variance of individual items

2.7.2 Increasing Test Reliability by Lengthening the Test.

If a teacher discovers that the coefficient of reliability is not too low, he can be saved from the problem of constructing another test by lengthening it. This can be done by applying the following formula adopted from Hopkins & Antes (1985) and Marshall and Hales (1977).

$$\text{Length desired} = \frac{(\text{Reliability desired}) \times (1 - \text{Reliability obtained})}{(\text{Reliability obtained}) \times (1 - \text{Reliability desired})}$$

The dependence of the test upon the number of test items assumes that the additional test items are drawn from the hypothetical item pool (have same difficulty level). The desired length of the test is interpreted as the number of similar tests needed to reach the desired level of reliability (Ebel & Frisbie, 1991; Marshall & Hales, 1977).

2.7.3 Scorer Reliability

Scorer reliability refers to the degree of agreement or consistency that exists between two or more scorers (Cohen et al, 1992). There is as much need for a measure of scorer reliability as there is for the more usual reliability coefficient as is a prerequisite for high-test reliability. Scorer reliability is particularly important for supply type test- items since consistent scoring of such responses is difficult to accomplish (Anastasi, 1982; Gregory, 1996; Popham, 1990). Nearly all Kiswahili test items used in TMTs in Kenyan secondary schools are of this type. Thus, Kiswahili teachers are expected to establish scorer reliability of their tests.

Scorer reliability can be established by correlating the values of two independent judgments on a common set of test responses (Hopkins & Antes, 1985; Hughes, 1997). The correlation coefficient can be determined by applying either Pearson Product-moment correlation coefficient formula or the simpler Spearman Rank-difference correlation coefficient formula (Lyman, 1991). The coefficient obtained reflects the degree of agreement between the readings and it is referred to as the scorer reliability coefficient (Anastasi, 1982; Hughes, 1997).

It is possible to make the reliability of subjective tests quite high by having a detailed scoring key that is faithfully followed and takes into consideration all the possible answers (Cangelosi, 1990; Wood, 1995). In a normal classroom situation, it is difficult to have the same set of papers scored by two separate teachers. It is more practical to have the regular subject teacher score the paper twice (Hopkins & Antes, 1985). When scorer reliability is to be calculated by correlating the scores assigned to a set of papers by the same teacher at two different scoring sessions, some precautions should be taken to reduce as much as possible the influence of the teachers memory and other extraneous factors on the scores assigned. A simple procedure propounded by Hopkins & Antes (1985) and Marshall & Hales (1977) can effectively be used. It is possible for instance to obtain a scorer reliability of over 0.9 for scoring of compositions (Hughes, 1997). It can therefore be argued that, if the scoring of the test is not reliable, then the test results cannot be reliable either (Dreckmeyr & Fracer 1991; Popham, 1990).

2. 8 Factors Affecting Test Reliability and Test Validity

The reliability of achievement test is affected by a number of factors. First is the length of the test. Mathematically derived estimates to test reliability confirm the fact that longer test tend to give more reliable results than shorter tests. This is derived from the premise that there is an increased degree of confidence associated with large samples as compared to small ones (Salvia & Ysseldyke, 1991). The Spearman-Brown formulae shown below indicates the theoretical relationship between score reliability and test length.

$$r_n = \frac{n(r)}{(n-1)(r) + 1}$$

Where

r_n = reliability of the scores from the new test

n = the number of times original test is lengthened

r = is the reliability of the original test scores

Table 3 shows the effect of consecutive doubling to the number of items beginning with a 5-item test with a reliability estimate of 0.20.

Table 3:

Relationship between Test length and Reliability

No. of Items	Reliability
5	0.20
10	0.33
20	0.5
40	0.67

Source:Adapted From Salvia & Ysseldyke (1991)

The psychological assumption involved in the Spearman Brown formula is that examinee responses to the test will not change. That is, factors like fatigue and boredom will not affect the motivation to do well (Ebel & Frisbie, 1991; Popham, 1990).

The second factor that affects reliability is the test content. Homogenous test content is expected to reveal higher reliability coefficients than tests in areas where subject matter is diverse and loosely organized. A 20-item test about the use of objectives is likely to provide more reliable scores than 20-item test of general Kiswahili grammar. According to Gregory (1996) and Hughes (1997) the high reliability for homogenous content results from the interdependence of principles and facts making for consistency of the responses.

Thirdly, item characteristics can also affect test reliability. A Kiswahili test that is made up of tasks that are too difficult for students may cause them to guess at responses and will result in inconsistent measures. This is because achievement test items often presume that the students taking the test have had exposure to concepts and skills measured by the test (Salvia & Ysseldyke, 1991). On the other hand tests that are too easy do not discriminate and thus do not allow for meaningful estimates of test reliability. The ability to discriminate depends heavily on the technical quality of the test items. Determination of discrimination index and its relation to reliability is discussed under Item Analysis. When a teacher works to improve the discrimination of the individual items in classroom tests it becomes the most effective means of improving reliability and hence, test quality (Ebel & Frisbie, 1991)

The difficulty of a test item affects reliability in two ways. First, an item that is answered correctly by all learners or missed by all contributes nothing to reliability. Secondly, an item with intermediate difficulty is potentially capable of contributing more to increased reliability. Therefore, a good norm-referenced achievement test should not include items that vary widely in difficulty (Popham, 1990)

The fourth factor that affects reliability of test scores is the accuracy in scoring. Any errors made during scoring a test will affect the reliability of that test. Lack of scoring accuracy especially for essay tests has been a major contributing factor to low

reliability (Hopkins & Antes, 1985). If the scorer cannot be consistent in scoring, test reliability will be drastically reduced. As mentioned earlier objectivity in scoring by using a detailed scoring key, following essay test scoring procedures and estimating scorer reliability can help overcome major difficulties associated with scoring essay tests.

Fifthly, group homogeneity in the classroom can affect reliability of a test. There are circumstances in which the students in a class are very similar to one another in their achievement of the objectives in an instructional unit. The standard deviation on the unit test becomes very small and the reliability coefficient becomes very low also. This may be a situation in which a high-quality test, when given properly, cannot yield scores of very high reliability (Popham, 1990; Salvia & Ysseldyke, 1991). Though group homogeneity may be a plausible explanation for low reliability at times, a teacher should not be quick to ignore the signs of faulty test items before settling on group homogeneity as the most likely reason.

Sixthly, students' motivation at the time of taking the test can greatly affect their scores. Indifference, lack of motivation, or under enthusiasm for whatever reasons can depress test scores in the same way that anxiety or over enthusiasm may (Ebel & Frisbie, 1991; Gregory, 1997). When motivation influences individuals in a group differently and inconsistently across testing occasions, random errors are likely to influence the scores.

Classroom conditions, teacher rapport, and home conditions contribute to the motivational level of students. Students who suffer from low motivation may have little drive and may answer items at random or with little attention to the task which has been presented on the test (Hopkins & Antes, 1985). Although test anxiety is not considered to be a major problem in testing, lack of motivation must be looked upon as a serious threat to reported reliability with certain populations of students and with individual students in all populations. Thus, test consistency (reliability) is a function of consistency in students' behaviour (Gregory, 1997).

The seventh factor that affects test reliability is time limits. A test which measures how fast someone can react to test tasks (speeded test) is expected to have higher

reliability coefficient than one which all or nearly all students can complete in the time provided for the test (power test). In the class, a speeded test has a restrictive time limit, which guarantees that few subjects complete the entire test (Hopkins & Antes, 1985; Gregory, 1997). Scores of a speeded test depend not only on how many items examinees can answer, but also on how fast they answer them. Thus, to estimate the reliability of scores on a speeded test, one must have estimates for both ability and speed. To control for this, students should comfortably complete all items on a classroom test. In most cases the concern of objectives is not how quickly students can perform a task but rather whether they can complete the task to satisfaction or not. However there are special skill tests, which involve speed as desirable factor, such tests would include typing, short hand, reading and others.

Lastly, cheating opportunities can lead to inflated reliability coefficient. Occurrence of cheating by students during a test contributes random errors to the test scores. Some learners provide correct answer for questions to which they actually do not know the answers. Copying of answers, getting access to copies of the test prior to its use, use of cheat sheets, and the passing of information amongst the learners all give unfair advantage to some and cause their scores to be higher than they would, thus, leading to inaccurate and less meaningful scores (Ebel & Frisbie, 1991). Nearly all examinee and administration related factors can be controlled by the teacher during testing. However, test-related factors are mostly controlled during the test construction process.

Whenever a test fails to measure what it purports to measure, validity is threatened. Thus, any factor that results in measuring 'something else' affects tests validity. Reliability and lack of proficiency on part of the test maker threaten validity. Reliability is necessary but not a sufficient condition for valid measurement (Davies, 1991; Salvia & Ysseldyke, 1991; Lyman, 1991). All valid tests are reliable, but reliable tests may or may not be valid. Such a test may only be 'reproducing' what it was not intended to measure. Therefore, all factors that affect reliability affect validity. In addition, lack of proficiency in constructing quality tests can adversely affect the test validity. Some of the poor testing practices amongst the teachers which threaten test validity are; -

- Unclear instructions that do not state clearly how learners should respond to the test items.
- Too complicated sentences structure and vocabulary. It becomes difficult for the learners to grasp the question. Ultimately, it looks as if the classroom test is testing vocabulary and not the content.
- Ambiguous test items. Ambiguity leads to misinterpretation of questions especially by high achievers. In most cases ambiguity favours low achievers, thus causing negative discrimination (Ebel & Frisbie, 1991; woods, 1995).
- Improper arrangements of test items. It is advisable to arrange test items from the easiest to the most difficult starting with difficult items has detrimental effects on motivation and may lead to learners spending too much time on one question,
- Identifiable pattern of answers especially in multiple – choice and true false items.

2.9 Item Analysis

Item analysis refers to the procedures of evaluating the effectiveness of the items in a test. The purpose of the item analysis is to identify those items that need to be improved or replaced (Frisbie & Ebel, 1991; Gronlund, 1982; Hughes, 1997). Popham (1990) and Marshall and Hales (1971), refer to item analysis as empirical assessment of item quality. Both validity and reliability of any test depends ultimately on the characteristics of its items. Thus, high validity and reliability can be built into a test in advance through item analysis (Anastasi, 1982; Davies, 1991). Item analysis is practically relevant to the construction of Kiswahili TMTs. Teachers can improve their item writing skills and eventually accumulate high quality test items in an item pool (Anastasi, 1982; Frisbie & Ebel, 1991). Nearly all proponents of item analysis agree that examining item difficulty (facility) and item discriminating power establishes the quality of individual test-items.

2.9.1 Item difficulty

For most testing purposes, the difficulty of an item in a classroom test is defined in terms of percentage or proportion of students who answered the item correctly. This is

commonly referred to as p value (Frisbie & Ebel, 1991; Gronlund, 1982; Hughes, 1997; Popham, 1990). The p value is basically defined in terms of the relative frequency with which those taking the test choose the correct response. In addition, the item difficulty is a characteristic of both the item and the population taking the test (Gregory, 1996; Murphy & Davidshofer, 1988).

The test p value can range from 0 to 1.00. Higher p value indicates that more examinees answered the item correctly. For example, in a Kiswahili test an item with a p value of 0.9 would be one that was answered correctly by almost all examinees. Similarly, an item with a p value of 0.1 would be one that most examinees failed. Perhaps the most useful implication of the p value is that it provides a common measure of the difficulty of the test items. The basic assumption of measurement is that there are systematic individual differences in the construct or defined universes of content being measured. Tests represent a method of quantifying these differences (Marshall & Hales, 1971; Murphy & Davidshofer, 1988). Thus, when nobody chooses the correct answer (a p value of 0) or when everybody chooses the correct answer (a p value of 1), there are no individual differences in the score of that item.

Test items with a p value of 0 or 1.00 may affect the test mean but have no effect whatsoever on the test reliability or validity, nor on the decisions that are based upon the test scores. A test designed to obtain maximum differentiation amongst the examinee should be of 0.5 difficulty index. Since most achievement tests attempt to measure a variety of facts and understanding, it is generally best to select items of approximately 0.5 difficulty (Ebel & Frisbie, 1991; Hughes, 1997). However, there are some circumstances that may dictate how difficult a test should be. For instance, if a test is to be used for selecting students who should enrol for science subjects at form three, the science test should contain many very difficult items. On the other hand, a test used to designate children for a remedial instructional programme should contain many easy items.

2.9.2 Item Discrimination

Item discrimination power is the extent to which an item separates those who really know the answer from those who do not know (Frisbie & Ebel, 1991; Wood, 1995). In principle, an item discrimination index reflects the relationship between

examinees' responses on the total test and their responses on a particular test item (Popham, 1990). This is commonly known as D value.

In norm-referenced tests, the ability of a test to highlight differences in achievement on the attribute being measured depends on the discrimination power of the test items. The higher the D value the better the item discrimination power (Gregory, 1996; Hopkins & Antes, 1985). Item discrimination index can vary from -1.0 to 1.0. A positively discriminating index indicates that the test item is answered correctly more often by those who score well on the total test than by those who score poorly on the total test. A negatively discriminating index indicates that an item is answered correctly more often by those who score poorly on the total test than by those who score well. Negative D value is a warning signal that the test item needs replacement or revision.

An item with a D value of zero provides no discrimination since in both the low-score group and high-score group the same number of students got the item correct. Therefore, an item that is not discriminating between the low and the high achievers should be revised or eliminated (Ebel & Fribie, 1991; Murphy & Davidshofer, 1988). Low D value should be investigated to see if they are resulting from very easy or very difficulty items. Since they are doing very little to place students in relative standing on the trait being tested, consideration must be given to changing the difficulty or discarding the item. However, Popham (1990), Hopkins and Antes (1985) and Ebel (1972) have provided the following experience-based guidelines for indicating the quality of norm-referenced test items on Table 4.

Table 4:

Levels of item discrimination

Discrimination Index	Item evaluation
.40 and above	Very good items
.30-.39	Reasonably good but possibly subject to improvement
.20-.29	Marginal items usually needing and being subjected to improvement
.19 and below	Poor items, to be rejected or improved by revision

Source: Popham (1990)

Item analysis of a Kiswahili test can therefore indicate which items may be too easy or too difficult and which may fail for whatever reason to discriminate properly between high and low achievers. Item analysis data also helps the teachers detect specific flaws and thus provide further information for improving test items (Gregory, 1996; Gronlund, 1982; Murphy, 1996). Popham (1990) adds that most often item analysis only identifies problems and the teacher searches for the problem causes and possible solutions.

In a classroom situation, item analysis provides useful information for class discussion of the test. For instance, answers to difficult items can be pointed out to the students rather than being defended as fair. Moreover, item analysis provides data that helps students improve their learning. According to Gronlund (1982), the frequency with which each incorrect answer is chosen reveals common errors and misconceptions, which provide a focus for discussion. In addition, item analysis provides insights and skills that lead to the preparation of better test items in the future. The process helps teachers become more aware of defective items and how to correct them (Popham, 1990).

2.10 Theoretical Framework

The theoretical framework of the study was based on General Systems Theory and the Instructional theory particularly Blooms Model of Mastery Learning. General

Systems Theory developed by Ludwing Von Bertalanffy posits that a system is a whole, which consists of several interrelated subsystems characteristically independent of each other (cited in Simiyu, 2001). On the other hand, Instructional theory postulates that courses should be arranged in instructional units, use of regular materials and methods to teach the class, testing the students to identify who have achieved mastery of each unit, advancing those who have achieved mastery, giving remedial instruction to those who have not and starting all students in the next unit of instruction at the same time (Klausmeier, 1985; Rensnick & Klopfer, 1989; Poid & Haladyna, 1985)

It is from these two theories that the proponents of systematic instruction have coined their approach. Any form of systematic instruction has three common elements (a) statements describing the intent of instruction (objectives), (b) instruction that is designed to help the students achieve the intended outcomes of instruction (content) and (c) testing that is explicitly related to both intent of instruction and instruction itself (Poid & Haladyna, 1982). The three elements borrowed from the instructional theory as the major components of instruction are viewed as unique enterprises, but there is substantial interdependence among them, hence forming a system. This information can be represented by three concentric circles as shown in Figure 2.

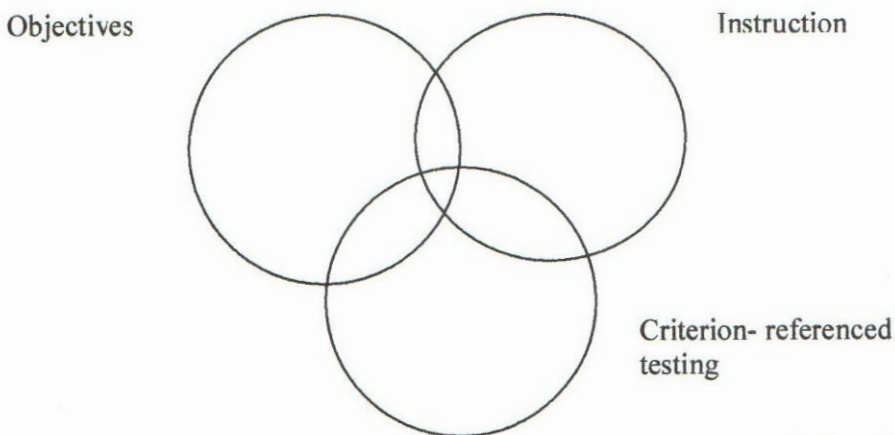


Figure 2: Aspects of systematic instruction (Adapted from Poid & Haladyna, 1982)

Based on this approach, the teacher is the one who develops the instructional objectives, comes up with instructional strategy and tests the learners on the mastery

of the content. It is on the basis of the tests, which are built from items that logically reflect the intent of instruction that teachers make decisions regarding which areas of the lesson has not been mastered (Poid & Haladyna, 1982). One of the most important tasks for a classroom teacher is to ensure that learners achieve instructional objectives. Teachers must monitor the progress of both the class and the individual students in order to make good decisions about where to begin teaching, when to move to the next unit of instructional content, whether to re-teach the present unit or whether a particular student or subgroup of students need special help to master the learning task (Thorndike, 1997). Besides, feedback from classroom tests promotes an opportunity for the teacher to modify instructional methods or materials to facilitate learning.

In respect to this study, the quality of these decisions will depend on the quality of the data gathered by the Kiswahili tests. Kiswahili teachers need to construct tests with high content validity through building table of specification, basing tests on instructional objectives and using Blooms taxonomy to classify their instructional objectives. Teachers also need to estimate the reliability of their tests and determine the quality of the test items through item analysis. Such Kiswahili tests will always yield quality data. Therefore, a quality test constructed by observing psychometric procedures will definitely give accurate and reliable information, which will help the teacher make sound decisions about the instructional process. Thorndike (1997) asserts that the study of educational assessment procedures should yield an understanding of the tools and techniques that will provide information that is more accurate and also provide the basis for judging the degree of confidence that can be placed on decisions made from such information. Thus, if efforts are made to improve the quality of Kiswahili TMTs, the quality of the instructional content and the effectiveness of instructional objectives will also improve. In support of this view, Grounlund (1982) contends that, in order to realise the full potential of TMTs as learning aids it is necessary to make testing an integral part of the instructional process.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes and justifies the research design used for this study. It also describes the population, sample and sampling procedures, instrumentation and administration of the research instruments, data collection and data analysis procedures

3.2 The Research Design

The research design for the study was descriptive survey. Descriptive survey design was adopted to obtain pertinent information concerning the quality of TMTs and factors that inhibit their accurate development. A survey uses questionnaires and, or interview schedules to collect data from participants in a sample, about their characteristics, experiences and opinions, in order to generalize the findings to a population (Best & Khan, 1992; Gall & Borg, 1996; Fraenkel & Wallen, 2000; Mugenda & Mugenda, 1999). Consequently, surveys are conducted to collect detailed description of an existing phenomenon with the intent of employing the data to explain the current conditions and practices (Gall & Borg, 1996; Kathuri & Pals, 1993; Ndagi, 1984). These characteristics fitted very well with the nature and purpose of the research design used in the study.

3.3 Population

The target population for this study was the Kiswahili teachers of Bahati Division. There were 86 teachers of Kiswahili in 39 secondary schools in Bahati Division of Nakuru district. Out of this population only 76 were trained (D.E.O Nakuru, 2003). All the trained teachers participated in the study as respondents. For each teacher, one of his/her end of term examinations was assessed. In essence, 76 TMTs were studied

irrespective of the class that the teacher was teaching. This is because test construction procedures are the same for all levels of classes in secondary school. Bahati division was chosen because there were as many public schools as there were private. The Division is also comprised of all categories of schools namely: District, Provincial and a National school. As such, a representative sample of teachers experiences from these different categories of schools was captured.

3.4 Sample and Sampling Procedures

This study involved all the teachers of Kiswahili in Bahati division. Taking the whole population was one way of enhancing representativeness with any error being attributed only to the measuring instruments. Whatever conclusions were drawn were very sound since they were describing and interpreting the parameters of that group directly rather than trying to describe the population characteristics based upon statistics from a sample (Bhaskar, 2002). In addition, all 20 heads of departments (with a Kiswahili bias) were selected for an interview. The 20 heads of department were purposely selected among the 39 heads of department in the 39 secondary schools.

3.5 Instrumentation

The researcher developed a questionnaire, an interview schedule, and a checklist intended to elicit information from teachers of Kiswahili on the current status of the quality of their tests and what may be hindering them from constructing accurate TMTs.

3.5.1 The Questionnaire

The questionnaire was given to all the Kiswahili teachers selected for the study. The items in the questionnaire were based on what the experts emphasize as the main components of a quality test. The question items consisted of multiple choice, Yes/No items followed by Open-ended questions. Open-ended questions are particularly important in this study because they will give the respondents an opportunity to respond freely in giving their answers (Mugenda & Mugenda, 1999; Ndagi, 1984).

3.5.2 Checklist

The checklist was applied to all the teachers on whom the questionnaire was administered. A checklist was used to gauge the respondents' frankness and application of test construction techniques. The checklist consisted of list of items whose physical presence was to be verified as highlighted in the teachers' questionnaires. These include; list of instructional objectives for the content to be tested, a table of specification and item banks. It was also used to collect test scores and specific information from Kiswahili tests which will enable the researcher to conduct item analysis, calculate the average reliability coefficient and percentage of items appearing at each level of cognitive domain. Checklist items are a matter of 'fact' not of 'judgement'. They are important tools in gathering facts for educational surveys, instructional procedures game facilities etc. (Koul, 1993)

3.5.3 Interview Schedule

An interview was conducted with 20 heads of departments. The purpose of interviews is to supplement data collected through other methods and thus allow the researcher to follow up a respondent's answer to obtain more information and clarify vague statements (Cohen & Manion, 1997 Gall & Borg, 1996). Specifically, confirmation interview was conducted. The researcher used the interview schedule to gather information on whether psychometric procedures were followed by Kiswahili teachers as they constructed their tests. Moreover, this data collection tool consisted of items that helped the researcher to probe into the factors that were perceived to inhibit accurate test construction by teachers. The use of the three instruments was necessary in order to investigate the research problem thoroughly. The questionnaire is shown in appendix A, the interview schedule appears in appendix B, and the checklist in appendix C.

3.6 Reliability and Validation of Research Instrument

Validation of the instrument involved piloting the questionnaire on 15 teachers of Kiswahili in secondary schools outside the study area. This was done in Njoro Division to ensure that the targeted population had no prior knowledge of the study. This also avoided contamination. The draft instruments were discussed with the research supervisors and any ambiguity removed. This enhanced content validity by ensuring that the instruments captured all requirements of a quality test instrument. This is what Borg and Gall (1996) refers to as the use of experts to enhance validity.

A measure of internal consistency was used to establish whether the items used to measure test quality and factors inhibiting quality test construction will be reliable or not. A reliable instrument is one that will provide similar results if used with same respondents on the different occasions (Cohen & Manion, 1997). The Cronbach alpha formula was used to calculate reliability coefficient. This formula was best suited for the study since the instruments contained both select and supply type items. The questionnaire should meet a reliability coefficient of 0.7 and above to be accepted. This is the acceptable level for survey research (Best & Khan, 1992; Mugenda & Mugenda, 1999). Some adjustments were done to some questions until the criterion was met.

3.7 Data Collection Procedures

The permit to conduct the research was sought from the Ministry of Education, Science and Technology. The 76 teachers were given the questionnaire by the researcher to solicit their responses. The researcher contacted all the 20 heads of departments for an interview. The researcher picked the questionnaires in person to ensure that they were duly filled. This was done a week after they were submitted. The checklist was later applied to ascertain the teachers' honesty and availability of records that indicate whether teachers follow procedures that promote quality tests.

3.8 Data Analysis

The completed questionnaires were evaluated for errors before subjecting them to analysis. The Statistical Package for Social Sciences (SPSS 11.5) was used to analyse the data collected. SPSS is the most commonly used set of computer program in educational research. It is comprehensive, integrated collection of computer programs for managing, analysing and displaying data (Gall & Borg, 1996). Tables, means, frequencies and percentages were used to summarize the data as shown on table 5. In essence, qualitative method of data analysis was used to analyse the data.

Table 5:

Variables and their Analysis

Research questions	Independent variables	Dependent Variable	Statistics used
Do teachers establish the validity of their tests?	Instructional objectives	Construction of Quality test	Percentages and frequency tables
	Table of specification	Construction of Quality test	Percentages and frequency tables
	Blooms Taxonomy	Construction of Quality test	Percentages and mean
	Subject Specialists	Construction of Quality test	Percentages and Graph
	Item bank	Construction of Quality test	Percentages and Frequency tables
Do teachers establish the reliability of their tests?	Reliability Coefficient	Construction of Quality test	Mean
	Scorer Reliability	Construction of Quality test	Percentages and Frequency tables
Do teachers establish the quality of their test items?	Item Difficulty	Construction of Quality test	Percentages and Graph
	Item discriminating power	Construction of Quality test	Percentages and Frequency tables
What may hinder teachers from constructing quality tests?	Constraints in Constructing accurate test	Construction of Quality test	Percentages and Frequency tables

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents and discusses the results on the quality of Kiswahili Language Teacher-made Tests by teachers in Bahati Division of Nakuru District. The questionnaire was the basic tool for data collection. A checklist was used to analyse all the tests constructed by teachers in the sample schools. The checklist was used to gather information about the availability of test records. It also contained a score sheet, which was used to collect learners scores needed to compute item analysis. In addition, interviews conducted with 20 heads of language department (With a Kiswahili bias), helped in collaborating and elaborating the issues that teachers could not communicate effectively through the questionnaires. Data was analysed using the Statistical Package for Social Science (SPSS) version 11.5 to generate tables of descriptive statistics. The results and discussions for the study are presented under the following four subheadings abstracted from the research questions:

- Validity of Teacher-made tests
- Reliability of Teacher-made tests
- Item Analysis of Teacher- made test
- Factors hindering teachers from constructing quality tests

4.2 Validity of Kiswahili Teacher- made Tests

A four point criteria was used to establish whether TMTs were valid or not. These were whether teachers were: -

- Basing tests on instructional objectives
- Using Blooms Taxonomy of Educational Objectives
- Using table of specification
- Using other subject specialists
- Using Item banks

4.2.1 Basing Tests on Instructional Objectives

Teachers are expected to base their tests on instructional objectives. This is because tests based on instructional objectives will measure how well the instructional objectives were achieved. The results from the study on Table 6 indicate the different sources used by the teachers to develop their test items.

Table 6:

Sources of Kiswahili test items

Test source	Responses (%)		Total
	Yes	No	
Instructional Objectives	19.7	80.3	100
Past Papers	36.8	63.2	100
Lesson notes	65.8	34.2	100

Source: Field Data

The Table shows that teachers mainly used three sources for their test items. These were identified as instructional objectives, past papers and lesson notes. However, it is evident that most teachers (80.3%) did not use instructional objectives as a guide in test construction. The highest number (65.8 %) used their lesson notes while another 36.8 % used the past papers. Only a small percent (19.7) indicated that they were guided by instructional objectives. The findings are consistent with observations made by Airasian (1991) that the success of instruction is undermined by the construction, selection and use of test items, which are not related to what was taught.

TMTs are meant to measure how well the instructional objectives have been achieved by the learners. Tests which are based on objectives allow the teacher to monitor a learner's progress. If the learner has difficulties in an area of the curriculum, the objectives in the area can be broken down into smaller steps and the learner can receive additional teaching in this area. Thus, when instructional objectives become the target of assessment, and test items are generated to match the objectives, the probability that items produced by two or more teachers will correspond is high, unlike when varied sources of content are used (Wood, 1995).

By using the past papers as a source of test items, it means that the quality of the items has not been established. If items have not been analyzed for discrimination and difficulty, such items have no value as aids to future tests (Mulder, 1993; Stanley & Hopkins, 1972). Sometimes it is inevitable that copies of old tests might find their way into the students' files. Students may also know where a teacher will likely lift questions from. In such incidences, there is a danger of students only investing in just passing but not the mastery of the intended content. (Idol & Jones, 1991). Anastasi (1991) and Hopkins and Antes (1985) argue that, if the test is not composed of tasks that measure the degree to which the student has achieved the objectives, the test can be said to have no content validity. Consequently, tests constructed by teachers are not likely to be precise reflections of the instructional content teachers have taught.

4.2.2 Using Bloom's Taxonomy of Educational Objectives

Teachers are expected to use Bloom's taxonomy of Educational Objectives as a guide to test construction. It enables them to ensure that all levels of cognitive domain are tested. Table 7 shows the results of teachers' responses about their use of Bloom's Taxonomy of Educational objectives to guide their test construction.

Table 7:

Teachers' responses on the use of Bloom Taxonomy of Educational Objectives

Response	N	%
Yes	1	1.3
No	75	98.7

Source: Field Data

Results on Table 7 above indicate that nearly all the teachers (98.7 %) did not use Blooms Taxonomy to guide them in test construction. Only one (1.3 %) teacher out of the 76 attempted to use the Bloom Taxonomy. These are regrettable findings because teachers seem to neglect the use of taxonomy of educational objectives as a useful guide for test construction. The taxonomy of educational objectives serves as a convenient checklist to certify that the learning outcomes emphasised during the teaching receive similar emphasis in testing, and that the learning outcomes are stated

in terms of specific students performance (Dreckmeyr & Fracer,1991; Marshal & Hales, 1977).

Gronlund (1982) argues that the cognitive domain of the taxonomy is especially useful in planning the achievement tests. It focuses on a comprehensive and complete list of mental processes to be considered when identifying learning outcomes and it provides a standard vocabulary for describing and classifying learning outcomes. When teachers fail to use the taxonomy to classify objectives, they may come to realize that most of their objectives require only simple remembering or recall of information, while they actually intended students to understand and apply knowledge (Ebel & Frisbie, 1991). A further analysis of the tests constructed by the teachers showed that most teachers did not commit themselves to ensuring that their tests measured all the levels of the cognitive domain. The percentage of items represented at each level is shown in figure 3.

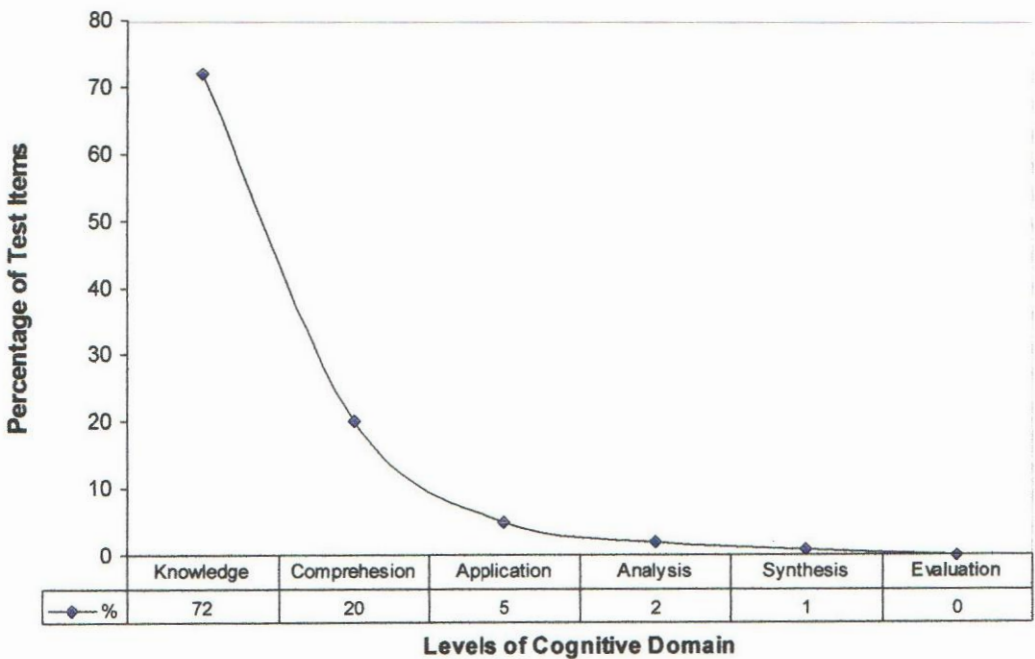


Figure 3: Percentage of items in Blooms Taxonomy of cognitive domain in Kiswahili tests
 (Source: Field Data)

The figure above shows that 97 % of all the test items tested the lower level cognitive skills of knowledge, comprehension, and application. Only 3 % tested the higher level

cognitive skills of analysis and synthesis. There was no single item testing evaluation. In some cases (Appendix D & E) all the items in a test were found to assess only the lower level cognitive skills. The results indicate that questions that asked students to recall information were much more common than those, which required comparisons, inferences or evaluation. This agrees with Gandiye (1991) who observed that school based assessment in Tanzania when compared to external examinations, showed test items which focused less on higher level abilities such as analysis, synthesis and evaluation. Similarly, Fleming and Chambers (1983) analysed more than 8,800 test items constructed by teachers in U.S.A and found out that nearly 80 % of them dealt only with knowledge of facts and specifics at the lower level in the Bloom Taxonomy (in Lefrancouis, 1991). In addition, an analysis of teacher-made tests by Idol and Jones (1991) indicated that questions that required the use of higher order thinking skills were more of an exception rather than the rule. Blooms Taxonomy of educational objectives is the best framework through which teachers can be sure that all the six levels of cognitive domain are tested and that teaching and testing is not only for memorization and recall (Ebel & Frisbie, 1991; Gronlund, 1982). In essence, Bloom taxonomy helps a teacher to select test items from all levels of educational objectives. When tests measure a variety of educational objectives, the teacher will be emphasising to the learner that the learner must devote his attention to all these areas. Eventually, the learner will learn that a mass memorization of factual information is not sufficient in the learning process (Hopkins & Antes, 1985; Smith, 1982).

4.2.3 Using the Table of Specification

A further investigation was carried out to establish whether teachers used the Table of Specification in constructing their tests. Table 8 below gives the teachers' responses.

Table 8:

Teachers' responses on the use of table of specification

Response	N	%
Yes	0	0
No	76	100

Source: Field Data

From the results above (Table 8), it is evident that all the teachers (100 %) did not use the Table of Specification in constructing their tests. The Table of Specification guides the work of test construction. It shows what the test items will be measuring, the level of cognitive skills tested and the relative weight to be given to each level depending on the emphasis given during instruction. It also provides an opportunity to other teachers to determine whether or not the items classified against each cognitive skill truly belong where they are placed. The absence of the table of specification among all the teachers perhaps suggests why their tests concentrated more on testing the lower level cognitive skills. Interview data collected from HODs yielded similar results as shown in Table 9.

Table 9:

H. O. D Responses on test construction procedures

Items	Responses	
	Yes	No
Do teachers in the department establish test validity through:		
• A table of specification?	0	20
• Blooms taxonomy of educational objectives?	0	20
• Items bank?	0	20
Do teachers in the department estimate test reliability by establishing:		
• Reliability coefficient?	0	20
• Scorer reliability?	0	20
Do teachers in the department conduct item analysis for their tests	0	20

Source: Field Data

These results agree with Anderson (1989) who analysed 120 TMTs and found out that in 90 % of the cases, no empirical test development procedure i.e. item analysis, table of specification or statement of objectives was given. Ebel and Frisbie (1991) and Popham (1990) argue that a table of specification is the test's blue print. It simultaneously lists the test content and level of cognitive skills required to be tested to enable the test to be consistent with the instructional objectives. Thus, the table specifies the topics to be tested, the nature of the questions to be used, how many

questions will relate to each topic and the sort of cognitive process to be sampled (Lefrancois, 1991). In addition, a careful use of the table of specification can help a teacher ensure the validity of his tests, both by content and by the level of cognitive skills tested (Salvia & Ysseldyke, 1991). Moreover, poor quality tests tend to direct learners to limited aspects of course content and on the part of a teacher, such tests lead to rewarding superficial learning (Hughes, 1997; Wood, 1995)

4.2.4 Use of Subject Specialists

A second judgement from a subject colleague is important in order to build content validity of a classroom test. However, figure 4 below shows that only 44.7 % of the teachers consulted their colleagues. The greater proportion that is, 55.3% did not.

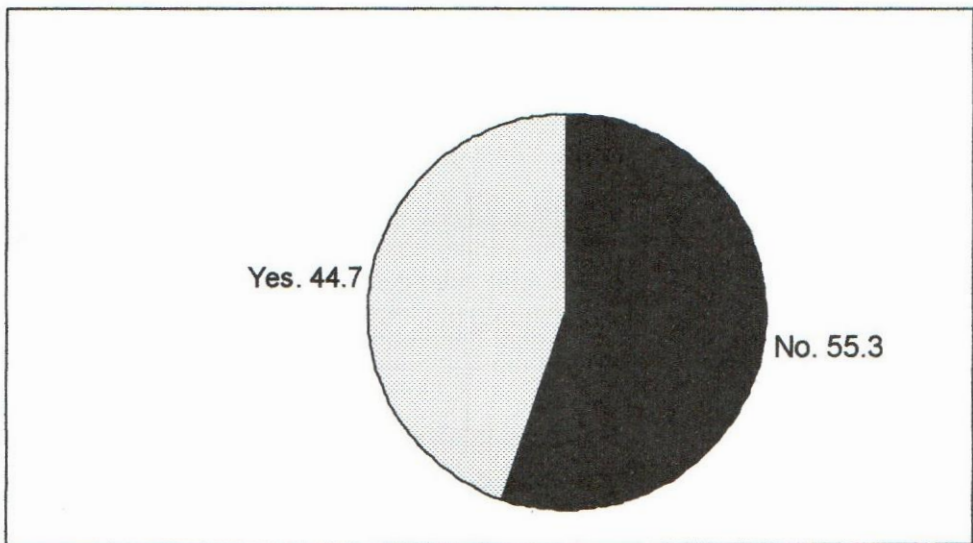


Figure 4: Use of other Subject Teachers in Test Construction (Source: Field Data)

One of the strategies followed in gathering content-related evidence of validity is to subject a test to a series of post-facto judgements. To use this approach, the classroom teacher asks another teacher in the subject area to review the appropriateness of the test's content (Popham, 1990; Davis, 1991). Airasian (1991) argues that the success of measuring the extent to which the instructional objectives have been achieved is undermined by the construction, selection and use of tests which are either not related to what was taught or which trivialize the breath and depth of the concepts and behaviour taught. Such inadequacies could be reduced greatly if a teacher would have

a competent colleague review the test and suggest corrections or improvements. Further, Hughes (1997) argues that if an individual teacher's judgement is always relied upon, too often the content of the test will be determined by what is easy rather than what is important to test.

4.2.5 Using Item Bank

The results on the responses whether teachers built Item banks during their testing procedure are shown in Table 10. It appears from these results that 100 % of the teachers indicated that they did not build item banks for their tests.

Table 10:

Kiswahili teachers' use of item bank

Response	Frequency	%
Yes	0	0
No	76	100

Source: Field Data

The results tabulated from the interview with heads of departments (Table 9 p55) yielded similar results. It therefore means that teachers do not control the quality of their tests. The revelation that most of them preferred to use past papers and lesson notes as the sources of their tests, implies that teachers could have been perpetuating the use of test items whose quality had not been determined.

An item bank makes it possible to select items in accordance with the table of specification. Since item analysis information is included for each item, it is the only sure way that the item selected for use is of high quality (Airasian, 1991; Lyman, 1991). A test bank assumes increasing importance as we shift from test items which measure knowledge of specific facts, to those which measure understanding, application and thinking skills. Items in these latter areas are difficult and time consuming to construct. With all the demands on a teacher's time, it is almost impossible to construct effective test items in these areas each time a teacher prepares a new test. (Airasian, 1991; Gronlund, 1976).

4.3 Reliability of Teacher-made Tests

Establishing test reliability is an attempt to build on the quality of the whole test. However, it can be seen on Table 11 that all the teachers (100%) did not attempt estimate reliability of their tests.

Table 11:
Teachers' responses on whether they estimated reliability of their tests

Response	N	%
Yes	0	0
No	76	100

Source: Field Data

Similarly, the results given by the heads of department who were interviewed (Table 9 p55) indicated that teachers in their departments did not establish the reliability of their tests. The checklist applied on the Kiswahili teachers' tests indicated that the tests reliability index ranged from 0.13 to 0.49 Alpha. The indices were grouped as shown on Table 12.

Table 12:
Reliability index of Kiswahili TMTs

Reliability	N	%
0.6 – 0.69	0	0
0.5 – 0.59	0	0
0.4 – 0.49	17	22.4
0.30 – 0.39	12	15.8
0.20 - 0.29	32	42.0
0 – 0.19	15	19.8

Source: Field Data

The Table shows that 61.8% of the tests had reliability of between 0 and 0.29. Another 38.2% had a reliability coefficient ranging between 0.30 and 0.49. Although the findings showed the highest reliability coefficient was 0.49 (Appendix F & G), this falls far below the expected reliability coefficient of 0.65. For a general TMT,

experts in educational measurement have agreed that the reliability coefficient should be at least 0.65 if the scores are the only information available used to make decisions (Ebel & Frisbie, 1991). The findings seem to collaborate Davis (1991) observations that TMTs are infamous for their lack of reliability and many have a reliability coefficient approaching zero. Probably most reliability coefficient, fall in the range of 0.20 – 0.50.

Teachers need to construct their tests in such a way as to ensure that the scores obtained in a test on a particular occasion are likely to be very similar to those which would have been obtained if it had been administered to the same students with the same ability but at a different times (Hughes, 1997; Popham, 1990). Furthermore, an acceptable level of reliability coefficient indicates how much confidence we can place in our test results. If we are going to use test results as a basis for making decisions about pupils, then consistency in pupils’ achievement should be of paramount concern. Thus, if such decisions, some of which are irreversible are to be made, we must obtain the most reliable evidence concerning pupils’ learning. These decisions are so important and the consequences so significant that teachers need to devote considerable time and expense to establish and increase reliability of their tests (Hughes, 1997; Salvia & Yssedyke, 1991).

Further investigation on the reliability of Kiswahili revealed that there were no attempts by teachers to estimate scorer reliability even though all their tests items were the supply type. Table 13 show how teachers responded on establishing scorer reliability.

Table 13:

Kiswahili teachers’ responses on whether they estimated scorer reliability of their tests

Response	N	%
Yes	0	0
No	76	100

Source: Field Data

It is evident from the results that 100% teachers did not correlate their scores. Although they indicated that they scored their entire tests alone, none scored the test twice to verify on the scores awarded to the learners. Scorer reliability is established by correlating two sets of scores from two independent scores on one test. Kamp (1969) agrees with this by arguing that most teachers appreciate the necessity for at least two independent readings of each essay test but few attempts to work it out in practice. Scorer reliability is a prerequisite for high test reliability. The coefficient obtained reflects the degree of agreement between the scores and therefore the decisions made, based on these results tend to be sound (Hughes, 1991; Lyman, 1991). Thus, it can be argued that, if the scoring of the test is not reliable, then the test results cannot be reliable either (Dreckmyre & Fracer, 1991).

4.4 Item Analysis for Teacher-made Tests

High validity and reliability can be built into a test in advance through item analysis. Item analysis involves measuring the difficulty of test items and how items discriminate between low and high achievers. Teachers are therefore expected to conduct item analysis for each item in the test. However, Table 14 shows that 100% of the Kiswahili teachers did not conduct item analysis for their test items.

Table 14:

Kiswahili teachers' responses on whether they conducted item analysis

Response	N	%
Yes	0	0
No	76	100

Source: Field Data

This table shows that majority of the teachers did not follow procedures that promote item quality. This could be a reflection that item quality procedures are ignored by most teachers. Gullickson's study (in Anderson, 1992) is in harmony with these findings. In his study he found out that very few teachers undertook to improve their tests through item analysis. Furthermore, teachers did not value the use of statistical procedures on test items as a helpful strategy for improving the quality of their items.

Ayot (1984) shares this view when he observes that many teachers in Kenya seem not to follow psychometric procedures in language testing.

4.4.1 Item Difficulty

Students' scores that were analysed showed that the average p -value for all the tests was 0.71. Specifically figure 5 indicates that 13.2 % of the tests have a p -value of between 0.56 and 0.60, 31.6 % between 0.61 and 0.65 and another 15.8 % between 0.66 and 0.70.

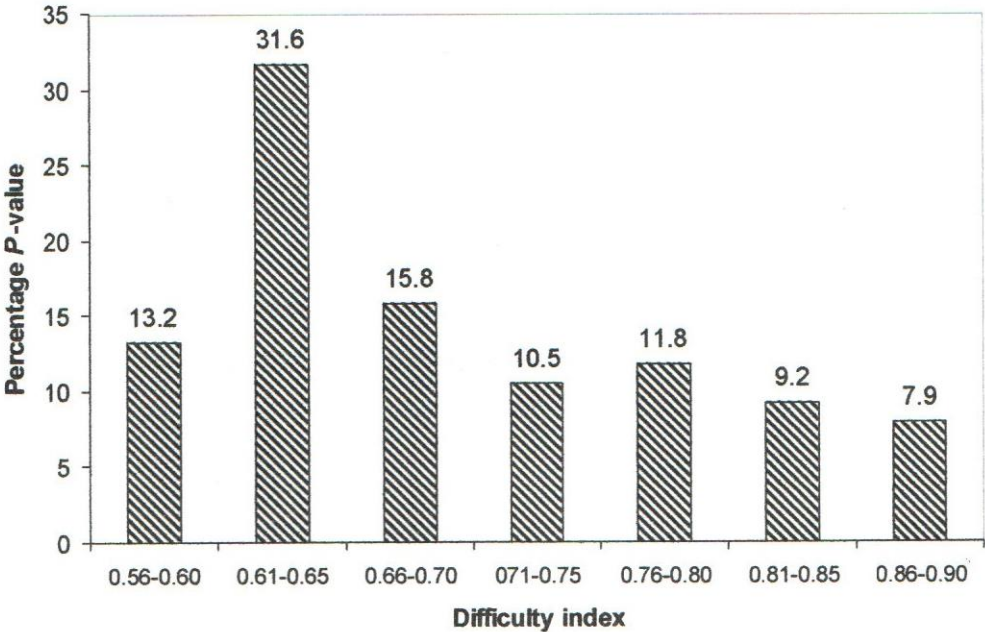


Figure 5: Percentage of P -value for the tests (Source: Field data)

In some specific tests, p -value of some items was found to be 0.0 (Appendix I question 17). This indicates that none of the learners scored the item correctly. In some other items p -value was found to be 1.0 meaning that all the learners scored the item correctly (Appendix I question 5). The basic measurement assumption is that there are systematic individual differences in the domain being measured. Thus, an item with p -value of 0.0 or 1.0 does not contribute to measuring individual differences and hence is almost certain to be useless (Hughes, 1997; Hopkins & Antes, 1985). When tests are designed for norm-referenced testing, p -value of near

0.5 is preferred over extremely easy or extremely difficult item. This is because items with p -value of near 0.5 have the maximum potential to be good discriminators.

Since the average p -value of TMTs that were analysed was approximately 0.71 it indicates that the TMTs were relatively easy. Most of the learners seem to have scored the items correctly. In a similar study by KNEC, it was found out that TMTs score for Teachers' Colleges candidates were in all cases much higher than the scores of the same candidates in KNEC examinations (Musau, 2004).

4.4.2 Item Discrimination

The scores obtained by each learner in each question were subjected to item analysis by the researcher and this revealed that the average Item discrimination index (D-value) was about 0.20. However, the D-value for individual test ranged between 0.03 – 0.26. D-value indicates effectiveness of an item in separating those who have grasped the concepts from those who have not. The Table 15 below shows the average D-value of the tests after they were grouped.

Table 15:
Percentage of D-value for the Kiswahili TMTs

Discrimination Index	Percentage
0.40 and above	0
0.30 – 0.39	0
0.20 – 0.29	56.2
0.19 and below	43.8

Source: Field Data

It is evident that 56.2 % of the tests had an average D-value of between 0.2 – 0.29 and 43.8 % have a D-value of 0.19 and below. None of the tests had an average D-value of above 0.3. According to the experts' evaluation of test items (Table 4 p42), items with D-value ranging between 0.20 – 0.29 are described as marginal items usually needing to be subjected to further improvement. Items with D-value of 0.19 and below are said to be poor items that are to be rejected or improved by revision. In

some specific tests the D – value was 0 (showing no discrimination), other items had negative discrimination (favouring low achievers) and nearly all the items having a D – value of below 0.29 (Appendix H & I). It appears that nearly all the test items for the tests that were investigated needed to be improved by revision. This suggests that many of the items constructed by the teachers did not attain the required quality. Gandiye (1991) in his studies on school-based assessments in Tanzania reported that TMTs had a poor discrimination index as compared to external examinations.

An achievement test has the principle function of distinguishing between different levels of achievement as clearly as possible. It is desirable for each item to have as high discrimination index as possible. Since an item answered correctly or incorrectly by all cannot discriminate at all, such items have no place in the achievement test (Ebel & Frisbie, 1991). Research has shown that when items discriminate negatively, that is, the low achievers are scoring higher than the high achievers in that particular item, the item is ambiguous. Ambiguity tends to confuse the better pupils than the poorer pupils, causing items to function less effectively (Gronlund, 1976; Hughes, 1997; Smith, 1984).

Item discrimination indices are indicators of problematic items. A low or a particularly negative D – value should alert the teacher to the possibility that an item is defective and needs corrective surgery or mercy killing (Popham, 1990). Nonetheless, it has been observed that the analysis data of the items that have been revised and tried out with another representative group showed an improvement in both test item difficulty level and discrimination power (Ebel & Frisbie, 1991).

4.5 Factors Hindering Teachers from Constructing Quality Tests

A number of reasons were advanced by the respondents as shown in Table 16 as making teachers unable to develop quality tests.

Table 16:

Factors hindering Kiswahili teachers from constructing quality tests

Responses	Yes	No
Skills were forgotten	26.3	73.7
Not a school requirement	64.5	35.5
Not an inspectorate requirement	59.2	40.8
Inadequate training	61.8	38.2
Heavy workload	92.1	7.9
Limited content	78.9	21.1

Source: field Data

Specifically, 26.3 % of the teachers indicated forgetting the skills, 64.5 % not a school requirement, 59.2 % not an inspectorate requirement, 92.1 % heavy workload, and 78.9 % were of the view that insufficient content in test and measurement course could have affected their construction of quality test. The heads of Language Departments interviewed felt that similar factors affected the construction of quality tests. The HOD responses are shown in Table 17.

Table 17:

H.O.D responses on factors affecting construction of quality tests

Item	Responses	Yes	No
Factors constraining constructions of quality tests	• Heavy workload	100	0
	• The quality of tests is never inspected. They only check on teaching records and materials	100	0
	• School administration is not concerned about the quality of the tests.	100	0
	• Inspectors do not give advice on construction of quality tests	100	0
	• Teacher education concentrates more on teaching skills and not testing skills	100	0
	• Limited scope in test and measurement course offered at the university	100	0
	• University supervisors during TP concentrated on teaching ignoring testing	100	0

Source: Field Data

In the school setting, it was observed that there were no clear guidelines either from the school administration or the Ministry of Education on how tests should be evaluated for their quality and who should do it. The heads of departments who were supposed to undertake this responsibility, acted only as custodians of past papers and students test scores.

The inspectors on their part were blamed for being unconcerned with testing in their routine inspection. It was found that they concentrated their efforts on the schemes of work, records of work and assessment records. They did not check on the quality of

the tests. Since tests are one of the means of measuring the impact of an instructional programme on students, they should be a legitimate concern of the inspectors. Where a need arises, the School Inspectors should consider consulting testing experts, probably from a nearby university or college. This would ensure that out dated, unreliable, or invalid tests are eliminated. In addition, they should ensure that tests are continually revised to keep up with curriculum changes from year to year (Mulder, 1993).

As it has been noted, 92.1 % of the Kiswahili teachers felt that their inability to construct quality tests resulted from a heavy workload as measured by the number of lessons taught per week. The analysis of lessons taught per week was as reflected on figure 6.

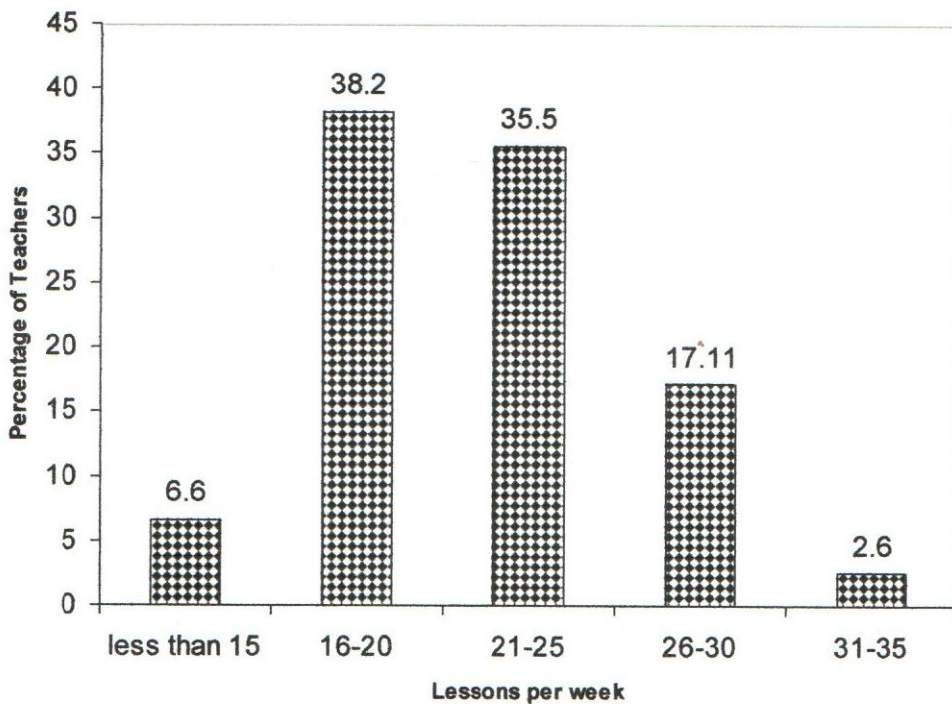


Figure 6: Number of lessons for Kiswahili teachers per week (Source: Field Data)

Surprisingly, 80.3 % of all the teachers had 25 or less lesson per week. The 25 lesson per week is below the minimum required workload. The Teacher Service Commission (T. S. C) has set 27 lessons a week as the minimum load for a teacher. In view of this, the situation will be worse if the recent pronouncement by the outgoing T. S. C

Secretary that the minimum lessons per week be increased from the current 27 to 32 will be implemented (Nation Reporter, 2004)

Teachers tend to put off test preparation to the last minute due to 'heavy workload'. A last minute test is likely to be a poor test. Further, such test cannot have a constructive influence in motivating and directing student learning than a good test prepared early in the course would (Smith, 1984). Moreover, teachers should not rely on a single summative assessment to evaluate students' achievement after a course of instruction. Rather teachers should test progress towards instructional objectives at least twice or thrice a week (Salvia & Ysseldyke, 1991). This certainly calls for teachers and schools to invest more time and resources in testing and test construction.

The admission that there is inadequate training on test construction is not surprising (Table 17). This is because there is much more emphasis on teaching skills that has left teachers incompetent in test construction. This is more specifically in the ability to measure students' attainment, to capitalize on the instructional purposes, and to develop fair and interpretable summaries of students' achievements (Ebel & Frisbie, 1991; Kamp, 1969). The findings from the interview with the heads of departments (Table 17) revealed that the teaching practice during training emphasised more on instructional process. The main concerns of the supervisors were schemes of work, lesson plans and the teaching methods and techniques employed by the student teacher. Little emphasis was put on assessment of the quality of the tests. In addition, student teachers were not given practicals on tests and measurement either individually or in groups. More emphasis was given to micro-teaching and development and use of media resources.

In addition, 78.9% of the teachers felt that the content of Test and Measurement course taught at universities was insufficient. During the interview, the heads of Language Departments expressed the view that Test and Measurement is taught just as a unit in Educational Psychology and even so, it is given little emphasis. Therefore, the content is presented as a requirement to pass collage examinations but not as a practical skill required to be used in the field. Thus, lack of emphasis of the course at the university coupled with the neglect of this area by the school administration and

the inspectorate sent a signal to students and practising teachers that following procedures that promote quality tests was not important. This agrees with Popham (1990) observation that teachers who do not possess the skills needed to evaluate tests will in all probability, continue to create and employ flawed assessment tools. Consequently, such teachers will often make inappropriate decisions about the learners who must suffer through these inadequate tests. Thus, it is possible to argue with confidence that testing is typically carried out by teachers whose formal training in assessment is inadequate and narrow in focus (Anderson, 1992).

CHAPTER FIVE

CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS

5.1 Introduction

This study investigated the quality of Kiswahili language Teacher-made tests. This chapter presents the conclusions and implications of major findings, recommendations and areas that need further research. The following conclusions and implications were reached under each research objective.

5.2 Conclusion and Implications

5.2.1 Validity of Teacher-made Tests

The five point criteria that was used to establish the validity of teacher-made tests were; basing tests on instructional objectives, using Blooms Taxonomy of Educational Objectives, using the table of specification, using other subject specialists, and using item bank.

5.2.1.1 Basing tests on instructional objectives

Although test validity is a major facet of enhancing test quality, the findings of the study showed that teachers did not employ procedures that promote test validity. Failure to follow these procedures greatly jeopardised the quality of their tests. The revelation that teachers did not base their tests on instructional objectives implies that there was no evidence that their tests had content validity. In addition, the information obtained from such a test may be of questionable value in making instructional decisions. Content validity is the core of all achievement tests. If achievement tests are based on objectives, rather than on teaching notes and textbook content, they will provide a truer picture of what has actually been achieved.

5.2.1.2 Using Blooms Taxonomy of Educational Objectives

The results of the study showed that most teachers did not classify instructional objectives in accordance with the Bloom Taxonomy of Educational Objectives. This in turn led to test items focusing less on higher-level abilities such as analysis, synthesis, and evaluation. This implies that the teachers could have been emphasising to the learners that learning is all about memorization and discouraging them to develop the use of more complex mental processes. Factual details learned in a course are most likely to be forgotten while the understanding of principles and their application to new situations would show little loss of retention with time. Thus mere memorization of information does not contribute to effective learning, and is likely to promote negative backwash. It is apparent that teachers need to construct tests that encourage the learners to focus on all areas of the course and all levels of cognitive domain, thus promoting beneficial backwash. Therefore, it can be concluded that although classifying objectives into various levels of cognitive objectives is essential for achieving high content validity, its use has been poor amongst the secondary school teachers.

5.2.1.3 Using the Table of Specification

The findings further indicated that teachers avoided the use of a table of specification in constructing their tests. The absence of the use of the table of specification by the Kiswahili teachers raised doubts about the content validity of their tests. This in turn implies that teachers who plunge directly into item writing are likely to produce a lopsided test. Without an advance plan therefore, some areas will be over represented while others may remain untouched. There is need for teachers to invest more effort on building the table of specification to promote content validity of their TMTs.

5.2.1.4 Using other Subject Specialists

From the findings to this study it is evident that majority of teachers did not seek a second opinion from other subject specialists in their schools in establishing content related validity. This implies that the individual teachers depended on their own subjective judgement about the content validity for their tests. In addition to the

observation that teachers did not use the table of specification in constructing their tests, the probability that TMTs would attain high content validity is very low. Too often, the teachers will end up testing trivialities, what is to test and score as opposed to the requirements of the syllabus.

5.2.1.5 Using Item Bank

The study also showed that most teachers did not build item banks for their test items. An item bank is the only sure way that items selected for use in a test are of high quality. Its neglect implied that most teachers were using test items whose quality has not been verified. Test items stored in item banks have their discrimination index and difficulty level analysed and found suitable. Since the results also showed that most teachers relied on past papers to construct their tests whose quality has not been checked, it is an indication that most TMTs perpetuated poor quality items. It is therefore necessary for teacher to build items banks for all their tests and those adapted from other sources. Consequently, teachers need to field test such items frequently to keep on building on their item banks.

It seems that despite the importance of establishing content validity in TMTs, teachers depended on their own perception and subjective appraisal of the validity of their tests. There is need to persuade them to be more committed to psychometric procedures that promote test validity.

5.2.2 Reliability of Teacher-made Tests

The results of the study showed that nearly all the teachers did not undertake to estimate the reliability of their tests. The TMTs were found to be unreliable with a reliability coefficient ranging from 0.13 to 0.49. In addition, there were no indications that teachers attempted to establish scorer reliability. Thus, the scores so obtained were not validated. In essence, the revelation that teachers did not estimate reliability of their tests and establish scorer reliability implies that the decisions made based on these scores could not be relied upon. Teachers appear reluctant to commit their energies to the rigorous procedures of calculating reliability coefficient of their tests.

5.2.3 Item Analysis for Teacher-made Tests

The findings of the study indicated that nearly all the teachers did not conduct item analysis to improve on the item quality. This implies that most of the teachers did not check what items required to be stored, revised or rejected; yet this is the very basis for a quality test. Average *P*-value showed that the tests were relatively simple. Furthermore, none of the items achieved the required *D*- value of 0.4. Consequently, nearly all the items analysed during the study ought to have been rejected or revised. In essence, it can be argued that failure to conduct item analysis resulted into poor quality test since item analysis is crucial to the overall quality of a test.

5.2.4 Factors Hindering Teachers from constructing Quality Tests

The findings of the study revealed that the development of quality tests amongst the teachers was affected by school administration not requiring that teachers construct quality tests, heavy workload on part of the teachers, the department of Quality Assurance and Standards in the Ministry of Education Science and Technology not requiring that teachers follow the established procedures and practises of tests construction and inadequate training at the university and colleges especially in practical skills required to develop quality tests

The failure of the school administration to make it mandatory for teachers to construct quality tests implies that teachers will continue administering defective instruments in measuring learners' achievement. Similarly, as long as the teachers feel that they are being overworked, the chances are that they will continue administering poor tests that are easy to construct and score in order to meet deadlines fixed by the school administration. Lack of emphasis by the Department of Quality Assurance and Standards on the construction of quality tests in their routine inspection implies teachers and the school administration will be viewing procedure of enhancing quality testing as unimportant in the education system. This means that the inspectors' effort to bring about remarkable improvement in instructional process might not be achieved since the collection of data on the performance of learners will continually be collected using poor quality tests. Inadequate training at the university and colleges especially in practical skills required to develop quality tests implies that student

teachers are likely to form a notion that testing is a peripheral component of instruction. As long as much emphasis is placed on construction of learning materials and application of teaching methods during microteaching and teaching practice sessions, teachers may be left uncommitted to adapt procedures that promote quality testing.

5.3 Recommendations

The conclusions and implications of this study suggest that teachers do not follow the psychometric procedures that promote valid and reliable tests. This could have been caused by inadequate training and lack of proper mechanism to enforce the construction of quality tests. As a result, the findings of this study have led to the following recommendations.

1. Cooperative evaluation needs to be adopted. Teachers in the same schools zones or district should be encouraged to adopt uniform procedures for the construction and scoring of the tests. This will lay the groundwork for standardized tests. This is possible if the following structures on the ground are expanded and utilized to the maximum.
2. The scores obtained from standardized TMTs need be integrated in the National Examinations. This is because TMTs prepared by the teacher are likely to fit the content and objectives of a particular course better than would a test prepared by anyone else. Externally imposed assessment (K. C. S. E) relies heavily upon a single test score and single cut-off points for decision-making. These day-to-day assessments by the teachers in the classroom are the most important yet are undervalued resource in the quest for educational excellence. The inclusion of TMTs in the National Examinations will in itself be a motivation for the teacher and encouragement in competency in testing.
3. The Ministry of Education could organize workshops and in-service courses as a form of staff development for teachers. Such in-service courses in test and measurement would play a critical role in acquainting teachers with new skills and knowledge on test development. This can be done easily by expanding, equipping and utilizing the existing Teacher Advisory Centres (T. A. C) at the

zonal levels. Due to the crucial role tests play in our nation's education enterprise especially with the ever changing curriculum, it is inevitable that the government should bear the responsibility for providing financial subventions specifically to foster the emergence of more sophisticated, hence more useful educational measurement technology.

4. Since testing is an integral part of the teaching and learning, equal investment in terms of time and resources should be accorded to both instruction and testing. Teachers will be required to spend more time in testing. The school administration to invest more in printed material, files, calculators, computers and personnel to assist teachers in statistical analysis of the tests.
5. There is need for universities and colleges to evaluate test and measurement courses they offer to ensure that teachers are well equipped with testing technology. Quality tests with evidence of validity, reliability and item analysis should be made part of the requirement of the projects for student teachers during their teaching practice. In addition, there is need to make test and measurement and measurement an independent subject rather than a unit in psychology. Its content will therefore be enriched and the course made more practical oriented.
6. The Ministry of Education Science and Technology needs to establish a strong process of inspection to ensure and monitor the use of quality TMTs in secondary schools. The inspectorate should also consider liaising with test consultants especially from local universities to get acquainted with the latest knowledge and research findings in tests and measurement. Equipped with this knowledge they will be able to meet testing challenges in secondary schools.

5.4 Suggestions for Further Research

The foregoing conclusions and recommendations suggest several directions for future research that the researcher believes deserve consideration.

- This research concentrated more on test construction. There is need to carry out research on the competence of Kiswahili secondary teachers in test administration and interpretation of test scores.
- There is need to establish test batteries (in Kiswahili and other subjects) and investigate on how accurately they could predict students' final examination scores with an aim of integrating these scores in the final examinations.
- Studies need to be carried out to establish what other factors could be affecting teachers in test construction; administration and interpretation of test scores.
- There is need also to carry out research on the attitude of both practising and training teachers on testing.
- An investigation need to be done to establish how equipped the inspectors are in terms of skills and resources to deal with testing practises in schools.

REFERENCES

- Airasian, P. (1991) Classroom Assessment. New York; Mc Graw-Hill inc.
- Anastasi, A (1982). Psychological Testing. New York; Macmillan Publishing Co inc.
- Anderson, L.W. (1992). The effective Teacher. New York: Random House.
- Anderson, L.W. (1989). The effective Teacher. New York: Random House.
- Ashworth, A.E (1982) Testing for Continuous Assessment. London; Evans Brothers limited.
- Atieno, R (2001.October,14 pg 24) New Website to Promote Kiswahili. Sunday Times; Kenya Times.
- Ayot, H. O ed (1984) Language For Learning. Nairobi: Macmillan
- Best, J.W & Khan, J.V (1992). Research in Education.New Delhi; Prentice- hall.
- Bhask, T (2002). Understanding Social Science Research. London; Sage Publications.
- Bolyard & Hatch (2003. December,23). Continuous Assessment. EQ Review Vol.1, No.1, pp 1-3. USAID
- Cangelosa, J. (1990) Designing Tests to Evaluating Student Achievement. New York; Longman
- Cohen, L & Manion, L (1997). Research Methods in Education. London; Routledge
- Cohen, R. J, Swerdik, M. E & Smith, D. K (1992). Psychological Testing and Assessment: An introduction to Test and Measurement. Mountain View; Mayfield Publishing Company.
- Constitution of Kenya Review Commission (2002). Draft Bill of the constitution of Kenya. Nairobi; Review Commission.
- Davis, A (1988). Procedures in Language Test validation. In Hughes, A. Testing English for University study. Oxford; Modern English Press.
- Davis, A (1991) Principles of Language Testing. London; Basil Blackwell.
- Dreckmeyr, M. A & Fracer, W (1991) Testing in Biology and Physical Science. Pretoria; Haum Tertiary.

- Duke, D. L (1990). Teaching: An Introduction. New York; McGraw-Hill Publishing Company.
- Ebel, R & Frisbie, D (1991). Essential of Educational Measurement. New Jersey; Prentice-Hall.
- Gall, M, Borg, W & Gall, J (1996) Educational Research. New York; Longman Publishers
- Gandiye, P. P (1991). Assessment of National Educational Goals and Objectives. The Case of Continous Assessment. Unpublished Paper Presented at IAEA 17th conference.
- Gregory, R. J (1996). Psychological Testing. Boston; Allyn and Bacon.
- Gronlund, N.E (1976) Measurement and Evaluation in Teaching 4th Ed .New York;
- Gronlund, N.E (1982). Constructing Achievement Tests. New Jersey; Prentice hall
- Hopkins, C & Antes, R (1985). Classroom Measurement and Evaluation. Illinois; Peacock
- Hughes, A. (1997). Testing for language Teachers. Cambridge: Cambridge University press.
- Idol, L & Jones, B (1991). Educational Values and Cognitive Instruction. New Jersey; Lawrence Erlbarum Associate Publishers.
- Kathuri, N.J & Pals, D.A (1993) Introduction to Educational Research. Njoro; Education Media Centre.
- Kellaghan, T & Greany, V (2004). Assessing Students Learning in Africa. Washington D.C; World Bank.
- Kimemia, J (2002) Language, Curriculum Process and Emerging Issues in Education. Nairobi; New-era International.
- Koul, L (1993). Methodology of Educational Research. New-Delhi; Vikas Publishing House.
- Lado, R. (1964) language Testing. London: Longman
- Lefrancois, G.R (1991). Psychology for Teaching. Belamont; Wasworth Publishing Company

- Lyman, H (1991) Test Scores And What They Mean. New Jersey; Prentice Hall.
- McMillan, J. H (2003). *Fundamentals Assessment Principles for Teachers and School Administrators*. Practical Assessment, Research & Evaluation, 7(8). Available <http://PAREonline.net/getvn.asp?v=7&n=8>.
- Marshall, J & Hales, L (1977). Classroom Test Construction. California; Addison-Wesley.
- Mbaabu, I (1991). Historia ya Usanifisha wa Kiswahili. Nairobi; Longman.
- McArhtur, J (1991) A Foundation Course for Language Teachers. Cambridge; Cambridge.
- Mucheru, O (2005, April 27). Principle of Assessment. Unpublished Seminar Paper. Kenya National Examination Council.
- Mugenda, O.M & Mugenda, A.G (1999) Research Methods: Quantitative and Qualitative Approaches. Nairobi; ACTS press.
- Miulder, J. C (1993). Statistical Techniques in Education. Cape Town; Haum Tertiary Publishers.
- MOEST (2000). Handbook for Inspection of Education Institutions. Nairobi. Government Press.
- Murphy, K. R & Davidshofer, C. O (1988). Psychological Testing. New Jersey; Prentice Hall
- Murphy, P. Ed. (1996). National Assessments. World Bank; Washington, D. C
- Musau, K (2004). Educational Measurement & Evaluation. A Guide for Teachers. Njoro; Egerton University
- Nation Reporter (2004 June 24 pg 6). Schools Lack 9000 Tutors, Say Ongwae. Daily Nation. Nairobi; Nation Media Group.
- Ndagi, J.O (1984) The Essentials of Research. Ibadan; University Press.
- Noll, V. H, Scanell, D.P & Noll, R. P (1972). Introductory Readings in Educational Measurement. Boston; Houghton Muffin Company.

- Nga'ng'a, N (1996) An Investigation Into The Relationship Between Form Four Teacher-Made tests and KCSE Examination. Unpublished Masters Thesis. Kenyatta University.
- Oguninnyi, M.B (1986). Educational Measurement and Evaluation. Lagos; Longman.
- Oirere, S (April.10, 1999 pg 16). *Are Exams the Best Way to Testing Ability?* Kenya Times. Nairobi
- Popham, J (1990). Modern Educational Measurement. New Jersey; Prentice Hall.
- Provincial Director of Education, Rift Valley (2003). School Inspectors' Report. Unpublished Document.
- Rensick, L.B & Kloper, L.E (eds) (1989) Curriculum. Pennyslivia; ASCD
- Rold. G. & Haladyna, I.(1986) A Technology of Test item writing.New York;AcademicPress.
- Salvia, J & Ysseldyke, J (1991) Assessment. New Jersey; Houghton Mifflin Company.
- Sattler, J. M (1992). Assessment of Children. San Diego; Jorome, M. Sattler Publishers inc.
- Shapiro, E (1989) Academic Skills Problems. New York; Guildford Press.
- Shiundu, J. S & Omulando, S. J (1992). Curriculum Theory and Practice in Kenya. Nairobi; Oxford University Press.
- Siringi, S (2005, Febraury 23, pg 5). *Study Exposes Cheating Ways of Academies.* The Daily Nation. Nairobi; Nation Media Group.
- Siringi, S (2001.October, 1 pg19) *Towards a Balanced Examination.* The Daily Nation; Nation Media Group.
- Smith, F (1984) Constructing and using Achievement Tests in the classroom. New York; Peter Lang Publishers.
- Stanley, J. C & Hopkins, K. D (1972). Educational and Psychological Measurement and Evaluation. New Jersey; Prentice-Hall inc.
- The Holy Bible. King James Version.Chap:12 (2001). London; Cambridge University Press

- Thorndike, R & Hagen, E (1977) Measurement and Evaluation in Psychology and Education. New York; John Willey & Sons inc.
- Thorndike, R. M (1997). Measurement and Evaluation in Psychology and Educations. New Jersey; Prentice Hall.
- Underhill, N (1995). Testing Spoken Language. New York; Cambridge University Press.
- Wood, R. (1995). Assessment and Testing. Cambridge; Cambridge University Press.

APPENDIX A

TEACHERS' QUESTIONNAIRE ON THE QUALITY OF TEACHER-MADE TESTS

The purpose of this study is to investigate quality of test in our schools. You have been randomly selected among others to participate in this study. Your opinions will contribute greatly in assessing classroom testing in order to help teachers improve the quality of their tests.

GENERAL INSTRUCTIONS

- a) Please read each item carefully before answering it
- b) All information given will be treated with high level of confidentiality
- c) This is not a test and therefore there are no right or wrong answers

Section A

PERSONAL INFORMATION

For this sections, put a tick in the square that corresponds with your particulars

a) GENDER Male Female

b) QUALIFICATION M.ED PGDE B.Ed Dip. Ed Others

Specify ____

c) LESSONS PER WEEK Less than 15 16-20 21-25
26-30 31-35 Over 36

d) TEACHING EXPERIENCE

Less than 3 years 3-5 years 6-10 years
11-15 years 16-20 years over 20 years

Section B

Answer all the questions in this section by ticking in the brackets and filling in the spaces provided.

1. How often does your school require you to give tests in a term?

- (a) Once (b) twice (c) thrice (d) more than thrice

2(i) Do you use any statistical procedures to establish whether your test consistently measures the intended outcomes (Test Reliability) Yes No

(ii) If yes, what method of establishing reliability do you use?

- (a) Test-retest (b) Equivalent-forms (Alternate) (c) Split-half
(d) Single test with Kuder-Richardson Formula (e) Single test with Cronbach alpha Formula

(ii) What is the average reliability coefficient of the last 3 tests?

- (a) Less than 0.44 (b) 0.45-0.55 (c) 0.56-0.65 (d) above 0.66

3(i) How do you score (mark) your tests?

- (a) Once and alone (b) same test re-marked by another teacher

(ii) If your answer is (a) above, do you remark the same test to determine whether there is variation of scores? Yes No

(iii) If yes, do you compute the correlation coefficient index? Yes No

(iv) If yes, what is the coefficient index of your last test?.....

4(i) If you answered 'No' in question 2(i) above, what makes you not to establish the reliability of your tests? (You can tick more than one answer)

(a) You were never taught the skills in college (b) You forgot the skills

(c) The school does not enforce the requirement (d) The Inspectors do not enforce the requirement (e) Others

(ii) If your answer is 'others' above, please explain.....
.....

5(i) Do you set all your tests alone? Yes No

(ii) If No, who else do you involve?.....

6 (i) What guides you in constructing your test?

(a) Instructional Objectives (b) Past papers (c) Lesson notes

d) Others

(ii) If your answer is (a) above, please fill in the table below using any 3 items from any of your end of term test.

FORM:		
Content	Objectives	Items
1.		
2.		
3.		

7 (i) Do you draw a table showing the composition of items in your test that range from simple recall to those that require evaluation (Blooms Taxonomy)

Yes NO

(ii) If yes, please indicate on the table below, the number of items from your end-term test, falling under each category of Bloom's Taxonomy of Educational Objectives. (Use tally method)

	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
End-of-term test (Date.....)						

8(i) Do you draw a table showing an agreement of the content, instructional objectives, test questions and the level of cognitive domain being tested when constructing a test? (A table of specification) Yes No

(ii) If yes, do you have a skeleton of a table of specification that you can make available to the researcher? Yes No

9 (i) Do you build an item pool (item bank) for test item samples?

Yes No

(ii) If yes, write the details of any of your stored item below.

Front of the Card	Back of the Card

10 (i) Do you use any statistical method to verify whether your test items were hard or easy or were able to discriminate between the low achievers and the high achievers (Item analysis)? Yes No

(ii) If yes, what is your acceptable level of:

a. Item difficulty index?.....

b. Item discriminating power index?.....

11. What do you think hinders teachers from establishing validity, reliability and conducting item analysis of their tests, in relation to?

i) Training of teachers

.....
.....

ii) School setting e.g. workload

.....
.....
.....

ii) The role of the Inspectorate

.....
.....
.....

12. (i) Do you think the integration of the teacher-made tests in the final examination

by KNEC will encourage teachers to construct more accurate tests? Yes

No

(ii) Give reasons for your answer.

.....
.....
.....

APPENDIX B

INTERVIEW GUIDE FOR KISWAHILI HEADS OF DEPARTMENT ON THE QUALITY OF TEACHER-MADE TESTS

PERSONAL INFORMATION

For this sections put a tick in the square that corresponds with your particulars

a) GENDER Male Female

b) QUALIFICATION M.ED PGDE B.Ed Dip. Ed Others

Specify _____

c) LESSONS PER WEEK Less than 15 16-20 21-25 26-30
31-35 Over 36

d) TEACHING EXPERIENCE

Less than 3 years 3-5 years 6-10 years 11-15 years

16-20 years over 20 years

QUESTIONS

1. Do you make any contribution towards ensuring that the teachers' tests are accurately constructed?

2. What are the main sources of test items for your teachers? _____

3a). Do teachers in your department undertake to establish the scorer reliability of their tests? _____

b) Do teachers in your department undertake to estimate the reliability coefficient of their tests? _____

c) Do you have an item bank in your department? _____

d) In the view of (a), (b) and (c) above, what would you identify as the major constraining factors? _____

4a) Is there evidence that teachers in your department do

i) Build the table of specification?

ii) Apply Blooms taxonomy of educational objectives? _____

iii) Conduct item analysis?

b) If none of 4(a) is done, what do you consider to be the major constraining factors? _____

5. What do you suggest should be done to ensure that teachers follow all the procedures for constructing accurate test? _____

APPENDIX C

CHECKLIST FOR EVALUATING THE QUALITY OF TEACHER-MADE TESTS

GENERAL INSTRUCTIONS

- a) Please read the items carefully before answering them
- b) All information given will be treated with high level of confidentiality

Section A

PERSONAL INFORMATION

For this sections put a tick in the square that corresponds with your particulars

a) GENDER Male Female

b) QUALIFICATION M.ED B.Ed PGDE Dip.Ed Others
specify _____

c) TEACHING EXPERIENCE Less than 3 years 3-5 years 6-10 years

11-15 years 16-20 years over 20 years

LESSONS PER WEEK less than 15 16-20 21-25 26-
30 31-35 Above 35

Section B

This section should be filled by the researcher.

1. Evidence of instructional objectives versus the content as a basis of test questions.

2. Presence of table of specification.

3a) Evidence of application of Blooms taxonomy of educational objectives in test construction _____

b) A table to assess inclusion of levels of cognitive skills in TMTs

	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation
End-of-term test						

4 (a) Evidence of estimating reliability coefficient.

b) Evidence of lengthening the test to increase reliability.

c) Table of details to enable the researcher calculate coefficient alpha and conduct item analysis for each teacher.

(A separate mark sheet will be used and a computer will be used to calculate coefficient alpha)

	Scores for each item										
Examinee	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
Kamau											
Kioko											
Wanjala											
etc											

5. Evidence of item analysis as regards;

ii) Item difficulty

iii) Item discriminating power

6. Presence of item bank (a file containing item record cards) _____

APPENDIX D
SAMPLE TEST 1

SHULE YA MTAKATIFU YUSUFU

S. L . P 14964,

NAKURU

KISWAHILI KIDATO CHA PILI MUHULA WA PILI

1. Kamilisha tashbihi zifuatazo. (al. 4)
 - i. Mwandiko mbaya kama
.....
 - ii. Kujiingiza katikati ya mambo
kama.....
 - iii. Adimika kama
.....
 - iv. Mzima kama
.....
2. Jaza nafasi zilizoachwa kwa msamiati ufaao. (al. 4)
 - i. ni majira baridi.
 - ii. ni majira ya mvua chache.
 - iii. ni siku ya tatu kabla ya kesho.
 - iv. ni wakati kuanzia saa sita hadi saa nane za
mchana
3. Taja aina ya maneno yaliyopigwa mstari chini yake. (al.4)
 - i. Mzee alitembea polepole
 - ii. Tulifika nyumbani kabla ya saa sita.
 - iii. Mtakwenda tu ingawa sina pesa.
 - iv. Ajabu! Unene wote ule hakuna aliyemshinda mbio.
4. Tunasema mlango u wazi. Tumia maneno yafuatayo badala ya mlango na uweke kiambishi kinanchofaa cha ngeli. (al. 4)
 - i. Chumba.....
 - ii. Pango
 - iii. Mashimo.....
 - iv. Nyumba

5. Toa maelezo ya sehemu za mwili zifuatazo. (al. 4)
- i. Kisigino.....
 - ii. Shavu.....
6. Taja methali zinazosisitiza haja ya uvumilifu. (al. 4)
- i.
 - ii.
 - iii.
 - iv.
7. Jaza mapengo katika sentensi zifuatazo. (al. 4)
- i. Sehemu ya nyuma ya jahazi huitwa na mbele huitwa
.....
 - ii. Huchuja damu pamoja na mkojo kutoka kwenye damu.
 - iii. Kazi ya ukarani wa pesa ni
8. Geuza sentensi zifuatazo kwa kutumia jinsi ya kufanyiza. (al. 6).
- i. Ile pombe ilimfanya alewe sana.
 - ii. Ndoo hii nitaifanya ijae maji.
 - iii. Dawa ilimfanya mtoto kupona.
9. Ziandike sentensi zifuatazo kufuatana na maagizo uliyopewa mwishoni mwa kila sentensi. (al. 6).
- i. Mahindi yasagiwe wagonjwa (Anza: Wagonjwa.....)
 - ii. Ukiondoka mapema utafika kesho jioni; (Anza uispoondoka.....)
 - iii. Ni lipi ambalo limekukasirisha? (Andika bila kutumia amba)

LIKIZO NJEMA

*****SC***** IDARA YA LUGHA*****

APPENDIX E

REPRESENTATION OF ITEMS IN BLOOM TAXONOMY FOR SAMPLE TEST 1

	Knowledge	comprehension	Application	Analysis	synthesis	Evaluation
Question 1	1	-	-	-	-	-
Question 2	1	-	-	-	-	-
Question 3	1	-	-	-	-	-
Question 4	-	1	-	-	-	-
Question 5	1	-	-	-	-	-
Question 6	1	-	-	-	-	-
Question 7	1	-	-	-	-	-
Question 8	-	1	-	-	-	-
Question 9	-	1	-	-	-	-
Total	6	3	0	0	0	0

APPENDIX F
SAMPLE TEST 2

MOI NDEFFO SECONDARY SCHOOL

KISWAHILI FORM II END TERM EXAM TERM II YEAR 2004

NAME.....ADM.....CLASS.....DATE.....

Soma kifungu hiki kisha ujibu maswali.

"Nikwambia nini mambo ya Bwana huyo, Nikikueleza kutwa hayeshi. Na utoshe huo wivu alikuwa nao; na litpasho hilo domo lake. Si wa nguo si wa kula, yeye mwambie wivu na hamaki tu - basi! Lakini navumilia tu. Nikisikia hamrore innanze hujingilie ndani nikafunga mlango, nikamwacha anaimba kama mkwezi wa minazi. Simjeli, namwona kama mwenda wazimu tu. Kitu kinachonikera zaidi ni wivu wako tu..... Si wivu aliojaliwa Bwana huyu. Hivi sasa hivi sijui kama hayuko darini anatusikiliza. Ekitoka nje kaniganda; nikiingia yuko nyuma; nikiulizwa hujambo vita; nikizungumza na mtu yumo pembeni anasikiliza! Mradi sina raha. Hapa siwezi kupumua mpaka atoke kwenda kazini. Pana siku yeye huja kuvuzia. Akiwa ndani naomba atoke nje. Afadhali nyinyi ambao hamjaolewa mna starehet!"

Bahati ambayo alikuwa yuko katika hamu ya kuolewa, alifadhaishwa na maneno ya Kidawa, lakini hakupenda kumwonyesha. Akamwambia, "Stahimili tu, Kidawa. Sasa utafanyaje? Kuna wasichana wazuri wamekaa majumbani, wanamtaka kama huyo wako na hawampati. Stahimili shoga yangu!"

"Ndiyo nestahimili hivi, shoga," Kidawa alisema shingo begani. "Nikienda shitaki kwetu baba ananitoa mbooi, ananiambia: 'tokomea huko huko, usije na maneno ya kipumbavu hapa, nyie wanawake anajulikana mliyoyo?"

Bahati akamuliza, "Je, hujapata mtoto bado?"
Akamjibu, "Ah, mwaka wa tano huu hata alama. Hata hayo baraka imeondoka."

"Mungu atakujelia upate nke na dume, uzau, Ujukuu, wawe na kheri na wewe. Mungu hamshehu mje wake!"

"Anina kwa baraka za Mungu na Ntuno, ulimi wako uwe kabuli," aliitikia kidawa....."Wewe je, Bahati, hujataka kuolewa?"

"Nina mohumba wangu. Karibu nitaolewa," Bahati alisema.

"Ndiye yule yule Idi, nini?"

"Ndiye yule yule!"

Kidawa akamjibu, "Haijembo, umeunga nwenzangu. Si kama wangu nimi, kutwa ananing'iniwa na mkanzu tu, hajui unaridadi, hajui kujipura!"

"Kidawa una jua?" Bahati alimkatiza, "Sasa nyingi nipe buibui langu."

"Mbona ^{mara} ~~mara~~, shoga yangu, lakini tukazungumza?" Bahati akamjibu. "La, sikai. Tutatona siku nyingine."

(a) Taja sifa za Bwana anayozungumziwa

(a1. 11)

(b) Wivu wake unadhihirikaje

(a1 4)

....2/

...3....

2. Taja viwakilishi ambishi vya nafsi katika umoja na uvitumie katika sentensi (al. 6)

3. Kamilish methali hizi (al.4)

(i) Uji wa moto haupuzwi kwa neka ya ..

(ii) ya meji haifumbatiki

4. Tambua vivumishi katika sentensi hii na utoje ni vya aina gani (al. 6)

(i) Mwalimu wetu aliuliza, "Kikombe gani kinaweza kuwa changu?"

5. Taja aina tano za viwakilishi na utoe mfano kwa kila aina (al. 10)

...4/

(c) Taja nambo matatu yanayokera kidawa (al. 6)

(d) Taja ^Stashibi moja iliyotundika (al. 2)

(e) Taja methali moja kutoka kwa kifungu (al. 2)

(f) Eleza maana na msamiti huu (al. 2)

(i) Kabuli

(ii) Kuvizia

SEHENU B.

1. Tumia viunganishi ulivyopatiwa ili kuunganisha sentensi hizi (al. 6)

(i) Asitoke mwanafunzi huu darasani. Sharti apate ruhusa ya mwalimu (Minghairi ya)

(ii) Nilimwona mwezi uliopita. Sikamwona tena (Token)

(iii) Anajua kuwa kusena uwongo ni vibaya. Anafanya tu (bali)

...4...

6. Kamusha sentensi hizi. (al. 6)

(i) Hii ni aina ya Insha ambapo unepewa mwelekezo

(ii) Mwenye nguvu mpisho.

(iii) Kwa hiyo lugha ya kisingerezo ni lugha ya walio wengi.

7. Tumia maneno haya katika sentensi ili kutofautisha maana. (Al. 4)

(i) Afyu

(ii) Avya

8. Akifisha kifungu hiki (al.5)

Hapana ni kifaduro skibisha mwingine La kifaduro ni ugonjwa wa watoto alisema mwingine.

9. Tumia vihisishi hivi katika sentensi (al. 3)

(i) Oh! Jamani!

(ii) Lo!

(iii) Ala!

APPENDIX G

SCORE SHEET AND RELIABILITY ANALYSIS FOR TEST SAMPLE 2

adm no	q1/2	q2/2	q3/2	q4/6	q5/2	q6/6	q7/62	q8/10	q9/2	q10/2	q11/2	q12/2	q13/2	q14/5	q15/3	total
1860	0	2	2	0	0	4	0	10	2	2	2	0	0	4	3	31
2007	0	0	0	0	0	5	0	10	2	0	0	1	0	5	3	26
2008	0	2	2	0	0	3	0	2	0	2	0	2	0	0	1	14
2010	0	2	2	0	0	4	0	10	2	2	2	2	0	4	3	33
2015	0	0	1	0	0	3	0	9	0	0	2	2	0	0	3	20
2016	0	1	1	0	0	0	0	6	0	2	2	2	0	0	0	14
2017	0	0	2	0	0	4	0	4	2	0	0	2	0	5	2	21
2025	0	2	1	0	0	3	0	8	2	2	2	2	0	1	2	25
2026	2	2	2	0	2	4	0	4	2	2	2	2	1	0	2	25
2028	0	0	2	2	0	3	0	10	0	2	0	1	0	4	3	27
2031	0	0	2	0	0	3	0	8	2	2	2	2	0	4	1	26
2032	0	0	2	0	0	4	0	9	2	2	2	2	0	4	3	30
2037	0	0	2	0	0	3	0	10	1	1	0	1	0	0	2	20
2039	0	0	0	2	0	4	0	4	1	2	2	2	0	5	3	25
2041	0	0	2	2	2	6	0	10	2	2	2	2	0	0	3	33
2042	0	2	2	0	2	6	0	10	0	2	2	0	0	5	3	34
2043	0	0	2	0	2	6	0	10	0	2	2	2	0	4	2	32
2046	2	2	2	0	2	5	0	6	0	2	2	1	2	4	0	30
2054	0	0	2	6	2	6	0	8	2	0	0	1	0	4	3	34
2055	0	0	2	0	0	3	0	8	0	2	2	2	0	4	3	26
2056	2	0	0	0	0	4	0	6	0	2	0	0	0	1	2	17
2065	0	0	2	2	2	6	0	10	2	2	0	2	0	4	3	35
2071	0	0	2	0	0	1	0	10	1	0	2	2	0	4	2	24
2072	0	2	2	0	2	6	0	6	2	0	0	2	0	0	3	25
2074	0	2	2	0	0	0	0	6	0	0	2	0	0	0	3	15
2080	0	2	2	0	0	4	0	8	0	2	0	2	0	0	3	23
2085	0	0	2	0	2	2	0	5	0	2	2	2	0	5	0	22
2087	2	2	2	0	0	0	0	3	0	0	2	2	0	0	3	16
2088	0	0	1	0	0	0	0	5	0	2	2	2	0	0	2	14
2090	0	0	2	0	2	4	0	7	2	2	2	2	0	0	3	26
2093	0	0	2	0	2	6	0	10	0	2	2	2	0	4	3	33
2095	2	2	2	0	2	2	0	8	0	2	2	2	0	2	3	29
2101	0	2	2	0	2	2	0	4	0	0	2	2	0	4	3	23
2102	0	2	2	0	2	6	0	8	2	2	2	2	0	4	3	35
2204	0	0	2	0	2	6	0	9	0	2	2	2	0	0	3	28
2207	0	1	2	6	0	0	0	9	2	2	2	2	0	0	3	29
2211	2	2	2	2	2	4	0	10	2	0	2	2	0	3	3	36

RELIABILITY ANALYSIS - SCALE (ALPH

A)
Reliability Coefficients

N of Cases = 37.0

N of Items = 15

Alpha = .4885

APPENDIX H
SAMPLE TEST 3

SHULE YA UPILI YA JOMO KENYATTA
MTIHANI WA MWISHO WA MUHULA
MUHULA WA PILI
KIDATO CHA 2
MUDA: SAA 1½

MALGIZO: JIBU MASWALI YOTE

JINA.....NAMBARI.....DARASA.....

A. UFAHAMU

1. Soma barua ifuatayo kisha ujibu maswali yatakatofuata.

S.L.P. 20,
MUGUNI
16.6.2004,

Sogora Msumeno Mwanangu,

Asalaam aleikum. Usiulize mimi ni nani, wala kustaaabu nimelipata vipi jina lako. Kila kitu kitakuwia wazi kutokana na maelezo yangu.

Watoto hao maskini wa Mungu, wana hadithi ndefu, ngumu, na ya kuhuzunisha sana, ambayo wenyewe hawajui, na singeweza kuwaeleza waje wakueleze na weno hadi uelwe, kwa sababu bado hawajaipata vizuri angalau lugha ya kujieleza, sigusii namna ya kueleza. Barua hii itawasaidia kuelezea matatizo yao.

Natanguliza kukusihii, ujikaze kiume kwa maelezo yangu haya. Nina maneno ambayo ni lazima nikutahadharishe kwa kuwa si madogo. Ni makubwa, tena makubwa sana basi - kwa ufupi, yale yanayostahili kuitwa majambo.

Niwie radhi, mwanangu, kuwa nimelazimika kukufichulia siri ya siri hatimaye, na kwa kufanya hivyo, bila shaka kuiperushaa mbali sana furaha yako. Kunradhi, hili hata kidogo silo lengo langu. Nimelazimika tu kuyaleta kwako mambo haya, mwanangu, kwa sababu sikuwa na njia nyingine. Maji yamewagika, na yakimwagika yamewagika, hayazoleki. Lakini kabla ya kuingilia maneno yenyewe, wewe na mimi hebu tuafikiane na huu usemo wa watu kuambiwa wakue, waje wayaone. Ya kuonekana siyo maembe au nini, mwanangu. Ni malimwengu ulimwenguni.

Ni hivi, mwanangu, nilikuwa safarini miaka kadha iliyopita. Mara, kufumbaa na kufumbaa, nikajikuta nimesimamishwa na sauti za watoto waliokuwa wakipokezana kulia. Nilipojihakikishia kuwa kweli walikuwa watoto hao, siyo majina au vichumbakazi, hikajiohoka kichochoroni, sauti zao zilipokuwa zikihanikisa kilio, Na, marahaba! kama nilivyobahatisha na kutarajia, nilikuta vitoto viwili vikiijililia maskini madanga ya wana, malaika wa Mungu vilikuwa vikitupatupa viguu na vikono vya huko na huko hewani kama vile makinda ya ndege yahisivyo kuwa mama yao yu karibu.

Mimi si mtu wa kuamini mashetani, mazingaombwe, ushirikina, wala uehavi kwa jumla. Hapo hapo, na kwa njia isiyoelweka, nikajiwa na huruma ya ajabu, iliyoniletea hata ubaridi mwilini. Na kwa wakati huohuo, nikapigwa na wazo la ghafla kuwa ilikuwa kazi ya shetani hii, hasa labda kwa sababu nauelewa sana ukatili wa wanawake siku hizi tunazoiita za maendeleo.

Mimi sikujaliwa kupata mtoto maishani mwangu kwa hivyo nilishawishika mara moja hapo, kuwa Mungu amenipelekea wito ambao nitajua kwa juujuu tu maana ya kulea.

Papo hapo kichochoroni, kulikuwapo vilevile na dawa fulani kiohupani, ilani ya kuzaliwa hospitalini kwa watoto hao na shilingi ishirini. Ziliwafaa wenyewe baadaye pesa hizo.

Niliwabeba mimi mwenyewe hadi nyumbani kwangu, watoto hawa. Huko nyumbani nampa shukrani nyingi mke wangu, kwa kazi kubwa aliyofanya ya kuwashughulikia kwa hali na hata kwa mali watoto hawa, mradi tu wakue waje wayaone ya kuyaona na wao.

Mbali na kazi ngumu sana ya kuona kuwa walikuwa wakikua kwa njia ya kawaida, kama vile watoto wengine mikononi mwa wazazi wao, nilijipa vilevile na jukumu la kukifanyia uchunguzi kitendo cha ukatili na cha unyama walichotendewa watoto hawa. Kwa usaidizi wa ilani yao ya kuzaliwa kwao, haikuwa kazi ngumu sana kwangu kufumbua fumbo ambalo sana hufumbiwa wajinga. Na ningeshawishika kuyafuata kisheria mambo haya, mwanamke katili, mama ya huyo, angegutukia kutiwa nguvuni; na bila shaka kufikia sasa angekuwa tayari ashajifunza maana ya utubora na kuheshimu maisha, ambayo kila kiumbe chayahathamini.

Kwa ufupi na kwa uwazi, kutokana na uchunguzi wangu, ni hivi. Mama watoto hawa alikata shauri kuwatupilia mbali wana hawa kwa sababu hakuwepata kwa njia halali ya kutungwa mimba na mumewe wa ndoa. Niligundua vilevile kuwa wewe ulikuwa uko ng'ambo nyakati hizo. Hivyo malimwengu yenyewe ndiyo haya mwanangu: kuwa mama watoto hao, kama inavyothibitisha ilani ya watoto kuzaliwa, ndiye huyohuyo mkeo mliyeapiana, bila shaka, kuwa mtaishi pamoja kwa mama na mabaya. Hivi ni kusema kuwa watoto wenyewe ambao kwa upande wangu nishawitimizia wajibu wa kuwalea mpaka sasa ambapo, ijapokuwa hawajakuwa watu, ndio hao wawili waliokujia na barua. Mmoja tulimpa jina Salome, na mwingine tukamwita Roda. Ombi langu kubwa zaidi kwako ni kwamba uahurumie na kuwaonea imani watoto hawa, ambao kwa vyovyote vile, hawana hatia yoyote. Ni majaliwa tu alelewe yalivyowaendea kinyuma, yakawacheza shere, na kuwatia doa hata kabla hawajazaliwa waja jhawa ambao kama waja wengine wenzao, hawana hiari.

Iwapo kutakuwa na haja ya kunitaka mazungumzo zaidi, utanipata Mpeketoni mjini, karibu na sokoni. Muulize mtu yeyote hapo, atakuonyesha nyumba yake ambamo wameishi nami watoto hawa tangu nilipowaokota.

Wasalaam,

Saidi Mwana Msondongoma

- (a) Msondongoma ana uhusiano gani na Sogoro Msumeno? (alama 1)
- (b) Madhumuni ya kuandika barua hii ni nini? (alama 1)
- (c) Ilani iliwafaa je watoto hawa baadaye? (alama 1)
- (d) Kwa nini Msondongoma aliamua kuwachukua watoto hawa? (alama 2)

- (f) Kitendo cha kutupa watoto wachanga ni hadithi tu au hutukia kweli? Ikiwa ni kweli, unafikiri kina mama wanaofanya hivyo huwa na sababu gani? (alama 2)

2(a) Eleza maana ya maneno mafuatayo kama yalivyotumika katika taarifa. (alama 4)

- (i) Kuwafikiana
- (ii) Malinwengu
- (iii) Hinikiza
- (iv) Malaika
- (v) Ushirikina
- (vi) Jukumu
- (vii) Makinda
- (viii) Majaliwa

(b) Mtu anapoomba msamaha husema kunradhi au: (alama 2)

- (i) radhi
- (ii) radhi.

(c) Eleza maana ya semi hizi kama zilivyotumika. (alama 5)

- (i) Jikaze kiume
- (ii) Kufumba na kufumbua
- (iii) Kutiwa nguvuni
- (iv) Kucheza shere
- (v) Kutia doa

(d) Kamilisha methali zifuatazo. (alama 2)

- (i) Maji yakimwagika
- (ii) Fumbo hufumbiwa mjinga

B. MATUMIZI YA LUGHA

3 (a) Onyesha vivumishi katika sentensi zifuatazo:

- (i) Mpira huu ni wa nani?
- (ii) Wanafunzi hawa na Walimu wale watatembelea miji hiyo.
- (iii) Matunda hayo ni machungu ajabu

(alama 5)

(b) Akifisha sentensi zifuatazo.

- (i) askari aliponiona alisema mikono juu mara moja (alama 2)
- (ii) Kisumu nairobi na mombasa ni miji mikuu ya Kenya. (alama 2)

(c) Andika sentensi zifuatazo katika hali ya Kiyakinisha. (alama 5)

- (i) Mama hapiki chakula kitamu.
- (ii) Mimi siandiki barua ndefu.
- (iii) Wandera havui samaki.
- (iv) Golikipa hashiki mpira.
- (v) Rais hanzizi kuhutubu.

(d) Chagua nomino mwafaka kukamilisha sentensi zifuatazo:

- (i) (upendo, Manukoto) yamekwisha.
- (ii) (Maji, Mkuki) yatachotwa na nani?
- (iii) (Chuo, Mlango) utafungwa.
- (iv) (Kuku, Shamba) Litanunuliwa.
- (v) (chakula, Kisima) kitapikwa leo jioni

(alama 5)

(e) Uunda Nomino kutokana na vitenzi vifuatazo:

<u>kilenzi</u>	<u>Nomino</u>
(i) Potea	_____
(ii) Kinywa	_____
(iii) Kosa	_____
(iv) Soma	_____
(v) Iba	_____

(alama 5)

4 (a) Andika sentensi zifuatazo kwa wingi:

- (i) Maji mengine yamemwagika.
- (ii) Mkoba mwingine umepotea.
- (iii) Mtu mwingine amekamatwa na polisi.
- (iv) Ugonjwa mwingine umezuka.
- (v) Chama kingine kimeanzishwa

(alama 5)

(b) Tunga sentenzi mbili kwa kila jozi ya maneno yafuatayo ili kutofautisha maana yake:

- (i) Saba.
Shaba.
- (ii) Kisasi
kizazi

(alama 4)

(c) Chagua neno lililo mwafaka katika mabano.

- (i) Shairi la majibizano huitwa (ngonjera, gojera)
- (ii) Walimu (waligoma, walingoma) mwaki uliopita.
- (iii) Mbona umehamia (bali, mbali) hivyo?
- (iv) (Nitakichambua/nitakichabua) Kitabu cha Walelisi.
- (v) Mama huyu (anawabagua, anawapakua) watoto wake

(alama 5)

(d) Kamilisha majina haya ya maundi. (alama 2)

(i) Thuraa ya

(ii) Bumba la

C. FASHI SIMULIZI

5 (a) Tegua vitendawili vifuatavyo.

(i) Sijui zendako wala atokako

(ii) Fatuma mohafu

(alama 2)

(b) Andika methali tatu za Kiswahili zinezotumiwa kuonya.

(i)

(ii)

(iii)

(alama 3)

(c) Shairi linaweza kuwa na vipande vinne. Kipande cha kwanza ni UKWAPI cha pili ni UTAO, cha tatu ni na cha nne

(alama 2)

(d) i) Eleza kwa kifupi tofauti kati ya Kisasili na mighani. (alama 2)

ii) Taja mfano mmoja wa kisasili. (alama 1)

HERI NA FANAKA

APPENDIX I
SCORE SHEET AND ITEM ANALYSIS FOR SAMPLE TEST 3

UPPER GROUP																			
Adm No	q1/5	q2/4	q3/2	q4/5	q5/2	q6/5	q7/4	q8/5	q9/5	q10/5	q11/5	q12/4	q13/5	q14/2	q15/2	q16/3	q17/2	q18/3	total
9013	5	2	1	3	1	4	4	5	5	4	5	4	5	1	2	3	0	3	57
9017	4	3	2	4	1	5	4	5	5	4	5	4	5	1	1	2	0	2	57
9043	4	4	2	3	1	5	4	5	5	3	5	2	5	1	1	3	0	3	56
9100	4	2	2	4	1	5	4	5	5	4	5	2	5	1	1	3	0	3	56
8955	3	3	1	3	1	5	4	5	5	4	4	4	5	2	1	2	0	3	55
8948	2	1	2	3	1	5	4	5	5	5	5	4	5	1	1	3	0	3	55
8931	3	3	2	5	1	5	4	4	5	5	5	3	5	1	2	2	0	0	55
9038	2	3	2	5	1	5	4	5	5	3	5	3	5	1	2	3	0	1	55
8886	4	3	1	3	1	5	4	5	5	4	3	4	4	1	1	3	0	2	53
8943	3	2	2	3	1	5	4	4	5	4	5	3	4	0	1	3	0	3	52
8959	5	1	2	1	1	4	4	4	5	3	4	4	5	1	2	3	0	2	51
9012	5	3	2	2	1	5	4	5	5	3	4	4	5	1	0	2	0	0	51
9050	3	2	1	2	1	5	4	5	5	4	5	2	5	1	2	3	0	0	50
8904	3	1	1	2	1	5	3	4	5	5	5	4	5	0	1	2	0	2	49
total	50	33	23	43	14	68	55	66	70	55	65	47	68	13	18	37	0	27	752
LOWER GROUP																			
Adm No	q1/5	q2/4	q3/2	q4/5	q5/2	q6/5	q7/4	q8/5	q9/5	q10/5	q11/5	q12/4	q13/5	q14/2	q15/2	q16/3	q17/2	q18/3	total
9129	4	1	1	0	1	5	4	5	5	4	5	1	5	0	1	3	0	3	48
9126	2	2	1	2	1	5	4	4	5	4	5	4	5	1	1	2	0	0	48
8889	2	1	2	3	1	5	4	3	5	5	5	1	5	1	2	1	0	1	47
8966	3	2	2	2	1	5	4	5	5	2	5	2	4	1	1	2	0	1	47
9115	5	1	1	3	1	5	3	0	5	3	5	2	5	1	1	3	0	2	46
9063	5	2	2	4	1	5	2	4	5	3	3	2	5	0	1	2	0	0	46
8934	3	2	1	2	1	5	2	5	5	4	4	4	5	0	0	3	0	0	46
9124	3	1	2	3	1	5	3	5	5	2	4	3	5	0	1	3	0	0	46
8985	4	2	2	3	1	5	1	5	5	0	5	3	5	0	1	3	0	0	45
8941	3	2	2	3	1	5	4	1	5	4	4	2	5	1	0	2	0	0	44
9059	2	2	1	2	1	5	3	4	5	3	5	3	3	0	1	2	0	2	44
8909	5	1	1	3	1	5	4	4	5	0	4	1	4	1	1	2	0	0	42
9102	5	2	0	1	1	5	4	0	5	4	4	1	5	1	1	2	0	0	41
8936	5	1	1	3	1	5	3	4	5	0	4	2	3	1	0	3	0	0	41
Total	51	22	23	34	14	70	45	49	70	38	62	31	64	8	12	33	0	9	
df	-1	11	0	9	0	-2	10	17	0	17	3	16	4	5	6	4	0	18	
D-value	-0.01	0.16	0.00	0.13	0.00	-0.03	0.10	0.24	0.00	0.24	0.04	0.23	0.06	0.07	0.09	0.06	0.00	0.26	

APPENDIX J
RESEARCH PERMIT

MINISTRY OF EDUCATION, SCIENCE AND TECHNOLOGY

Telegrams: "EDUCATION", Nairobi
Telephone: Nairobi 334411
When replying please quote

Ref. No **MOEST-13/001/34C 341/2**
and date



JOGOO HOUSE "B"
HARAMBEE AVENUE
P.O. Box 30040-00100
NAIROBI

30th September, 2004, 20.....

David Macharia
Egerton University
P.O. BOX 536
NJORO

Dear Sir

RE: RESEARCH AUTHORISATION

Please refer to your application for authority to conduct research on "Quality of Kiswahili Language Teacher made Tests: A case study of Bahati Division". I am pleased to inform you that you have been authorised to carry out research in Bahati Division in Nakuru District for a period ending 30th September, 2005.

You are advised to report to the District Commissioner and the District Education Officer Nakuru District Education Officer Nakuru District before commencing your study. It is noted that the research is a requirement in part fulfillment for the award of M.Ed Degree by Egerton University. Upon completion of your research project, you are expected to submit two copies of your research report to this Office.

Yours faithfully


T. MOTURI
FOR: PERMANENT SECRETARY

Cc
The District Commissioner
Nakuru District

The District Education Officer
Nakuru District

(d) Kamilisha majina haya ya kawaida. (alama 2)

(i) Thurca ya

(ii) Bumba la

C. FASIMI SIMULIZI

5 (a) Tegua vitendawili vifuatavyo.

(i) Sijui zendako wala atokako

(ii) Fatuma mohafu

(alama 2)

(b) Andika methali tatu za Kiswahili zinezotumiwa kuonya.

(i)

(ii)

(iii)

(alama 3)

(c) Shairi linaweza kuwa na vipande vinne. Kipande cha kwanza ni UKWAPI cha pili ni UTAO, cha tatu ni na cha nne

(alama 2)

(d) i) Eleza kwa kifupi tofauti kati ya Kiasili na mighani. (alama 2)

ii) Taja mfano mmoja wa kiasili. (alama 1)

HERI NA FANAKA