

**VALIDITY AND RELIABILITY OF TEACHER MADE
TESTS IN KENYA: A CASE STUDY OF
PHYSICS FORM THREE IN
NYAHURURU DISTRICT.**

KINYUA, DANIEL KIRAGU.

**UNIVERSITY OF NAIROBI
NYAHURURU LIBRARY**

**UNIVERSITY OF NAIROBI
SCHOOL OF EDUCATION
DEPARTMENT OF PSYCHOLOGY**

©

2012

A RESEARCH PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENT FOR THE DEGREE OF MASTER OF EDUCATION IN THE
DEPARTMENT OF EDUCATION PSYCHOLOGY,

UNIVERSITY OF NAIROBI
2012.

© Copyright 2012

All rights reserved. No part of this work may be reproduced, stored in a retrieval system or transmitted in any form or means electronic, photocopying, recording or otherwise without prior written permission of the author or the University of Nairobi.

Declaration

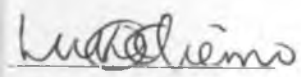
This project report is my original work and has not been submitted for approval in any other university



Kinyua Daniel Kiragu.
(E58/62315/2010)

Date 23 / 11 /2012

This project report has been submitted with my approval as the university supervisor.



Dr Luke Odiemo
Department of psychology
University of Nairobi

Date 22 / 11 /2012

DEDICATION

This project report is a special dedication to my Father in Heaven and my whole family, especially my wife Tabitha Wangari. You all gave me time and resources to work on my study

I would like to also dedicate this research project report to the University of Nairobi, Department of psychology especially the coordinator of the course Measurement and Evaluation Dr. Karen Odhiambo for her immeasurable support in the course of study and all the lecturers who took me through the various courses to fulfill the requirements of the school.

ACKNOWLEDGEMENT

I would like to acknowledge the sincere support and guidance that has been offered by my supervisor, Dr Luke Odiemo, for guiding and giving all the necessary advice and help which lead to the successful production of this project. You are a hero for this.

Special acknowledgment to all the respondents who willingly filled the questionnaires and the key respondents. My gratitude to the teachers who helped me in analysis of the exams to determine their validity. Your participation is greatly appreciated. My gratitude goes to the Head teacher of Thiru secondary school for willingly giving me permission and to all the teachers who were willing to stand in for me whenever I was out.

Special gratitude to my brother George Mwaniki who typed and printed the work. You have been of great help to me. I cannot forget the entire Body of Christ Fellowship, Mwariki. Your help was enormous. To all who participated in this project, God bless you all abundantly

ABSTRACT

This study was carried out to determine the factors affecting validity and reliability. Validity is the quality of a test to measure what it purports to measure. Reliability is the property of replicability of the results of a test. These are the two main and important properties of a test.

The study specifically sought the effect of each of the factors examined on validity and reliability levels of the different tests. It was conducted in Nyahururu District of Laikipia County, Kenya

The study involved 42 teachers and 15 key informants in the various schools in this district.

The study was guided by the following questions: 1) Does experience affect validity and reliability of teacher made tests?; 2) Does training on test construction and analysis affect reliability and validity?; 3) Does level of education affect validity and reliability of teacher made tests?; 4) Does use of Bloom's taxonomy affect validity and reliability?; 5) Does moderation affect the validity and reliability of tests?; 6) Does length of tests affect reliability and validity of teacher made tests?

This study used a mixed descriptive survey research design to collect and analyze the data. Data was collected through questionnaires and interviews with the key informants i.e. directors of studies or deputy principals. All these were applied to elicit opinion of all teachers and directors of studies. The information from the questionnaires was presented in figures and percentages in tables while the information obtained from interviews were analyzed using qualitative techniques.

The findings of the study revealed that the experience of teachers, training on test construction and analysis, level of education, use of Bloom's taxonomy, moderation of tests and length of tests have an effect on the validity and reliability of the tests. The findings also revealed that each of the factors have a varying level of significance on the validity and reliability of the tests.

The study concluded those teachers made tests are valid and reliable. The various factors come into play in determination of the level of validity and reliability. There is a lot that needs to be done especially in teacher training in order to improve on the test constructed by the teachers.

TABLE OF CONTENT

Page

Declaration-----	iii
Dedication-----	iv
Acknowledgement-----	v
Abstract-----	vi
Table of content-----	vii
List of tables-----	xiii
List of figures-----	xiv

CHAPTER ONE: INTRODUCTION

1.0. Background information -----	1
1.1. Statement of research problem-----	7
1.2. Objectives of the study-----	9
1.3. Research questions-----	9
1.4. Justification of proposed research-----	9
1.5. Scope and limitations-----	11
1.6. Purpose of the study-----	11

CHAPTER TWO: LITERATURE REVIEW

2.0. Introduction-----	12
2.1 Context-----	12
2.2. Theoretical framework -----	12

- 2.3. Factors affecting validity and reliability of
 - teacher made tests-----14
 - 2.3.1. Training test construction-----14
 - 2.3.2. Number of items-----16
 - 2.3.3. Use of table of specification (bloom’s taxonomy) -----17
- 2.4. Effect of valid and reliable formative tests
 - on performance in summative assessment-----19
- 2.5. Validity criteria-----23
- 2.6. Hypothesis-----26
- 2.7. Conceptual framework-----26

CHAPTER THREE: METHODOLOGY

- 3.1. Introduction-----28
- 3.2. Research design-----28
- 3.3. Target population-----29
- 3.4. Sample selection and sample size-----29
- 3.5. Research instruments-----29
 - 3.5.1. Validity of research instruments-----30
 - 3.5.2. Reliability of research instruments-----30
- 3.6. Data collection procedures-----30
- 3.7. Data analysis-----30

3.7. Ethical considerations-----33

CHAPTER FOUR: DATA ANALYSIS, PRESENTATION AND DISCUSSION

4.1. Introduction-----34

4.2. Response return rate-----34

4.3. Demographic characteristics of respondents-----35

4.3.1. Distribution of respondents by age-----35

4.3.2. Distribution of respondents by gender-----36

4.3.3. Distribution of respondents by location from school-----37

4.3.4. Distribution of respondents by their relationship
with other staff-----37

4.4. Factors affecting validity and reliability-----38

4.4.1. Teachers experience and the effect on validity
and validity-----39

4.4.2. Education level and effect on validity and reliability-----43

4.4.3. Training on test construction and analysis and effect
on validity and reliability-----47

4.4.4. Use of bloom's taxonomy and effect on validity and
reliability-----53

4.4.5. Moderation of tests and effect on validity and
reliability-----56

4.4.6. Length of tests and effect on validity and reliability-----60

4.5. Conclusion-----64

**CHAPTER FIVE: SUMMARY OF FINDINGS, CONCLUSIONS AND
RECOMMENDATIONS**

5.1. Introduction-----65

5.2. Methodology-----65

5.3. Justification of methodology used-----66

5.4. Reliability of the tests-----67

5.5. Validity of the tests-----67

5.5. Findings-----67

5.6.1. Does teachers experience affect validity and
reliability of their tests-----67

5.6.2. Does the level of education affect the validity and reliability of tests-----68

5.6.3. Does training on test construction and analysis effect validity and
reliability of tests-----68

5.6.4. Does moderation of tests affect their validity and reliability-----68

5.6.5. Does use of Bloom’s taxonomy affect the validity and reliability of tests-----69

5.6.6. Does the length of tests affect the validity and reliability of tests-----69

5.7. Contributions to the body of knowledge-----69

5.8. Recommendations-----70

UNIVERSITY OF NAIROBI
LIBRARY

5.9. Practical use of the findings-----	71
5.10. Suggestions for further research-----	71
References-----	72
Appendices	
Appendix 1. Researcher’s introductory letter-----	74
Appendix 2. Questionnaire for teachers in Nyahururu district-----	75
Appendix 3. Questions for the key informant-----	80
Appendix 4. Determination of sample size for research activities-----	81

LIST OF TABLES

Table 1 Questionnaire response rate-----	34
Table 2. Age of respondents in years-----	35
Table 3 Distribution of respondents by gender-----	36
Table 4 Distribution of respondents by location from school-----	37
Table 5 Distribution of respondents by their relationship with other members of staff-----	38
Table 6 Teachers experience and corresponding measures of validity and reliability-----	39
Table 7 Distribution of respondents on experience -----	40
Table 8 Education level versus levels of reliability and validity-----	43
Table 9 Distribution of respondents by their education level-----	44
Table 10 Training on test construction and corresponding values of reliability and validity-----	47
Table 11 Summary of the results on training on test construction and analysis-----	48
Table 12 Distribution of respondents on need for further training on test construction And analysis-----	50
Table 13 Distribution of respondents in terms of use of Bloom's taxonomy and validity -----	53
Table 14 Usage of Bloom's taxonomy and corresponding values of reliability-----	54
Table 15 Moderation of tests and corresponding values of validity and reliability-----	57
Table 16 Distribution of respondents in terms of exam moderation-----	58
Table 17. The more the number of items in an exam the higher the level of reliability and also validity-----	61

LIST OF GRAPHS

Graph 1. Summary of the number of years one has been teaching-----	41
Graph 2. Summary of education level of respondents-----	45
Graph 3. Summary of adequacy of training on test construction and analysis-----	49
Graph 4. Summary of need for further training on test Construction-----	51
Graph 5. Summary of the usage of Bloom's taxonomy in preparing exams-----	55
Graph 6. Distribution of respondents in terms of exam moderation-----	59
Graph 7. Summary of length of tests-----	63

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND INFORMATION

The problem in this study is that the validity and reliability of teacher made tests is not known in Kenya. A good teacher made test should be valid and reliable and if these qualities are not known then it becomes almost impossible and even deceptive to use teacher made tests for evaluation purposes.

Measurement experts and many educators believe that every measurement devices should possess certain qualities. Perhaps the two most common technical concepts in measurement and evaluation are validity and reliability. Any kind of assessment, whether traditional or authentic, must be developed in a way that gives the assessor accurate information about the performance of the individual.

Validity is the quality of a test which measures what the test is supposed to measure. Validity refers to the degree to which the evidence support that these interpretations are used is appropriate (American Psychological Organization, 1999). It is the accuracy of truthfulness of measurement. It is the degree to which common sense, or theory supports any interpretations or conclusions about a student based on his or her test performance (Hunter and Schmidt F. L,1999). Simply is how one knows that a Mathematics test measures student's mathematical ability not their reading skills. There are various types of validity i.e. Face validity, Content validity, Criterion related validity, and Construct validity.

Reliability refers to the consistency of the scores obtained. That is, how consistent the scores are for each individual from one administration of an instrument to another and from one item to

another. Reliability is a measure of how stable, dependable, trustworthy and consistent a test is in measuring the same thing each time (Worthen et al, 1993). For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessments, when the response was scored and who scored the response. On the other hand, on an unreliable test, a student's score may vary based on factors that are not related to the purpose of assessment (James H. McMillan, Virginia Commonwealth University, 2000). There are many types of reliability which include Test-retest method, Parallel form reliability, Split-half method, inter-rater reliability, and internal consistency.

The process of assessment and evaluation is always part and parcel in the teaching and learning process in every educational institution whether in basic education or in tertiary education. The use of teacher made tests in evaluation in all institutions of learning ranging from Early Childhood Education up to University level is a rule rather than the exception. Teacher made tests are written or oral assessments. They are not commercially produced or standardized. It is a test the teacher designs specifically for his or her students. Teacher made tests can be classified as aptitude, progress, achievement and proficiency tests, (Valette, 1977). To determine how much learning has occurred teachers can, for example, have students take exams, respond to oral questions, do assignment exercises, write papers, solve problems, and make oral presentations.

Tests can be important parts of the teaching and learning process if they are integrated into daily classroom teaching and are constructed to be part of the learning process-not just the culminating event. They allow students to see their own progress and allow teachers to make adjustments to their instruction on daily basis. "But one of the most serious problems of evaluation is the fact that a primary means of assessment, the test, itself is often severely flawed or misused (Hills,

1991. Pg541). "While large-scale standardized tests may appear to have great influence at specific times without question, teachers are the drivers of the assessment systems that determine the effectiveness of schools" (Stiggins, 1994).

According to a study carried out in the Seychelles which is a developing country like Kenya it revealed that the overall quality of the teacher made tests is far from ideal. The study used a sample of 23 teachers of grade 5A in Primary School. A sample of mathematics tests constructed by Primary 5A teachers was analyzed. It concluded that teachers should be cautious in using the results of these tests to make decisions about the education of the students under their care. In-service teachers need assistance in test construction and other measurement areas so that they eventually lessen their dependence on tests to measure students' performance. They should be encouraged to develop and implement other assessment instruments. It would be advisable to include a measurement course in the pre-service teacher training program.

There are, however, a number of limitations inherent in the study. First, the sample of tests analysed was small. Secondly, the study considered only mathematics tests at Primary 5A level. Information about the quality of tests across other year levels and in other subjects' areas is unknown. Moreover, the study has examined only one test constructed by each of these teachers. It cannot be claimed that the tests they constructed always resembles the picture received from this study.

Another limitation of this study concerns the procedures used to assess the content-related validity of the tests. The processes were somewhat subjective. If others were to judge the content-related validity evidence of the tests, it is possible that they would arrive at a different point of view. This decreases the reliability of the study. In spite of these weaknesses, the results

clearly indicate that there exists a problem about test construction when teachers are not trained to construct tests and this affects their reliability and validity.

Another study by Anita M. Parr in Ohio, U.S.A showed that there is no significant difference in achievement between a test taken on a Friday and a test taken on the following Monday. This researcher attempted to diminish the threat of test-retest reliability but was not able to eliminate it completely. The familiarity of the test format and the information tested may have altered the results by providing a higher mean score on the posttest. This can help to establish the validity and reliability of teacher-made tests in the biology classroom in Monroe Central High School.

Teacher-made tests are one of the major contributors to the overall grade point averages of students at Monroe Central High School. Establishing validity for teacher-made tests can help to establish the validity of student's grade point averages as an indicator for achievement. The researcher found that establishing the validity of teacher-made tests will improve the accountability of education and add credibility to Grade Point Averages as a measure of student achievement. The problem with this study is that it only investigates only test-retest reliability and does not report anything about the tests validity yet we know that a reliable instrument needs to be a valid instrument.

The length /number of items is a crucial factor of test reliability. Carefully written tests with an adequate number of items usually produce high reliability (Justin, D.V; John,R.G.,1996) since they usually provide a representative sample of the behavior being measured and the scores are apt to be less distorted by chance factors, for example, familiarity with a given item or misunderstanding of what is expected from an item (Linn & Gronlund,1995).

The training one has had on test construction also affects the validity and reliability of tests constructed by teachers. One study by Mayo (1967) found that graduating seniors in 86 teacher

training institutions did not demonstrate a very high level of measurement competence. Most teachers believe that they need strong measurement skills (Wise, Lukin & Roos, 1991). While some report that they are confident in their ability to produce valid and reliable tests (Oescher & Kirby, 1990; Wise, et al., 1991), others report a level of discomfort with the quality of their own tests (Stiggins & Bridgeford, 1985) or believe that their training was inadequate (Wise, et al.). Indeed, most state certification systems and half of all teacher education programs have no assessment course requirement or even an explicit requirement that teachers have received training in assessment (Boothroyd, et al.; Stiggins, 1991; Trice, 2000; Wise, et al.).

In a research to determine the profile of the professors' level of appropriateness in test construction in University of Perpetual Help Laguna, it was noted that the practical rules of test item construction is formulated on the basis of years of experience in preparing items and empirical evaluation of responses. Thus, the level of appropriateness in test construction correlated with the actual years of experience in teaching was investigated. There were 33 college professors who were surveyed in this study. The survey determined their tendency to follow general principles, guidelines and procedure in test construction. The results indicate that majority of the professors (54.54%) had an average level of appropriateness in test construction with a mean value of 24.76. The result also revealed that there is no significant relationship between the level of appropriateness in test construction and the actual years of teaching experience. This is due to the adequate training and exposure that new professors had in test construction. The level of appropriateness has an effect on the validity and reliability since the tests could give distorted information. The level of appropriateness is determined using Bloom's taxonomy. Thus the use of Bloom's table of specification in determining the items to include in a test has a great effect on the tests validity and reliability. However, this research used professors

from only one university and there would be a need to include others to avoid subjectivity on the results obtained. This would lead to a better generalization on these tests.

In Kenya just like other parts of the world the use of teacher made tests is the norm from kindergarten to universities. A study by Mwangi Ndirangu (2010), suggests that there is need to diversify final school assessment than relying on a single examination at the end of the learning cycle, to assess student learning. He suggests that all assessment contains some error or imprecision. Relying on the results of one examination may lead to incorrect decisions being made about a learner. Incorporating teacher made tests (TMTs) as part of the wider continuous assessment has been suggested as likely to provide a better framework for assessing student academic achievement and potential. This study sought to investigate whether teachers embraced the psychometric procedures to enhance validity of their tests. This would validate the test results inclusion into students' final school assessments. The study adopted a descriptive survey design and involved all trained secondary school Kiswahili teachers in Bahati Division, Nakuru North District, Kenya. Kiswahili is one of the subjects examined at Kenya Certificate of Secondary Education examination. All the trained Kiswahili teachers (76 in number) participated in this study. A questionnaire, an interview schedule and a checklist were the main instruments for data collection. Using simple random sampling, 20 heads of department were selected for an interview. The major findings of this study were that teachers did not use the established psychometric procedures in test construction. The resulting tests had a low content validity. Consequently, the results from such tests could not be effective indicators of learners' achievement. The findings are likely to be important to all stakeholders concerned with ways of improving student assessment.

Typically, teacher-based assessment is presented in the literature as having higher validity than external assessment (Justin D.V, John R.G, 1996). Similar assertions have been made by Anita M. Parr (2003) and Carlo Magno (1996). Due to its continuous nature, teacher-based assessment often allows for important achievements to be measured that could not be captured in a final examination, such as extended projects, practical assignments or oral work.

However, teacher-based assessments are often perceived as unreliable. These assertions are made by Ouster (2003), Prof . Mwangi Ndirangu (2010), and Frey et al (2005). Test items and grading standards may vary widely between teachers and schools, so that the results of internal assessment will lack external confidence and cannot be compared across schools. There might also be a high risk of bias in teacher-based assessment, *i.e.* the assessment is unfair to particular groups of students.

This indicates that a combination of teacher-based and external assessments would be most suitable to ensure maximum validity and reliability. Learning outcomes that can be readily assessed in external examination should be covered this way, whereas more complex competencies should be assessed through continuous teacher-based assessment. The validity and reliability of Physics in form three is not yet studied in Kenya. This is the essence of this study which seeks to establish the reliability and validity of these tests.

1.2 STATEMENT OF THE PROBLEM

In the recent past a lot of attention has been given to the validity and reliability of teacher made tests. Many researchers have examined the extent to which teacher made tests are valid and reliable. Also the factors that affect reliability and validity have been studied and analyzed. Means of improving the quality of teacher made tests have been well documented.

Apart from many studies having been carried out on teacher made tests, this has not yet been done in Nyahururu District of Laikipia County. Also the factors affecting the reliability and validity in other areas may be different from those of this area. Some studies have been carried out in other subjects apart from Physics. Therefore the factors affecting other subjects are not the same as those in Physics. The level of validity and reliability of teacher made tests in Physics is not yet known in Kenya.

High quality assessment system in education is an obvious need for all schools. A teacher spends most of his/her time in assessment. Time is allocated at the end of each portion and /or terms for examinations. Teachers augment these results by cumulative test during each term or semester (Peterson and Walberg, 1970). Since teacher made tests are an integral part of assessment then their reliability and validity needs to be established.

It is important to determine the level of validity and reliability because many researchers report that the level of teacher's questions has a direct relationship to the cognitive level that students have to employ to arrive at satisfactory responses to questions (Victor Y. Billeh, 1974). He continues to note that through proper tests (that are valid and reliable) teachers can have a significant influence on directing and developing the cognitive processes of their students, therefore knowledge on validity and reliability is indispensable.

Teachers place a lot of weight on their own tests in determining grades and student's progress than they do on assessments produced by others or from other sources (Boothroyd et al, 1992). Due to this it is important to determine the validity and reliability of these tests.

Good tests teach students about test-taking. If teachers give poorly constructed assessments students are not learning an important skill that will assist them in correctly interpreting and

answering tests. High quality tests that are valid and reliable teach students about test taking (Harvey Craft, 2011). If teachers give poorly constructed tests students are not learning an important skill that will assist them in correctly interpreting and answering tests.

1.3. OBJECTIVE OF THE STUDY

The objective of the study will be to determine the factors that affect the validity and reliability of teacher made tests.

1.4. RESEARCH QUESTIONS

1. Does experience affect validity and reliability of teacher made tests?
2. Does training on test construction and analysis affect reliability and validity?
3. Does level of education affect validity and reliability of teacher made tests?
4. Does use of Bloom's taxonomy affect validity and reliability?
5. Does moderation affect the validity and reliability of tests?
6. Does length of tests affect reliability and validity of teacher made tests?

1.5. JUSTIFICATION OF PROPOSED RESEARCH

In Kenya today, with the implementation of the new constitution, decentralization of Kenya national examinations council is inevitable. The assessment of learners' achievement through teacher made classroom tests in the course of learning is raising interest as time goes by. Thus

there is a need to establish the reliability and validity of these tests so that if this is the direction that education takes in the end then it is proved to be a reliable means of assessment.

Also with the coming in of the new constitution of Kenya, educational services will somehow be decentralized to the counties government including but not limited to assessment. Thus the determination of the quality of tests is important to ensure uniformity in assessment process throughout the country. This will prevent disparities which may emerge from different testing criteria between the counties.

The finding in the research project will also help the curriculum designers of a teacher education (colleges and universities) to adjust or put more emphasis on the process of testing. This will ensure that the teachers learning in colleges and other institutions will have been adequately prepared to be able to carry out assessment effectively. The study is also very important in determining the extent to which teachers should be trained on testing and also the effectiveness and adequacy of this training.

The study will also help to increase the confidence the community and education stakeholders have in teacher made classroom tests, whether they are end of term or continuous assessments. It is also important since it builds the morale of teachers to know that their exams are well recognized and not taken just because it is end of term or middle of the term or any other time. This, of course, would mean that the teachers will need a thorough training on testing but if it is worth it if their exams are to be used to evaluate overall students achievement and at the end of a course

1.6. SCOPE AND LIMITATIONS

The study will be carried out in secondary schools in Nyahururu district of Rift valley province in Kenya. The study will cover the physics subject in form three only in Nyahururu district. Though this is the case the findings could still be implied in other subjects and levels in the secondary schools. The quality of information collected under the study will so much depend on the quality, willingness and sincerity of the respondents. However the study questionnaire will include a comprehensive introductory letter to elicit clarity as to the use of any information that will be provided.

1.7. PURPOSE OF THE STUDY

The problem is that the validity and reliability of teacher made tests in Physics Form three in Kenya is not yet known. Yet the results of teacher made tests are used in making important decisions about the education of their students. This study will seek to establish the extent to which teacher made tests are reliable and valid in Physics subject, Form three.

UNIVERSITY OF NAIROBI
"MUNYI IIRIAD"

CHAPTER TWO

LITERATURE REVIEW

2.0 INTRODUCTION

This chapter presents a review of related literature of this study. It also presents the theoretical review particularly the behaviorist theory. It also highlights the conceptual framework, and the related literature on the factors which affect reliability and validity of teacher made tests.

2.1 CONTEXT

The Kenyan education system is eight years of primary school, four years in Primary school and four years of University education. There are exams at the end of primary school and at the end of form four. As such teachers do prepare exams at the end of every term as measures of student attainment at each level. These exams are many and they give a lot of information about the learners. Therefore there validity and reliability needs to be established such that the interpretations made from the results are accurate. This study seeks to establish the factors that affect reliability and validity and also determine the effect of these factors on levels of reliability and validity .The study was carried out in Nyahururu District of Laikipia county in Kenya. It has Thirty six secondary Schools and 45 Physics teachers. It is located in the former Rift Valley province within the boundaries of former Central province.

2.2 THEORETICAL FRAMEWORK

The goal of reliability theory is to estimate errors in measurement and to suggest ways of improving tests so that errors are minimized. The central assumption of reliability theory is that

measurement errors are essentially random. This does not mean that errors arise from random processes. For any individual, an error in measurement is not a completely random event. However, across a large number of individuals, the causes of measurement error are assumed to be so varied that measure errors act as random variable.

If errors have the essential characteristics of random variables, then it is reasonable to assume that errors are equally likely to be positive or negative, and that they are not correlated with true scores or with errors on other tests. The basic assumptions are that the mean error of measurement is zero, true scores and error scores are uncorrelated and errors on different measures are uncorrelated.

In many educational and psychological measurement situations, there is an underlying variable of interest. This variable is often something that is intuitively understood, such as "intelligence." When people are described as being bright or average, the listener has some idea as to what the speaker is conveying about the object of the discussion.

In academic areas, one can use descriptive terms such as reading ability and arithmetic ability. Each of these is what psychometricians refer to as an unobservable, or latent, trait. Although such a variable is easily described, and knowledgeable persons can list its attributes, it cannot be measured directly as can height or weight, for example, since the variable is a concept rather than a physical dimension. A primary goal of educational and psychological measurement is the determination of how much of such a latent trait a person or a test possesses.

Item response theory is a psychometric theory that seeks to model an unobserved, latent variable using observed responses. According to item response theory the latent trait is measured using attributes that are measurable. The idea behind item response theory is to model the respondent's

proficiency on the unobserved latent variable of interest given the same respondent's proficiency on a set of observed (and measured) responses, which are related to the unobserved latent variable. The item response theory is an important aspect of this study. The latent traits is validity and reliability which can be determined by use of attributes like the length of tests, use of Bloom's taxonomy and also the training of teachers on test construction.

2.3 FACTORS AFFECTING VALIDITY AND RELIABILITY OF TEACHER MADE TESTS

2.3.1. TRAINING ON TEST CONSTRUCTION.

Most teachers believe that they need strong measurement skills (Wise, Lukin & Roos, 1991). While some report that they are confident in their ability to produce valid and reliable tests (Oescher & Kirby, 1990; Wise, et al., 1991), others report a level of discomfort with the quality of their own tests (Stiggins & Bridgeford, 1985) or believe that their training was inadequate (Wise, et al.). Indeed, most state certification systems and half of all teacher education programs have no assessment course requirement or even an explicit requirement that teachers have received training in assessment (Boothroyd, et al.; Stiggins, 1991; Trice, 2000; Wise, et al.). In addition, teachers have historically received little or no training or support after certification (Herman & Dorr-Bremme, 1984). The formal assessment training teachers do receive often focuses on large-scale test administration and standardized test score interpretation rather than on the test construction strategies or item-writing rules that teachers need (Stiggins, 1991; Stiggins & Bridgeford, 1985). One study by Mayo (1967) found that graduating seniors in 86 teacher training institutions did not demonstrate a very high level of measurement competence.

A study on social consequences of testing and the development of talent by Goslin (1967) found that about 60% of all teachers had only minimal exposure to training in tests and measurement techniques. The unsatisfactory quality of the majority of teacher made tests no doubt reflects this inadequacy in training. Moreover, it was found that teachers who have had little preparation in tests and measurement tend to make little use of the information obtained from standardized tests.

According to a paper on reliability and validity of tests constructed by Seychelles teachers, Seychelles teachers are rarely trained on test construction. Yet, the use of teacher made tests for assessing students is a common occurrence in schools. The results indicated that the tests have high internal consistency reliability.

All the Cronbach alpha coefficients are above 0.7. The highest is 0.95 and the lowest is 0.73. The mean Cronbach coefficient alpha for Paper 1 results is 0.89 and that of Paper 2 results is 0.88. The mean Cronbach coefficient alpha for all test results is 0.89. This indicates that in general the test results are highly reliable. High internal reliability indicates that items of the test consistently measure the same ability.

However the tests are low in content related validity. The judges observed little evidence of content-related validity in the tests. On average they were giving an overall rating of 2.9 per test. The most valid test received an average overall rating of 4.5. Only one test received this rating.

The tests contain a low percentage of effective items as judged by traditional norm-referenced item analysis techniques. The study recommends that a measurement course be introduced in the

teacher training program. It also recommends that in-service teachers lessen their dependence on tests to measure students performance and that other assessment instruments be developed and implemented as well.

As this may be true in Seychelles, it may not be the same case in Kenya. This is because there is a measurement course in teacher training program already in place. Well, it may not be necessarily adequate and so improvement may be needed. This study will seek to establish how much teachers are trained on test construction and their feeling on its adequacy.

2.3.2: NUMBER OF ITEMS

One way teachers can construct better teacher-made is to consider the number of questions/items that should be included on a test. Obviously, it is important to select test items that will measure whether students have achieved the significant learning objectives, benchmarks, or standards that have been targeted.

The length /number of items is a crucial factor of test reliability. Carefully written tests with an adequate number of items usually produce high reliability (Justin, D.V; John,R.G.,1996) since they usually provide a representative sample of the behavior being measured and the scores are apt to be less distorted by chance factors, for example, familiarity with a given item or misunderstanding of what is expected from an item (Linn & Gronlund,1995).

According to a study by Meshkani, Z., PhD; Hossein Abadie F., MSc. increasing the number of items does not necessarily increase the reliability but it is an important consideration. They found out that Length of examination affects the reliability, but this study showed without the considering the quality of test items, increasing the number of questions in order to increase reliability is a big mistake.

It is clear that increasing the number of items does not necessarily increase the validity and reliability of the teacher made tests. Craig S. Wells and James A. Wollack conclude that longer tests produce higher reliabilities. This may be seen in the old carpenter's adage, "measure twice, and cut once." Intuitively, this also makes a great deal of sense. Most instructors would feel uncomfortable basing midterm grades on students' responses to a single multiple-choice item, but are perfectly comfortable basing midterm grades on a test of 50 multiple-choice items. This is because, for any given item, measurement error represents a large percentage of students' scores. The percentage of measurement error decreases as test length increases. Even very low achieving students can answer a single item correctly, even through guessing; however it is much less likely that low achieving students can correctly answer all items on a 20-item test.

Although reliability does increase with test length, the reward is more evident with short tests than with long ones. Increasing test length by 5 items may improve the reliability substantially if the original test was 5 items, but might have only a minimal impact if the original test was 50 items.

2.3.3. USE OF TABLE OF SPECIFICATION (BLOOM'S TAXONOMY).

To avoid any lopsidedness and overrepresentations of tests it necessitates test Constructors to go back into the statement of objectives made for the term. The objectives will serve as the direction on what specific task is to be measured. To guard any fortuitous imbalances and disproportionate item distribution, test constructors draws up a table of specifications before any items are prepared. Such specifications should begin with an outline of both the instructional objectives of the course, the subject matter to be covered, and the cognitive skills measured – a three-way grid (Gronlund, 1990).

According to a research by Justin D, Valentin and John R. Godfrey on teacher made tests in Seychelles they concluded that little evidence of content-related validity was observed in the tests. For some tests, objectives of the syllabus were not well represented and the weighting of items was not balanced. Furthermore, the items mainly focused on the lower level of the cognitive domain of Bloom's taxonomy of educational objectives (Bloom, 1956). The latter is a consistent finding in teacher-made tests in general (Marso & Pigge, 1991; McMorris & Boothroyd, 1992; Oescher & Kirby, 1990).

The low content-related validity evidence of the tests may be explained by the lack of planning in test construction. Marso and Pigge (1992) noted that a major reason why quality of teacher-made tests is poor is because of the lack of preparation in the test construction. Tuckman (1988) described careful planning as one way of ensuring the content-related validity of a test. The use of a table of specifications is recommended in order to provide a base for careful sampling of items thus ensuring that the relationship between objectives and items is established within the test (Gay, 1991; Linn & Gronlund, 1995).

Tests submitted for this study were not accompanied with a plan or a table of specifications. The teachers reported that it would take them too much time to write the objectives of the tests they submitted. Marso and Pigge (1992) argued that lack of planning is due to lack of training. It is probable that teachers are unaware of the ways they can plan their tests. Another reason is that these teachers construct too many tests at once. All the teachers surveyed teach at least three subjects in different classes and are responsible for organizing testing for all their classes. It is speculated that the time available for preparing tests is insufficient.

Bloom's taxonomy is a multi-tiered model of classifying thinking according to six cognitive levels of complexity. The levels are depicted as a stairway, leading many teachers to encourage

their students to climb a higher level of thought. The lowest three levels are knowledge, comprehension and application. The highest three levels are analysis synthesis and creating. The taxonomy is hierarchical in that each level is subsumed by the higher levels e.g. a student functioning at applying level is assumed to have mastered the material at the knowledge and comprehension levels. It is easy to see how this arrangements leads to natural driver of lower and higher level thinking skills.

The use of Bloom's taxonomy in tests construction is of a rule rather than the exception. This helps to increase the validity and reliability of the teacher made tests. This study will seek to establish the extent to which teachers use Bloom's taxonomy in test construction and the effect this has on the validity and reliability of their exams.

2.4. EFFECT OF VALID AND RELIABLE FORMATIVE TESTS ON SUMMATIVE ASSESSMENT

There are several potential advantages in using teachers' judgements more widely as part of summative assessment for external as well as internal uses. First, teachers are making judgements about students' attainment in the course of their normal interactions during teaching and learning. Second, in this process teachers can build up a picture of students' attainments across the full range of activities and goals. This gives a broader and fuller account than can be obtained through any test that uses a necessarily restricted range of items and thus provides a more valid means of assessing outcomes of education (Crooks, 1988; Wood, 1991; Maxwell, 2004). Third, the teacher has the opportunity to use such accumulating information gathered in this way to help learning. Fourth, it can facilitate a more open and collaborative approach to

summative assessment, in which students can share through self-assessment and derive a sense of progress towards learning goals rather than performance goals.

That these potential advantages can be translated into reality is evident in practice in systems, such as those in Queensland and Sweden, where teachers' judgements are used for assessment on which important decisions for students are based.

At the same time, there are potential arguments against teachers having a significant role in summative assessment. In the first place, there is no doubt that there is evidence of unreliability and bias in teachers' assessment. Second, where the assessment is for external use (such as for certification by an awarding body), there would be additional work for the teachers and resources required for moderation procedures. Third, there is the possibility that the requirements of moderation procedures could constrain teachers' use of the full range of evidence available to focus only on what can be 'safely' assessed. There is warning here that summative assessment by teachers in some circumstances can have the same narrowing effect on the curriculum as do tests.

These opposing arguments gave rise to a search for the evidence in relation to the question: What is the evidence concerning the reliability and validity of assessment by teachers used for summative purposes and how might it be improved?

The following were the findings of Wynne Harlem, University of Cambridge (2005). She found that the extent to which the assessment tasks, and the criteria used in judging them, are specified are key variables affecting dependability. Where neither tasks nor criteria are well specified, dependability is low. Detailed criteria, describing progressive levels of competency, have been shown to be capable of supporting reliable assessment by teachers. Tightly specifying tasks does not necessarily increase reliability and is likely to reduce validity by reducing the opportunity for a broad range of learning outcomes to be included. Greater dependability is found where there

are detailed, but generic, criteria that allow evidence to be gathered from the full range of classroom work.

Bias in teachers' assessments is generally due to teachers taking into account information about non-relevant aspects of students' behaviour or being apparently influenced by gender, special educational needs, or the general or verbal ability of a student in judging performance in a particular task. Researchers claim that bias in teachers' assessment is susceptible to correction through focused workshop training. Participation of teachers in developing criteria is an effective way of enabling the reliable use of the emerging criteria. There is variation in the way that teachers gather information from students' regular work, but no evidence that this affects dependability. But it is important for teachers to follow agreed procedures for applying criteria to the evidence they collect. Consistency in applying criteria depends upon teachers being clear about the goals of the work and on the thoroughness with which relevant areas of the curriculum are covered in teaching.

The context of the school's support and value system has a role in how assessment by teachers is practiced. Conditions associated with greater dependability include the extent to which teachers share interpretations of criteria and develop a common language for describing and assessing students' work. Students find assessment of coursework motivating, enabling them to learn during the assessment process. But they need more help to understand the assessment criteria and what is expected of them in meeting these criteria.

The way in which teachers present classroom assessment activities may affect students' orientation to learning or performance goals. Changing teachers' assessment practices to include processes and explanations leads to better student learning. The introduction of teachers' assessment related to levels of the National Curriculum in England and Wales was perceived by

teachers as having a positive impact on students' learning, enhanced by teachers working collaboratively towards a shared understanding of the goals of the assessment and of procedures to meet these goals. Teachers find compensation for the time spent on assessment in the information they gain about their students and about learning opportunities for students that need to be extended.

The existence of criteria, and particularly involvement in identifying them, help teachers' understanding of the meaning of learning outcomes. But criteria that identify qualitative differences in progression towards learning goals need to be distinguished from externally devised checklists, which encourage a mechanistic approach to assessment.

Close external control of teachers' summative assessment can inhibit teachers from gaining detailed knowledge of their students. Opportunities for teachers to share and develop their understanding of assessment procedures enable them to review their teaching practice, their view of students' learning and their understanding of subject goals. Such opportunities have to be sustained over time and preferably should include provision for teachers to work collaboratively across, as well as within, schools.

There is considerable similarity in some of the implications from the research evidence in the three reviews, relating particularly to: the importance of providing non-judgemental feedback that helps students know where they are in relation to learning goals; the need for teachers to share with students the reasons for, and goals of, assessment; the value to teachers of using assessment to learn more about their students and to reflect on the adequacy of the learning opportunities being provided; teachers and students placing less emphasis on comparisons among students and more on individual development; and helping students to take responsibility for their learning and work towards learning goals rather than performance goals. All these

points are ones that favour formative assessment as well as improving the dependability and positive impact of summative assessment by teachers. It follows that the actions teachers need to take in developing their assessment for summative purposes overlap to a great extent with the actions required for practicing formative assessment.

2.5. VALIDITY CRITERIA

Validity is the extent to which an assessment measures what is needed for a particular purpose and the results, as they are interpreted and used, meaningfully and thoroughly represent the specified knowledge or skill. This definition highlights the importance of considering purposes and intended uses when developing and selecting assessment procedures. Those procedures must assess the knowledge or skill (or learning goal, outcome, or objective) that they claim to measure. The type of information produced must be useful for the intended purposes.

Linn, Baker, and Dunbar (1991) proposed eight criteria for evaluating validity in performance-based assessment. However, the criteria are also appropriate for forced-choice assessment and may help improve it more than traditional statistical validation procedures. These criteria are as follows:-

1) Consequences

This focuses on the effects of the assessment. This criterion may be of limited utility for individual assessment procedures. Educators find it more useful to examine all procedures that, together, will assess a particular content unit (e.g., learning outcome). While individual

assessment procedures must be appropriate for the content assessed, they may cover only a portion of it.

2) Content Quality

This focuses on the consistency with current content conceptualization. This criterion is especially important in content areas (e.g., science) in which knowledge sometimes grow rapidly. An assessment that represents an outdated conceptualization of the content assessed is not likely to produce useful information and will waste students' and teachers' time.

3) Transfer and Generalizability

This focuses on the assessment's representatives of a larger domain. For different reasons, this criterion is of concern in both forced-choice and performance-based assessment. In the former, the nature of questions may make them poor indicators of student ability to deal with concepts in the domain assessed. In performance-based assessment, the small number of tasks makes it essential that each task or group of tasks is representative of the domain assessed. Also, continued re-use of any type often may compromise generalizability, because teachers and students may focus on specific items or tasks from the test rather than on the larger domain.

4) Cognitive Complexity

This focuses on whether level of knowledge assessed is appropriate. Do the assessment tasks or questions represent the cognitive complexity of the knowledge or skill that it is intended to assess? (For example, if an outcome includes higher order or critical thinking skills--such as problem solving or synthesis--does the assessment measure them?). Does the assessment actually

require students to use higher-level knowledge or skills, or can students simply respond from memory without having to think?

5) Meaningfulness

The focus here is on the relevance of the assessment in the minds of students. Assessment procedures must be meaningful to students in order to produce valid, useful information. Assessment that is relevant to students' personal experiences is likely to motivate students to perform as well as possible. However, some assessments cannot be made relevant to problems students encounter in real life, and educators should realize that contrived assessments may poorly represent the knowledge or skills assessed.

6) Fairness

The focus here is on the fairness to members of all groups. For some assessment purposes (e.g., to measure the achievement of individual students in a content area), it may be important to consider whether all students have had similar opportunities to acquire the knowledge or skills assessed. For example, are some students at an advantage because ancillary skills (such as prior knowledge or reading ability) that are not relevant to the focus of the assessment enable them to score higher? On the other hand, if the purpose is to find out whether a group of students has achieved something (such as a learning outcome) and some students have not had an opportunity to learn it, the assessment is not unfair or invalid. Rather, instruction may not have been adequate.

7) Cost and Efficiency

The focus is on the practicality or feasibility of an assessment. The questions to ask here are: Is the assessment a reasonable burden on teachers, instructional time, and finances? And is the resulting information worth the required costs in money, time, and effort?

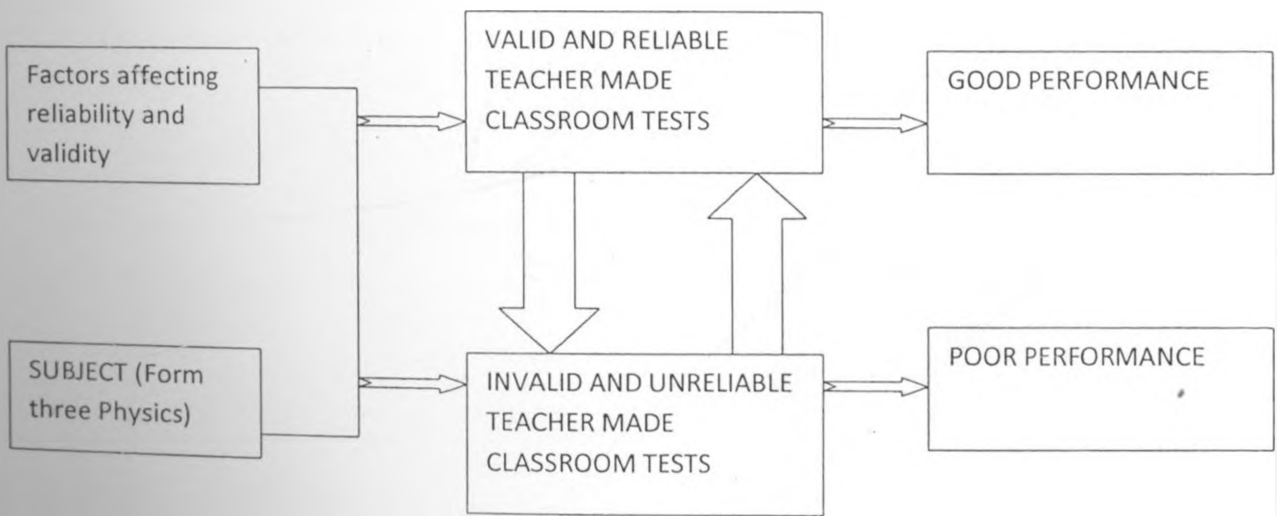
2.5. HYPOTHESIS

H₀: Teacher made classroom tests are valid and reliable.

H₁: Teacher made classroom tests are not valid and reliable.

2.6. CONCEPTUAL FRAMEWORK

This study hypothesizes that teacher made tests are valid and reliable. The various ways of determining internal consistency will be used to determine reliability. Also the methods of determining validity will be applied to examine the extent of validity of teacher made tests.



The effect of the hypothesized relationship between the dependent variable and independent variable is that teacher made classroom tests can be used even to evaluate effectively and predict with precision future performance of a student.

CHAPTER THREE: METHODOLOGY

3.1. INTRODUCTION

This chapter covers research design, target population, sampling procedure, research instruments, data collection procedures, data analysis techniques, and ethical considerations.

3.2. RESEARCH DESIGN

The purpose of this study is to establish validity and reliability of teacher made classroom tests. Research method chosen was used to collect data that was used to this particular end. There are various designs used here i.e. classical experimental, quasi experimental, case study and survey. The purpose of case study as this one was to seek and determine factors and relationships among the factors that would result to reliability and validity of teacher made classroom tests. Since the focus of a case study was typical selection of subjects was done carefully to ensure that the unit selected was typical to those it will be generalized to.

Descriptive research seeks to find answers to questions through the analysis of variable relationships. A descriptive survey research design was then used to evaluate the validity and reliability of teacher made classroom tests in Nyahururu district. This design was chosen because it helped the researcher to obtain the primary data required to determine the specific characteristics of teacher made tests. The design also provided a guide in the collection, analysis, and interpretation of observations quickly.

3.3. TARGET POPULATION

The data was collected from teachers in Nyahururu district. All the teachers of physics in the District was used. Also the teachers in charge of examinations or curriculum masters were used in key informant interview to give more details.

3.4 SAMPLE SELECTION AND SAMPLE SIZE

Under this research a simple random sampling technique was used to collect data among teachers in Nyahururu District.

A sample size will be determined using simple random sampling technique recommended by Krejcie, 1970 (Appendix 4). From 45 teachers a sample of 42 teachers was used.

3.5 RESEARCH INSTRUMENTS

Questionnaires (see appendix 1) were used consisting of closed ended questions to collect the primary data. Questionnaires were chosen because they are relatively cheap, it is free from the bias of the interviewer, respondents have adequate time to give well thought out answers and it is easy to communicate with respondents. The questionnaires were administered by the researcher. Also the researcher collected test papers for the term two examination 2012 results to analyze how it their fit into the criteria for validity and calculate their reliability.

The researcher also used structured interviews (see appendix 2) in which he used a set of predetermined questions. These were used to give more in depth information on the teacher made tests. The interviews were used because of their advantages to allow the researcher to probe for particular responses, clarifications and confirmation of information from respondents.

3.5.1. VALIDITY OF RESEARCH INSTRUMENTS

It refers to the appropriateness, meaningfulness, and usefulness of the inferences a researcher makes, (Mwangi, 2008). Under this study the researcher strived to make the language used as simple as it is practically possible. This ensured that the response gotten is as intended by the researcher.

3.5.2. RELIABILITY OF RESEARCH INSTRUMENTS

According to Mwangi (2008) reliability means consistency of scores or answers from one administration to another. The researcher strived to standardize conditions under which the data was being collected to improve on reliability. Also the researcher gave clear directions for filling the questionnaires.

3.6. DATA COLLECTION PROCEDURES

The researcher visited the participating schools in the research in person. After introduction to the head of institutions, request was made to collect the data required from the various teachers. Questionnaires were then distributed to teachers. Also the researcher requested teachers to provide examination results for form three 2012 examination for term two. The researcher also conducted an interview with the curriculum masters to give more information on the teacher made tests.

3.7. DATA ANALYSIS

The data was coded and run through the statistical Program for Social Sciences (SPSS). The qualitative data was analyzed using descriptive statistics. The researcher also examined all

completed questionnaires and filled the information in a frequency table. The magnitude of correlation between the variables was then calculated.

The researcher with the help of two other teachers evaluated various examination questions to determine their level in the criteria for validity. These are teachers with an experience of more than 10 years and have worked in various stations. They are also Kenya National Examinations Council examiners. The following classifications were used to determine validity of various examinations:-

V₁-Consequencies: 1.Very inconsequential, 2.Inconsequential, 3. Neutral, 4, Consequential and 5.Very consequential

V₂-content quality: 1.Very low content quality, 2.Low content quality, 3.Moderate content quality, 4.High content quality and 5. Very high content quality.

V₃-Generalizability: 1.Very specific, 2.Specific, 3.Neutral, 4.Generalizable And 5.Totally generalizable.

V₄-Cognitive complexity: 1.Very simple 2. Simple, 3.Fair, 4.Complex and 5.Very complex

V₅-Meaningfulness: 1.Very meaningless, 2.Meaningless. 3.neutral, 4.meaningful and 5.Very meaningful.

V₆-Fairness: 1.Very unfair, 2.Unfair, 3.Neutral, 4.Fair and 5.Very fair.

V₇-Cost and efficiency: 1.Very expensive and inefficient, 2.Expensive and inefficient, 3.neutral, 4. Cheap and efficient and 5.Cheap and efficient.

The levels of each factor were also classified as follows:-

1. 1-5 years-Novice who had little situational perception and discretionary judgment.
2. 6-10 years-Advanced beginner with all attributes and aspects treated separately.
3. 11-15 years-Competent whose plan guides performance as situation evolves.
4. 16-20 years-Proficient with situational factors guiding performance as situation evolves.
5. > 20 years-Expertise with intuitive recognition of appropriate decision or action.

After the above classifications for each of the factors the results were analyzed and averaged to get the value of each exam.

The quantitative data was analyzed using quantitative methods and presented by use of tables, frequencies, percentages, statistical measures of relationship between the dependent and independent variables. The researcher also carried out an analysis of examination results in order to calculate the reliability coefficient. The results were used to draw conclusions and in making recommendations.

3.8. ETHICAL CONSIDERATIONS

The data collection procedures and instruments under this research were preceded by an introductory letter which served to assure respondents that the information given will solely be used for the purposes of the study.

The questionnaires were administered through voluntary informed consent and the participants assured that no harm would result from any information given either in filling the questionnaires or answering any question. All the participants were treated with respect and courtesy. All other ethical considerations were taken into account.

CHAPTER FOUR

DATA ANALYSIS, PRESENTATION, INTERPRETATION AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the study findings which has been analyzed and discussed under the following thematic areas response return rate, demographic characteristics of respondents, teachers' experience, education level, training on test construction and analysis, use of Bloom's taxonomy, exam moderation, length of test and conclusion.

4.2 Response return rate

A total of 42 teachers and 15 key informants were used for this study within Nyahururu District of Laikipia County. Questionnaires were administered to them and interviews carried out for the key informants. However out of the targeted respondents 41 were able to return their questionnaires representing 97.6%. all the intended key informants were reached and they provided the needed information.

Table 1 Questionnaire return rate

Response	Frequency	Percentage
Returned	41	97.6
Not returned	1	2.4
Total	42	100.0

Only one person was not able to return the questionnaire representing 2.4%. however this has been considered insignificant as Hunt, (2001) under his studies in probability sampling argues that at 5% alpha 2.4% is not significant to affect the results of a study findings. The intended target was however achieved in the study area.

4.3 Demographic characteristics of the respondents

This section focused on the demographic profile of the respondents selected for this study. The section provided a basis for understanding and evaluating the composition of the respondents and to determine if the respondents age , gender, location and relationship with other members of staff affect the validity and reliability of teacher made tests.

4.3.1 Distribution of the respondents by age

UNIVERSITY OF NAIROBI
UNIVERSITY OF NAIROBI

The study was interested in age of teachers. The table 2 shows a summary of the results. The respondents were asked to state their age brackets.

Table 2: Age of respondents in years

Age bracket	Frequency	Percentage
21-30	3	7.3
31-40	21	51.2
41-50	16	39.0
51-60	1	2.5
Total	41	100.0

Table 2 shows that majority of the respondents are aged between 31-40 years represented by a percentage of 51.2%. At this age bracket teachers are experienced and are at their prime to carry out testing effectively. This was followed by respondents who were aged between 41-50 years represented by 39.0%. Then this was followed by those age between 21-30 represented by 7.3% and finally those aged 51-60 represented by 2.5%.

4.3.2 Distribution of respondents by gender

Majority of the respondents (80.5%) were males, while 19.5%were females as shown in table 3.

Table 3 distribution of the respondents by gender

Gender	Frequency	Percentage
Male	33	80.5
Female	8	19.5
Total	41	100.0

While there seems to be a big disparity between the male and female counterparts this did not affect the data collected since the gender of an individual does not affect the validity and reliability of their tests. This came up due to the fact that not many ladies in the past chose to take sciences though this trend may change as years go by. Also the disparity may be due to the fact that this is a hardship area and so few female teachers are posted or willing to come to such an area.

4.3.3 Distribution of respondents by their location from school

This study was interested in how far ne lives from the school. The respondents were asked whether they live near the school. The table 4 shows a summary of the results.

Table 4 Distribution of respondents by location from the school

Choice	Frequency	Percentage
Strongly disagree	2	4.9
Disagree	6	14.6
Neutral	4	9.8
Agree	27	65.9
Strongly agree	2	4.9
Total	41	100.0

Most of the respondents live near the school represented by 65.9%. Followed by those who live not so far (14.6%), those not far and not so near (9.9%), those living far and those living within school were equal at 4.9%.

4.3.4 Distribution of respondents on cordial relationship with other staff

The study was also interested in finding out the relationship between respondents and other members of the staff i.e. teaching and non-teaching staff. Table 5 is a summary of the findings.

Table 5 Distribution of respondents by their relationship with other members of staff

Choice	Frequency	Percentage
Strongly disagree	1	2.5
Disagree	0	0
Neutral	0	0
Agree	24	58.5
Strongly agree	16	39
Total	41	100.0

Most of the respondents (58.5% and 39%) have a good relationship with other members of staff.

Only 1(2.5%) has a poor relationship with other members of staff.

4.4 Factors affecting validity and reliability

This study was interested in determining the factors that affect reliability and validity of teacher made tests. It was established that there are various factors that affect the reliability and validity of teacher made tests.

4.4.1 Teachers' experience and the effect on reliability and validity

From table 6 it shows that as the experience of teachers increase there is a corresponding increase in validity.

Table 6 Teachers experience and corresponding measures of validity and reliability

Teacher's experience	Level of experience	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	RELIABILITY (Kuder-Richardson method)
	1	4	4	4	4	4	4	4	4
2	2	3	4	5	5	4	4	4	0.8
3	4	4	4	5	5	4	4	4	0.7
4	4	4	5	5	5	5	5	5	0.8
5	5	5	5	5	5	5	5	5	0.8

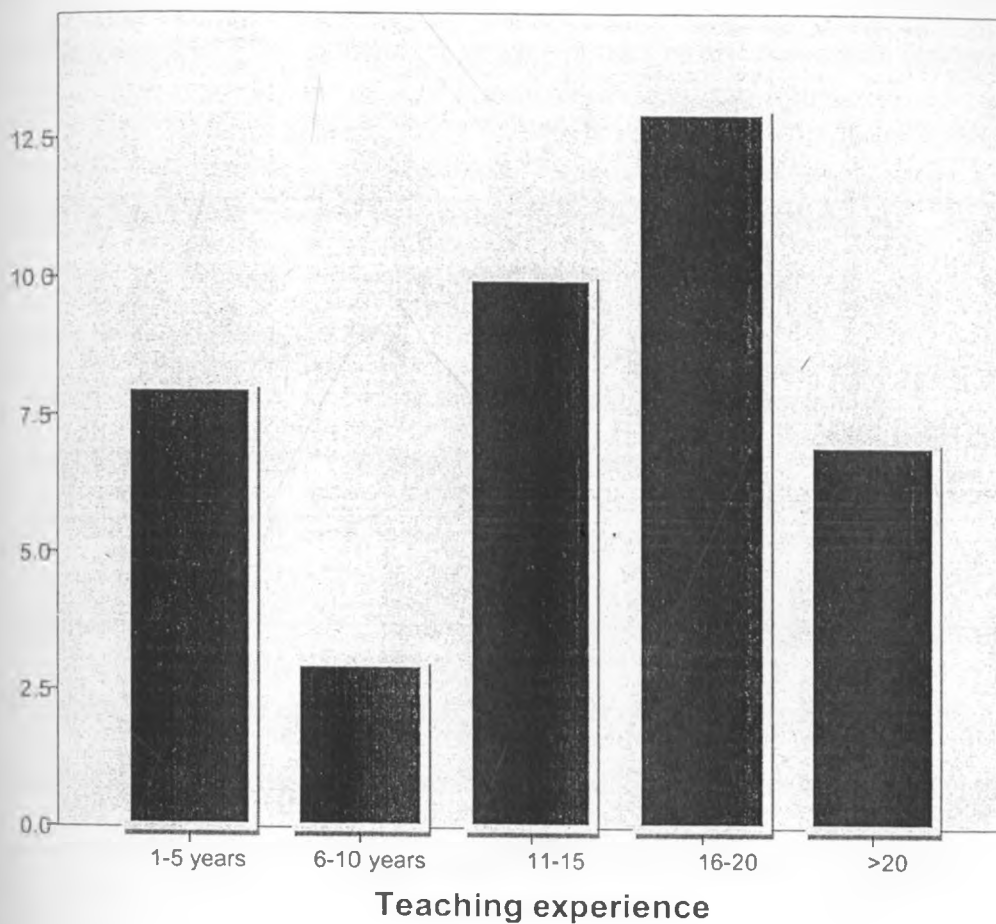
Though teachers experience affects and varies with the number of years one has been teaching, this does not seem to have the same effect on reliability. This implies that validity is affected so much by the experience of teachers rather than reliability.

The table 7 shows a summary of the number of years that teachers have been teaching. Teachers were asked how many years that have been teaching.

Table 7 Distribution of respondents on experience

Number of years	Frequency	Percentage
1-5	8	19.5
6-10	3	7.3
11-15	10	24.5
16-20	13	31.7
>20	7	17.1
Total	41	100.0

Graph 1. Summary of the number of years one has been teaching



Most of the respondents have an experience of 16-20 years(31.7%), followed closely by those with 11-15 years(24.5%), then those with 1-5 years(19.5%). those with more than 20 years follow with 17.1% and finally those with 6-10 years at 7.3%. this shows that most teachers have an

experience of between 11-20 years. This is good experience to be able to perform their responsibility of testing effectively.

The key informants were asked whether teachers experience affects the reliability and validity of the tests they construct. The following is an excerpt from of discussion with few interviewees.

Interviewer: Does the number of years one has taught affect quality of their tests in terms Of validity and reliability?

Key informant 1: Definitely it does.

Key informant 2: Yes it does, because initially it is kind if they are still in training.

Key informant 3: Yes it does and keeps improving all along.

Key informant 4: It is true it does

Most of them were of the opinion that as the number of years increases the validity and reliability will also improve though also it has a limit kind of where one attains optimum point.

The $\chi^2=6.683$, with $p=0.01$ and 4 df. Thus the significance of experience is high.

4.4.2 Education level and effect on reliability and validity

Table 8 shows that the level of education affects reliability and validity. The respondents were asked to state their level of education.

Table 8 Education level versus levels of reliability and validity

Levels	RELIABILITY							(Kuder-Richardson method)
	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	
1	-	-	-	-	-	-	-	-
2	4	4	4	4	4	4	4	0.68
3	5	5	5	5	5	5	5	0.74
4	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-

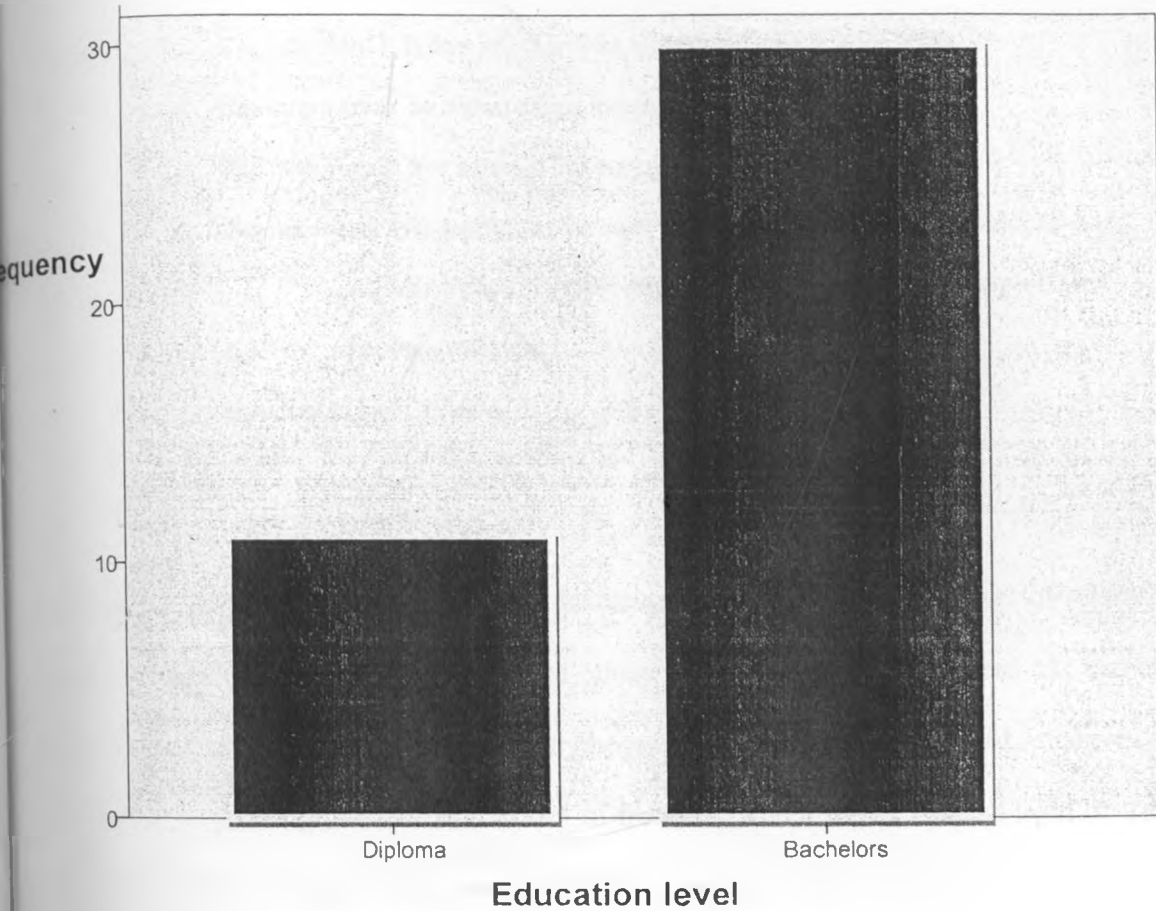
As the level of education rises also validity and reliability also get better. This shows that the level of education affects reliability and validity of teacher made tests.

The distribution of respondents in terms of their education level is shown in table 9 below. The respondents were asked to state their level of education i.e. Certificate, Diploma, Bachelors, Masters or Doctorate.

Table 9 Distribution of respondents by their education level

Level of education	Frequency	Percentage
Certificate	0	0.0
Diploma	11	26.8
Bachelors	30	73.2
Masters	0	0.0
Doctorate	0	0.0
Total	41	100.0

Graph 2. Summary of education level of respondents



Most of the respondents have bachelors degree (73.2%) while the others have diploma (26.8%).

Most of the teachers are well qualified to set tests that meet the criteria for validity and also have a high reliability. The key informants were asked whether they thought that education level affected quality of teacher made tests. The following is an excerpt of their responses.

Interviewer: Education level affects validity and reliability of teacher made tests.

Comment.

Key informant 1: It does affect and the more qualified one is the better.

Key informant 2: To a great extent it does.

Key informant 3: Yes it does affect but not so much.

Key informant 4: It depends on the level you are comparing. If between certificate and degree there is a lot of difference but not so big between diploma and degree holders.

Key informant 5: Yes it does.

Most of the key informants felt that the education level does affect the level of reliability and validity. It was felt that one could not take others where they have never been and therefore the better one is trained the more efficient they become in constructing valid and reliable tests. The $\chi^2=8.805$ at $p=0.01$ significance and $df = 4$ of freedom, which shows that the level of training is significant in determining validity and reliability of teacher made tests.

4.4.3 Training on test construction and analysis

This study was interested in identifying whether training on test construction and analysis has effect on validity and reliability.

Table 10 Training on test construction and corresponding values of reliability and validity

Level of training	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	RELIABILITY
								(Kuder-Richardson method)
1	3	3	3	3	3	3	3	0.64
2	3	3	3	3	3	3	3	0.68
3	3	3	3	3	3	3	3	0.76
4	5	5	5	5	5	5	5	0.8
5	4	4	5	4	4	5	4	0.79

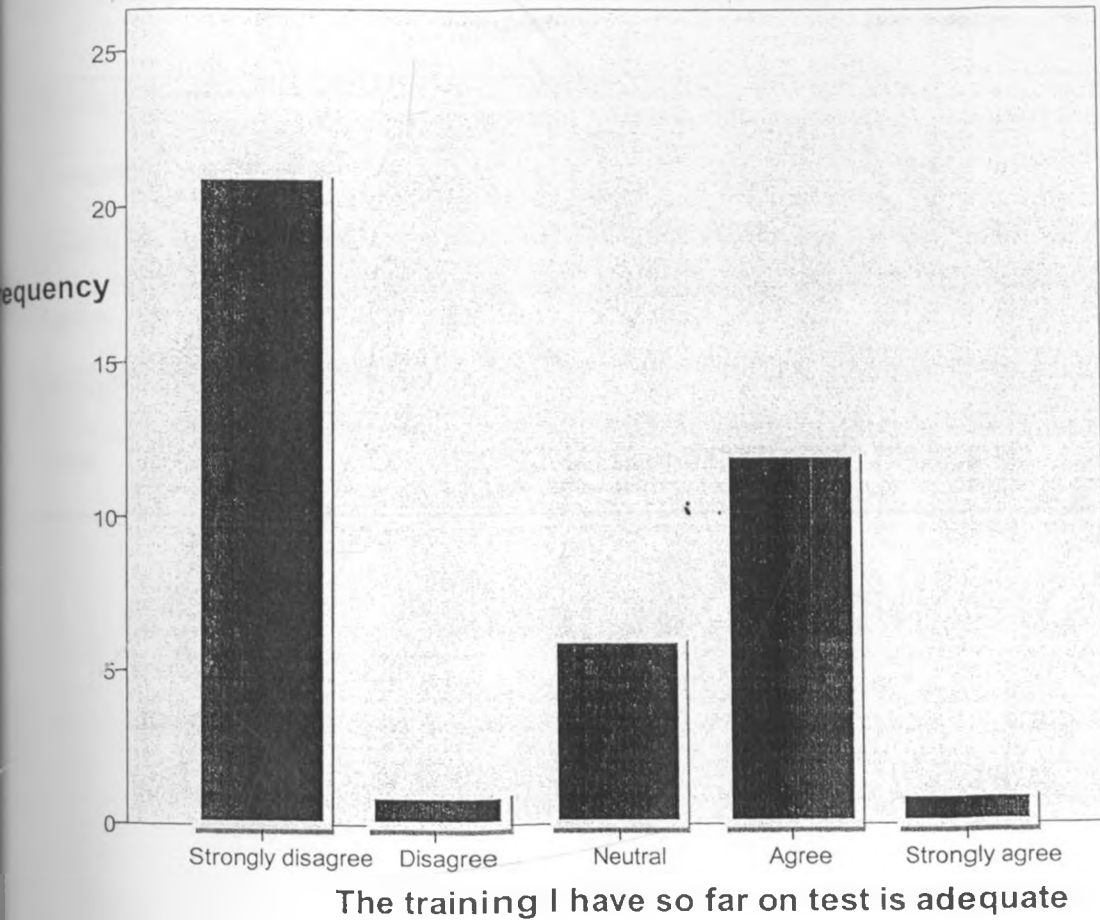
From table 10 as the level of training one have on test construction increases also the reliability increases. The trend is observed when we consider reliability though there is no much variation.

When asked whether the training one has had on test construction and analysis is adequate most of the respondents strongly disagreed (51.2%) while only 2.4% strongly agreed that their training is adequate. Table 4.11 shows a summary of the results.

Table 11 Summary of the results on training on test construction and analysis

	Frequency	Percentage
Strongly disagree	21	51.2
Disagree	1	2.4
Do not know	6	14.6
Agree	12	29.3
Strongly agree	1	2.4
Total	41	100.0

Graph 3. Summary of adequacy of training on test construction and analysis

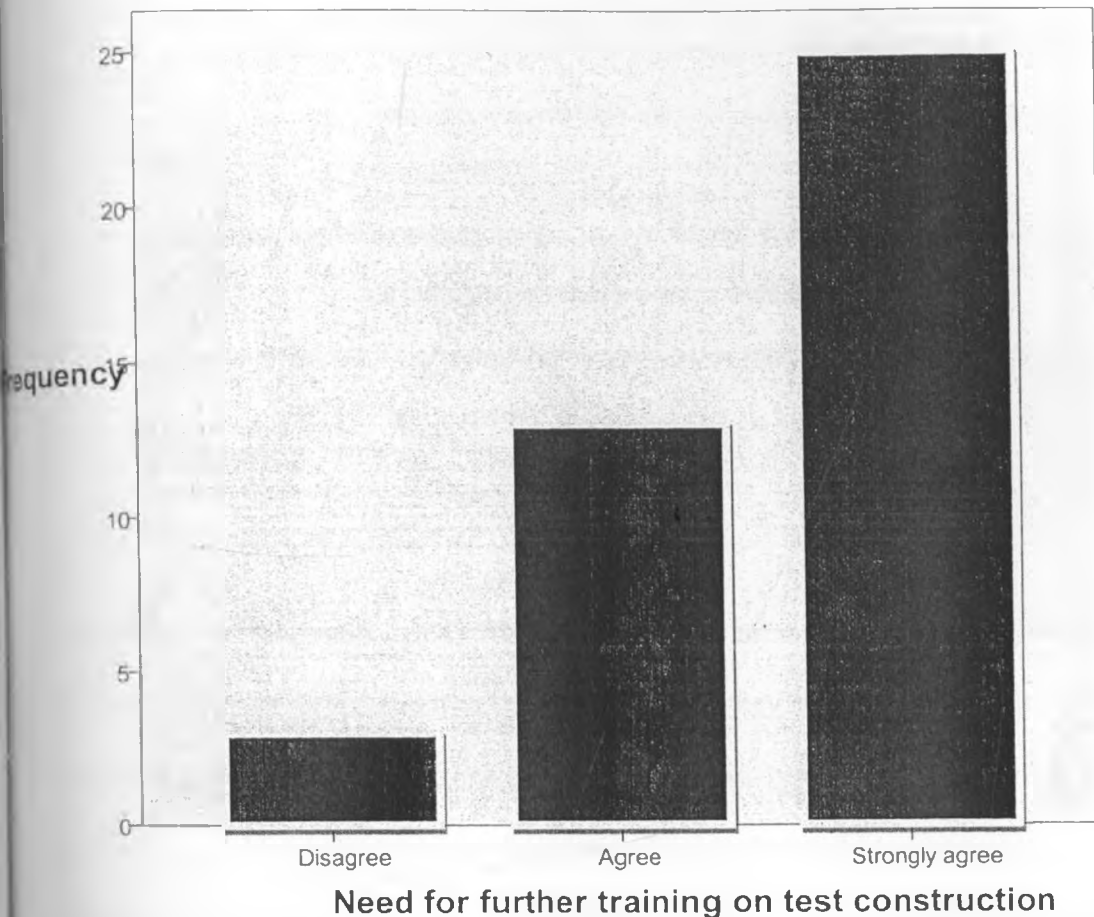


When the respondents were asked whether they need further test on test construction and analysis most of them strongly agreed (61%), while 31.7% agreed. Only 7.3% disagreed and none was neutral or strongly disagreed. The table below shows a summary of the results. The $\chi^2=34.976$ at 0.01 significance and $df =4$. There is high significance of training on test construction and analysis.

Table 12 Distribution of respondents on need for further training on test construction and analysis

	Frequency	percentage
Strongly disagree	0	0
Disagree	3	7.3
Neutral	0	0
Agree	13	31.7
Strongly agree	25	61.0
Total	41	100.0

Graph 4. Summary of need for further training on test construction



Most of the respondents are not satisfied in their current level of training in test construction and analysis to further enhance their skills.

The key informants were asked whether they think teachers have enough training on test construction and whether more training is needed. The following is an excerpt from a few key informants.

Interviewer: Do you think teachers are well trained on test construction or more is needed.

Key informant 1₁: There is a lot that needs to be done on training of teachers on test construction and analysis.

Key informant 2: Even the institutions from which teachers come from affect the level of reliability and reliability and a lot needs to be done to harmonize these disparities.

Key informant 3: The training leaves a lot to be desired. Teachers need a lot of in-service training on test construction and analysis.

Key informant 4: All the teachers need to keep sharpening their skills in test construction and analysis.

Most of the key informants agree strongly that teachers need more training on test construction and analysis. When the question was posed to them the key informants also felt that they themselves need to be trained on test construction and analysis. Some went further to comment that there is need for more training when one undergoes the teaching course in teacher training institutions. The $\chi^2=17.756$ at 0.01 significance level and $df = 4$ which indicates the importance placed on further training on test construction.

4.4.4 Use of Bloom's taxonomy

The use of Bloom's taxonomy in constructing tests is more of a rule rather than exception. From table 13 it shows that use of bloom's taxonomy improves the reliability and validity of the teacher made tests.

Table 13 Usage of Bloom's taxonomy and corresponding values of reliability and validity

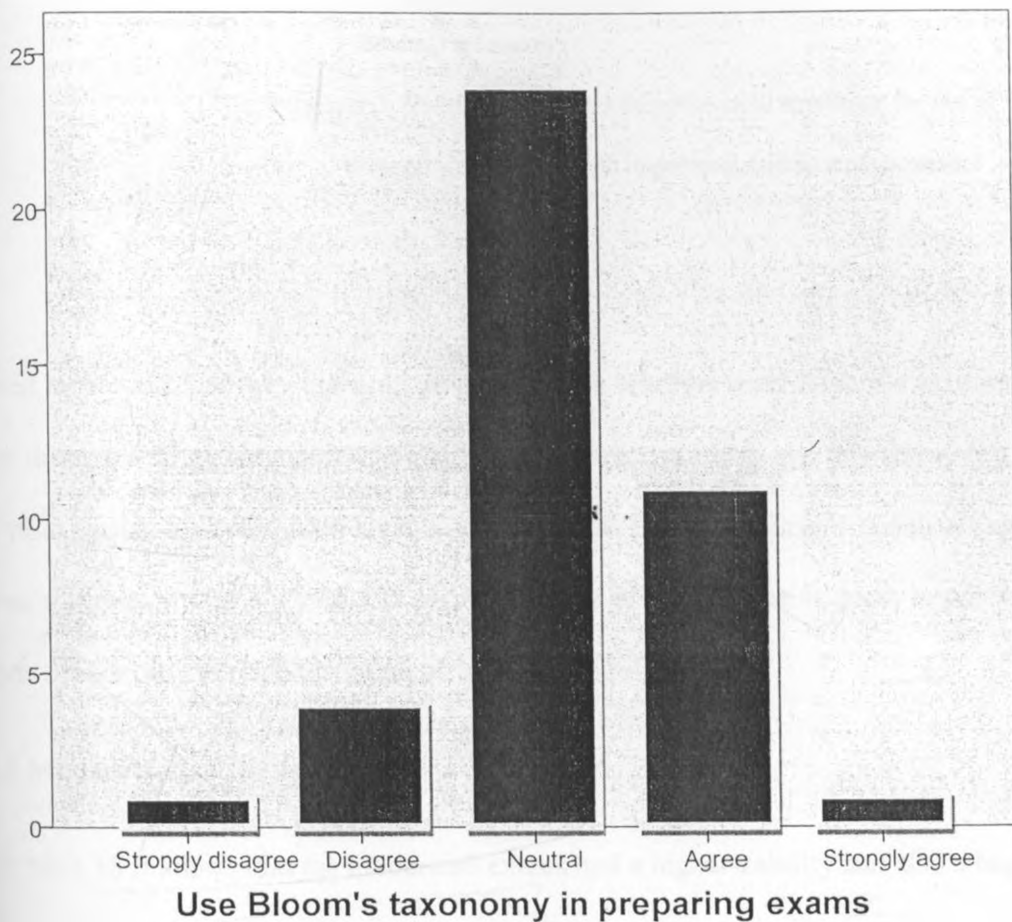
Level of usage of Bloom's taxonomy	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	RELIABILITY (Kuder-Richardson method)
	1	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3	0.60
3	4	4	4	4	4	4	5	0.7
4	4	4	4	4	4	4	4	0.79
5	4	4	4	4	4	4	4	0.82

Those exams set in accordance to the levels of Bloom's taxonomy have a high reliability and also validity. The only surprising thing is that most teachers never really seem to be sure whether they use Bloom's taxonomy. The respondents were asked whether they use Bloom's taxonomy in the process of constructing their exams. The following table is a summary of the results obtained.

Table 14 Distribution of respondents in terms of use of Bloom's taxonomy

	Frequency	Percentage
Strongly disagree	1	2.4
Disagree	4	9.9
Neutral	24	58.5
Agree	11	26.8
Strongly agree	1	2.4
Total	41	100.0

Graph 5. Summary of the use Bloom's taxonomy in preparing exams



The key informants could not be further from the truth since when asked whether teachers use Bloom's taxonomy most were not sure about. The excerpt below is a summary of some responses from key informants.

Interviewer: Do teachers use Bloom's taxonomy in preparing tests.

Key informant 1: Not really. Most teachers do not seem to understand it.

Key informant 2: No. Teachers are not able to identify objectives in there tests using
Bloom's taxonomy.

Key informant 3: Not really. Teachers do not seem to appreciate the use of Bloom's
taxonomy or recognize its importance when preparing tests.

Indeed most said that they did not really think that teachers used Bloom's taxonomy. It seems either there are no mechanisms of determining whether teachers use this taxonomy or there is a lack of expertise to do so. Also most would say it takes a lot of time to construct an exam using Bloom's taxonomy. The $\chi^2=46.195$ at $p=0.01$ and $df = 4$. There is great importance in using Bloom's taxonomy to raise the value of validity and reliability.

4.4.5 Moderation of the tests

From table 15 it shows that the moderated exams had a higher validity and also a high reliability. This shows that the input of the members of department is very important in improving the validity and reliability of the tests constructed by teachers.

Table 15 Moderation of tests and corresponding values of validity and reliability

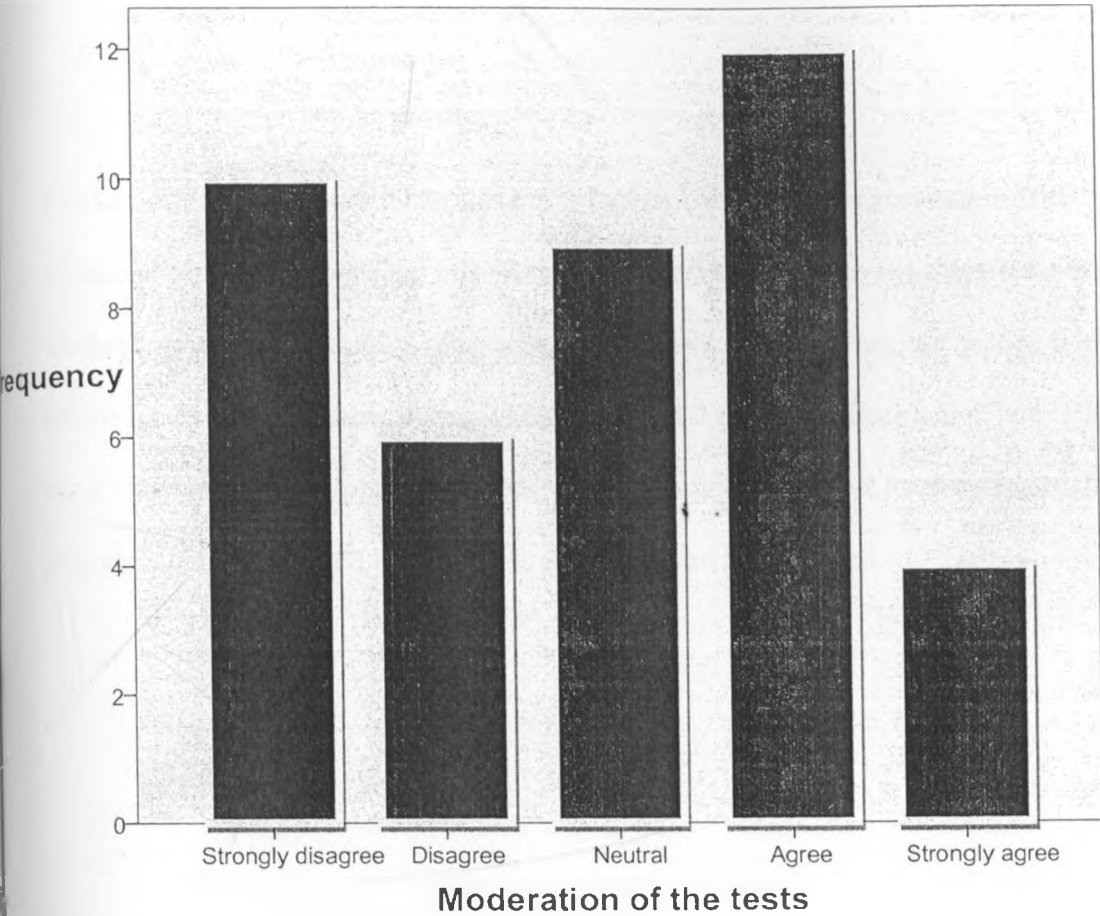
Level of moderation	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	RELIABILITY (Kuder-Richardson method)
1	4	4	4	4	3	4	4	0.8
2	3	5	4	3	5	5	2	0.85
3	5	5	5	5	5	5	3	0.85
4	5	5	5	5	5	5	5	0.89
5	5	5	5	5	5	5	5	0.8

The respondents were asked whether they moderate their tests. The table 16 is a summary of the results.

Table 16 Distribution of respondents in terms of exam moderation

	Frequency	Percentage
Strongly disagree	10	24.4
Disagree	6	14.6
Neutral	9	22.0
Agree	12	29.3
Strongly agree	4	9.9
Total	41	100.0

Graph 6. Distribution of respondents in terms of exam moderation



It seems that not many respondents subject their exams to moderation and this goes to affect the level of reliability and validity of teacher made tests.

Most of the key informants accepted that their respective institutions do not moderate exams but it is very important. The following is an excerpt of the discussion with a few key informants.

Interviewer: Is moderation important and do you do it in your institution.

Key informant 1: It is very important though we do not carry it out.

Key informant 2: Very important but we do not have time to do it.

The key informants reported that this never happens and most teachers assume that if they take questions from past papers and mix them with their own it serves to moderate the exams. Also teachers view that moderation is an unnecessary burden and yet they are trained to construct exams in universities and colleges. The $\chi^2=4.976$ at 0.01 significance and $df = 4$. This shows that not a lot of interest was placed on moderation of tests though most teachers really show less importance to moderation thinking it is there standard being tested.

4.4.6 Length of the tests

As the length of a test increases the validity and reliability also do improve. This is shown by Table 17. The more the number of items in an exam the higher the level of reliability and also validity.

Table 17 Length of tests and level of validity and reliability.

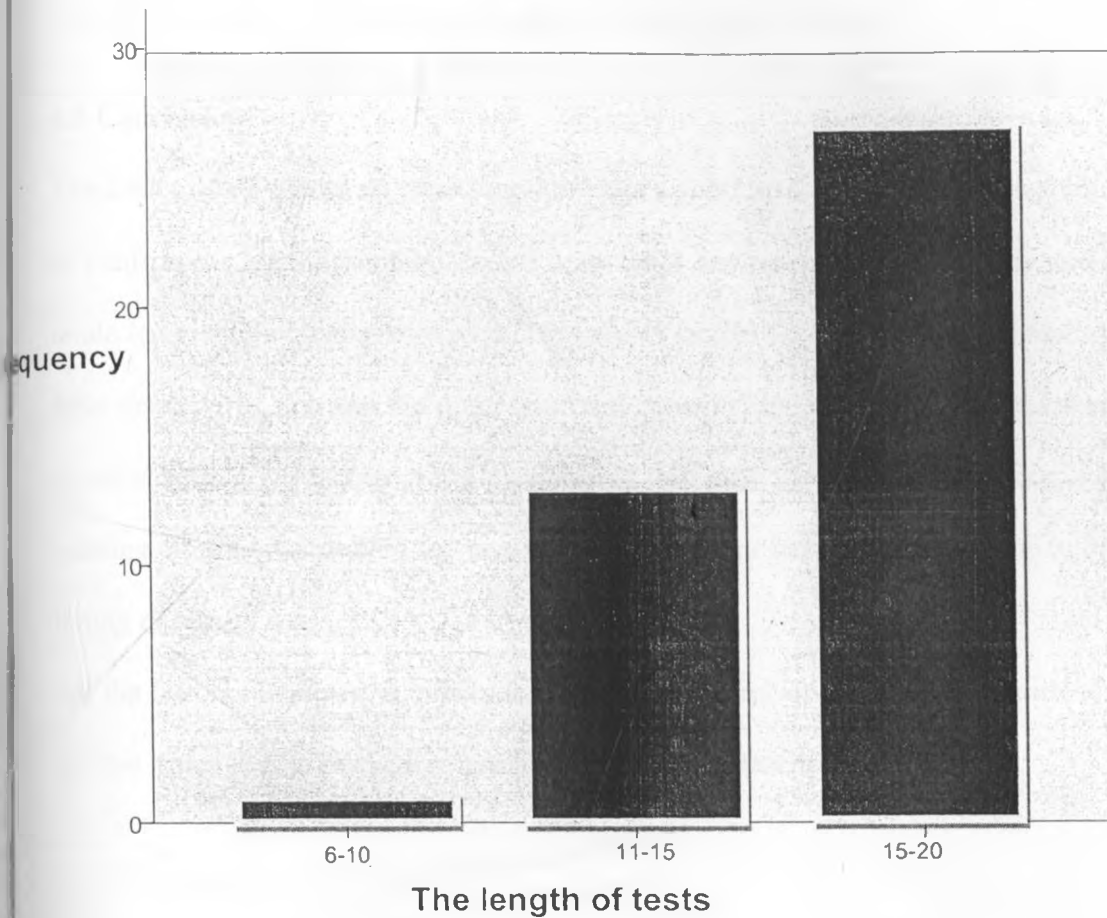
Length of the	Number								RELIABILITY
	of items	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	(Kuder-Richardson method)
1	3	3	3	3	3	3	3	3	0.6
2	3	4	3	5	4	3	3	3	0.68
3	3	5	4	5	4	3	4	4	0.8
4	3	4	4	4	4	4	4	4	0.8
5	5	5	5	5	5	5	5	5	0.8

The respondents were asked how many items they normally use in their exams. Table 4.18 shows the results obtained for the number of items used in an examination by the respondents.

Table 18 Results for the number of test items used by respondents

	Frequency	Percentage
1-5	0	0.0
6-10	1	2.4
11-15	13	31.7
15-20	27	65.9
>20	0	0.0
Total	41	100.0

Graph 7. Summary of length of tests



Most of the respondents construct tests with 16-20 items (65.9%), followed by 11-15 items (31.7%) and finally 6-10 items (2.4%). The key informants also felt that the number of items have an effect on the validity and reliability. When asked whether the number of items affect validity and reliability the key informants had a plain answer of yes it does. Most of them felt that the number if items served to remove or to a great extent reduce the examiners biases and

also encourage the learners. Though this is true, it is also better to consider the process of coming up with the questions so that it will not be just for the sake of having many items in the tests.

$\chi^2=24.78$ at 0.01 and $df=4$. This is an indicator that most teachers place a lot of importance on the length of test to as a factor that increases validity and reliability.

4.5 Conclusion

The tests made by teachers were found to have a good level of validity and reliability. This goes to confirm the hypothesis that teacher made tests are valid and reliable. The quality of teacher made tests can be greatly improved if the above factors are taken into consideration. Also more time needs to be put into the process of test construction and analysis by the teachers. It was established that the factors that play a role in this include teachers experience, education level, training on test construction and analysis, use of Bloom's taxonomy, moderation of the tests and length of test.

All the factors interacted to influence the validity and reliability of teacher made tests. There is a lot that also needs to be done to maximize the level of teacher made tests.

CHAPTER FIVE

SUMMARY OF FINDINGS, CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a summary of the findings, conclusions and recommendations. They are presented under the following thematic areas i.e. objective of the study, internal and external validity of research instruments, reliability of research instruments, methodology, findings of research questions, contributions to the body of knowledge and suggestions for further research.

The objective of the study was to determine the factors that affect validity and validity of teacher made tests. Also the study was interested in finding out the level of validity and reliability of various tests constructed by teachers visa vie the factors that affect validity and reliability.

5.2 Methodology

Since this study was a case study the method selected was for a case study. The purpose was to determine factors and relationships among the factors that would result in validity and reliability of the tests. Data was collected carefully using subjects who were carefully selected to ensure that the results can be generalized to the whole population.

The method used to collect data was descriptive survey since it helps to obtain primary data required to determine the characteristics intended for the teacher made tests. This method enabled the researcher to get the required data easily and also make interpretations of the observations made. The researcher visited the schools to collect the needed data through questionnaires and there was total support from the teachers in providing the required

information and filling in the questionnaires. Also the teachers readily provided their exams for term two. When asked to provide marks for their students most of them were suspicious of the intentions of the research and they required a lot of information on the use to be made of the marks. The key informants were more than willing to take part in the discussion and they provided very valuable information to this study.

The tests were directly analyzed to determine their degree of content-related validity and the internal consistency reliability of their scores. The statistical program for social sciences (SPSS) was used for all the statistical analyses performed on the tests.

5.3 Justification of the Methodology Used

Marso and Pigge (1992) reported that much of what is known about teachers' testing practices have been obtained through studies using teachers' self-report data-gathering procedures. These self-report studies provide a valuable but limited understanding of teachers' actual testing knowledge and skills. This study was conducted to provide a better understanding of teachers' tests and add more knowledge to the existing literature on direct observations of teacher-made tests. To increase homogeneity in the research, the study was limited to one year level and one subject area.

5.4 Reliability of the tests

A Kuder-Richardson coefficient was calculated for the results of all the tests. All the Kuder-Richardson coefficients are above 0.6. The highest is 0.89 and the lowest is 0.60. The mean Kuder-Richardson coefficient for all test results is 0.78. This indicates that in general the test results are highly reliable. High internal reliability indicates that items of the test consistently measure the same ability.

5.5 Validity of the Tests

Two teaching colleagues and the researcher, labeled as Judge 1 and Judge 2, and judge 3 rated the validity evidence of the tests using a scale as indicated in data analysis technique. A rating of 1 indicates very validity on a given criteria and a rating of 5 indicates very high validity on the given criteria for validity. The criteria used to assess a particular test were consequences, content validity, generalizability, cognitive complexity, meaningfulness, fairness and cost and effectiveness.

5.6 Findings on research questions

5.6.1 Does teachers experience affect reliability and validity of their tests?

This study found that the quality of teacher made tests in terms of their validity and reliability is affected by teachers' experience. The teachers with more experience prepared tests which were more valid and reliable. This agrees with the findings of others. According to a study by Carlo Magno (2003) the more experienced teachers prepared exams with high validity and reliability. This agrees totally with the findings in this study. Teachers with long years of teaching and testing may also have gained the appropriate skills and knowledge in test construction.

5.6.2 Does the level of education affect validity and reliability of teacher made tests?

The findings of this study were that the education level of the teachers also affected the validity and reliability of the tests. Teachers who are trained on test construction and analysis prepared tests that were more valid and reliable. According to literature the best trained teachers on test construction and analysis prepare tests that have a high degree of reliability and validity. According to Stiggins,1994 the education level affects reliability and validity of teacher made tests. Marso and Pigge (1992) argue that lack of planning for good tests is due to lack of training. This agrees with the findings of this study in that the better one is trained then their exams are more valid and reliable.

5.6.3 Does training on test construction and analysis affect validity and reliability?

This study established that the training one has had on test construction and analysis affects the validity and reliability of the tests constructed by teachers. Carlo Magno found out that the training one has on test construction is important in determining the validity and reliability of the tests. This agrees with the findings if this study. Thus the findings can be generalized to all the teacher made tests.

5.6.4 Does moderation of tests affect validity and reliability of the tests?

The moderation of the tests which was also found to have a direct relationship with validity and reliability. This study found out that moderated exams have a high level of reliability and validity. This emerged as a by the way but later became an important factor affecting validity and reliability. Moderation of tests had a $\chi^2=4.976$ at 0.01 significance and $df = 4$ and therefore proved to ba an important factor affecting validity and reliability.

5.6.5 Does use of Bloom's taxonomy have effect on validity and reliability of tests?

Use of Bloom's taxonomy was found to have a direct influence on validity and reliability. The exams given by those respondents who used the table of specification were found to have high validity and reliability. The findings of Gay (1991), Linn and Gronlund (1995) were that it is important to plan using the table of specification in order to ensure proper sampling of items to meet conditions of validity and reliability.

5.6.6 Does the length of test affect its validity and Reliability?

The length of the tests also affected the quality of the tests. According to Craig S. Wells and James A. Wollack longer tests produce higher reliabilities and validities. Thus the longer an exam is the more likely that it will be reliable and valid. This trend was also observed in this study such that most respondent agreed with the fact that the length of tests affected their reliability and validity. Indeed most teachers construct exams with many items in order to increase there reliability and validity and this was observed in this study. The teachers who considered these factors well in preparing tests, their tests they have a high level of reliability and validity.

5.7 Contributions to the body of knowledge

This study has made a great contribution in that the factors affecting reliability and validity are now well known and documented. The study also related each of the factors with their level of reliability and validity. The knowledge of relationship between each of the factors on validity and reliability was established. These factors can be arranged in order of importance as shown by their χ^2 values as use of Bloom's taxonomy, training on test construction and analysis, length of tests, education level, experience and moderation of tests in that order.

5.8 Recommendation

The following recommendations were made in order to improve the validity and reliability of teacher made tests.

1. Because of varied experience levels teachers should be encouraged to consult with others in order to improve the validity and reliability of their tests. This will enhance collaboration which will go a long way in improving the quality teacher made tests.
2. There should be in service training of teachers on test construction and analysis. Most of the respondents agreed that they need further training on test construction and analysis. The government should be advised to come up with means of ensuring that teachers improve there skills in test construction and analysis.
3. Before tests are administered to students there is need for moderation to be carried out by the departments. This enhances the level of validity and reliability of teacher made tests. Also the input of others is important in improving the quality of teacher these tests.
4. Much time should be invested in the process of constructing tests. This will serve to enhance quality of the tests and at the same time serve to ensure that all factors that affect the validity and reliability are put into consideration.
5. Institutions that train teachers need to improve their training on tests and measurement. The trainee teachers need to be given more skills in the process of test construction and analysis.

5.9 Practical use of the findings

The findings in this study are useful in determining the level of validity and reliability of teacher made tests. This study is useful to teachers and other stake holders in education sector who can use it in evaluating appropriateness of teacher made tests.

5.10 Suggestions for further research

The following areas are suggested for further research:-

- i. To determine the impact of teacher valid and reliable teacher made tests on the performance of the candidates at higher level of education.
- ii. To establish the item discrimination and difficulty indices of teacher.

REFERENCES

- Carol A. Dwyer**,(1998) Assessments in Education. New Jersey:Prentice-Hall inc.
- Carlo Magno**, (2003), Level of Appropriateness in Test Construction. UPHL institutional Journal.
- Crooks; Hills; Stiggins, Griswold, and Wikelund**, , Psychology Applied to Teaching, Houghton Mifflin co. (1997)
- Green and Stagger** (1986/1987) Psychology Applied to Teaching, Houghton Mifflin co. (1997).
- Gronlund, N.** (1990).Measurement and Evaluation in Teaching (6th ed).New York: Macmillan Publishing Company.
- Harold, G.L; Christine, H.M.; Leroy, W.N.;** (1970). American Education Research Journal, (1970). Amarican Educational Research Association.
- Hunter and Schmidt F.L.** (1990), Methods of Meta analysis.
- Justin, D.V; John, R. G;** (1996).The reliability and Validity of tests constructed Seychelles teachers, A Paper, Austrarian Association for Research in Education.
- Linn, R. L & Gronlund, N. E.** (1995). Measurement and Assessing in teaching (7th ed).New Jersey:Prentice-Hall inc.

Moskal, Barbara M. and Jon Lee, Scoring Rubric Development: Validity and Reliability. A

peer reviewed electronic journal ISSN-7714-2000.

Mwangi Ndirangu , A study of the Psychometric Quality of Continuous Assessment Tests and

their suitability for Contributing to the Final School Examination Scores, Egerton

Journal, Humanities and Education, 2010.

Peterson, L. P. and Wablg, J. H. (eds).(1970).Research on Teaching: Concepts, Findings and

Implications: Berkley: Mccuthan Publishing Corporation.

Stiggins (1994), Psychology Applied to Teaching, Houghton Mifflin co. (1997)

W. James Popham, Why Standardized tests do not measure educational quality.

Worthne, B.R., Borg, W.R., and White, K. R. (1993). Measurement and Evaluation in the

school. N. Y, Longman.

Worthen et al, (1993). Interpreting Test scores

APPENDIX 1

Daniel Kiragu Kinyua

Masters student

University of Nairobi

28/08/2012

Dear respondent

I am a post graduate student at the University of Nairobi.

In order to fulfill the requirements for the award of masters degree in education (measurement and evaluation) at the University of Nairobi, am conducting a research study entitled; Validity and Reliability of teacher made tests in Kenya: A case study of Physics form three in Nyahururu District.

You have been selected to assist in providing the required information as your views are considered important in this study.

I am therefore kindly requesting you to fill this questionnaire. Please note that information given will be treated with utmost confidentiality and will only be used for the purpose of this study.

Thank you in advance for accepting to take part in this study.

Yours faithfully,

Daniel Kiragu kinyua

APPENDIX 2

QUESTIONNAIRE FOR THE PHYSICS TEACHERS IN NYAHURURU DISTRICT

PART A (Please tick and fill in where appropriate)

Date: _____

My Gender : Male Female

My age is: 21-30 31-40 41-50 51-60

School: _____

I have been teaching for:

1-5yrs 6-10yrs 11-15yrs 16-20yrs >20yrs

I live near the school

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

I have a cordial relationship with other members of staff (teaching and non-teaching)

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

My level of education is

Certificate	Diploma	Bachelors	Masters	Doctorate

PART B

1. Level of training on test construction

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The training I have had so far on test construction is adequate					
I would like further training on the process of test construction					
I would like further training on test analysis					
I have attended in service training on test construction					
I am able to calculate item					

difficulty and discrimination indices of the tests I give					
The training I have had so far on test construction is inadequate					
I normally analyse my examinations after scoring					
We usually moderate our tests					
I was inducted in test construction					
I am aware that moderation of tests is important					

2. Use of Bloom's taxonomy

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I am aware of the existence of Bloom's taxonomy					
I am aware of the levels of Bloom's taxonomy					
I often use Bloom's taxonomy in preparing the exams I give					
Most of the questions I give are in low order level of Bloom's taxonomy					
Most of the questions I give are in the high order level of Bloom's taxonomy					
I never use Bloom's taxonomy in preparing the exams I give					

3. Length of the test

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The exams I give have less than ten questions					
I usually give an exam with between ten and fifteen questions					
I take a lot of time to prepare the exams I give					
The length of tests I give is adequate					
I give an exam with more than fifteen questions					
I am aware that the length of a test affects its quality					

APPENDIX 3

QUESTIONS FOR THE KEY INFORMANT

1. Does the training one has had on test construction affect the quality of Physics exams form three in your school according to you?

2. Are you aware that the length of a test affect validity and reliability of teacher made tests?

3. Are teachers able to identify objectives as per Bloom's taxonomy in their exams?

UNIVERSITY OF NAIROBI
NYKUYU LIBRARY

4. Do teachers use Bloom's taxonomy in preparing their exams in Physics form three?

5. Do you carry out induction in test construction in your school?

APPENDIX 4. DETERMINATION OF SAMPLE SIZE FOR RESEARCH ACTIVITIES

Table for Determining Sample Size from a Given Population

<i>N</i>	<i>S</i>	<i>N</i>	<i>S</i>	<i>N</i>	<i>S</i>
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	357
100	80	500	217	6000	361
110	86	550	226	7000	364
120	92	600	234	8000	367
130	97	650	242	9000	368
140	103	700	248	10000	370
150	108	750	254	15000	375
160	113	800	260	20000	377
170	118	850	265	30000	379
180	123	900	269	40000	380
190	127	950	274	50000	381
200	132	1000	278	75000	382
210	136	1100	285	1000000	384

Note.—*N* is population size and *S* is sample size.

Adapted from R.V. Krejcie and D.W. Morgan, "Determining sample size for research activities". Educational and Psychological Measurement, 30(3) By sage publications, inc