RELIABILITY AND CONTENT VALIDITY OF COMMERCIAL TESTS AND THEIR CORRELATION TO PUPILS' PERFOMANCE IN MATHEMATICS KAJIADO COUNTY, KENYA.

LILLY KAVUTHA SANGALE

UNIVERSITY OF NAIROBI
SCHOOL OF EDUCATION
DEPARTMENT OF PSYCHOLOGY

2014

A Research Project Submitted in Partial Fulfillment of the Requirements of Masters Degree in Measurement and Evaluation University of Nairobi

DECLARATION

This research project is my original work and has not been presented for an award of any degree in any university.

SIGN …….. DATE……

Lilly K. Sangale
E58/73623/2012

This research project has been submitted for examination with my approval as University of Nairobi supervisor.

SIGN…… DATE……

Dr. Luke O. Odiemo

Department of psychology

School of Education.

University of Nairobi

DEDICATION

I dedicate this research project to my late dad Joel Mulwa who instilled in me a strong moral compass and encouraged me to pursue my studies to Doctorate level.

ACKNOWLEDGEMENTS

First to The Almighty God, for the gift of life, time, strength and resources, I give you thanks. To my supervisor Dr. Luke O. Odiemo, for his fatherly guidance, scholarly and constructive advice throughout the project may God bless you .I also wish to express my gratitude to all lectures who taught us different units with great commitment, again the tireless effort of the coordinator Dr. Karen Odhiambo. To my fellow colleagues of measurement and Evaluation thanks for the new ideas and encouragement throughout and especially when I was unwell and hospitalized, God bless you all. My appreciation goes to my dear family for being there for me, to my husband Shadrack Sangale who assisted me morally and financially while also taking care of our young children- Salaash, Salaton and Sanaipei who never understood why mummy was ever saying "naenda shule". Am so much grateful to Cephas Mwaura commonly known as "Ceph" of Cybertron Kitengela for his recommendable job in typing and editing this document. To the entire Nkuos family who supported me in prayers and encouraged me through all my studies, God bless you.

ABSTRACT

This study was carried out to establish the reliability and content validity of commercial tests and their correlation to pupils' performance in mathematics: the case of public primary schools, Isinya District, Kajiado County. It involved nine teachers who were the key informants selected from different schools and were actively involved in mathematics teaching in class 8 or held responsibilities like mathematics panel heads, in their schools in Isinya District. The teachers analyzed the nine commercial papers used for research in this project, and also indicated the Mathematics mean scores attained by the class 8 pupils in the years 2013 and 2012. A mixed descriptive survey research design was applied and data was collected through questionnaires with key informants. Quantitative data analysis was applied to survey data collected via questionnaires. The frequency distribution was described while data from questionnaires was qualitatively analyzed using tables and Fleiss Kappa method. The findings of the study revealed that the poor reliability and content validity of commercial papers compared to K.C.P.E, had an effect on the subject mean scores leading to wrong conclusions' by the mathematics teachers or the school administration. Other factors like training in test construction have a varying influence on the validity and reliability of Commercial tests. Experienced teachers who had prior training in test development (33.3%) applied a number of this knowledge when analyzing the commercial papers. This was evident from the findings of the study that commercial papers 006,008 and 009 score rated low with k= 0.115, 0.157 and 0.162 respectively. The commercial exams mean scores did not have any correlation with the standardized exam K.C.P.E which had an average mean difference of 6.2%, it was concluded that commercial tests are generally invalid and unreliable. Due to lack of test construction knowledge by teachers as revealed by the study (66%). This is a contributing factor for the teachers choosing to use commercial tests. The study recommended research could be done on reliability and content validity of commercial papers and their correlation to pupils' performance in other subjects. A further research may be done to unveil the reasons that hinder public schools from purchasing standard set exams.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background

In many learning institutions, tests are administered by teachers and instructors to their learners for different purposes, like tracking and selecting students for promotion to next grade, admission into academic secondary schools or universities. (Ghosh, 2004).The "vision 2030" in view of education matters is to have globally competitive quality education, training and research for sustainable development, Government Of Kenya,(2007).Therefore to achieve this among the many things to put in consideration is assessment. Mwanzia and Miano, (2007) point out that for assessment to play a role in fostering quality education, it must pay attention to the goals of education in terms of what is taught and learnt and the levels at which the knowledge and skills are assessed.

Assessment is therefore an integral component of learning and teaching. Harlen, (2005, p.207) views assessment to be all processes employed by academic staff to make judgment about the achievement of students in units of study and over a course of study, these processes include making decisions about what is relevant evidence for a particular purpose, how to collect and interpret the evidence and how to communicate it to intended users (students, parents, university administrators, and others). According to one of the Canadian provinces Manitoba, the ultimate goal of assessment is to help develop independent, lifelong learners who regularly monitor and assess their own progress. (Manitoba Education and Youth, 2003).

Having assessment as a very important aspect in mind, teachers should therefore consider the type of test tools to use in the assessment process. In France there is the

‘mirror effect’ theory devised by Thelot, 1998, cited by Pons, 2008, explained that standardized assessment should have a ‘mirror effect’. This means that the assessment should confront players in the education system with the results of their actions, but need not necessarily provide them with explanations. Teaching staff need feedback on their methods to be able to make improvements, if they have not achieved their intended aims. ‘This mirror effect must be achieved providing results without necessarily providing explanations, as those are not always available.’(Theolet, 1998).

The mirror effect model is based on purely symbolic sanctions. Having this as a bases of teachers having summative exams for predictive purposes for the formative exams, then the summative assessment should be reliable and have the right content validity to be fit to have the mirror effect. According to Blue Paper I 2002, when planning an assessment strategy, there are a variety of issues to be considered, first being understanding that assessment is a form of communication which is to the learners, teacher, to the curriculum designer, to the administrators and to the employers. Teachers use assessments and evaluation to-:provide student and parents with ongoing feedback, plan further instructional and learning activities, set subsequent learning goals and identify students who may require intervention,(British Columbia Ministry of Education,2004).

Tests are tools of assessment used to give feedback to learners themselves and teachers. Classroom assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve student’s achievement of intended instructional outcomes. (Popham, 2008) Classroom assessment provides valuable information that allows teachers to adapt instructional procedures, valuable information that allows teachers to adapt

instructional procedures to the learning needs of their students (Kovalik, 2002 as cited by Eggen, Kauchack, 2004). To learners, assessment increases motivation by helping them learn more and better their current grades attained.

In Kenya, primary school education takes 8 years where each level takes 1 year and after administration of Continuous assessment tests (CATs) and an end of year exams by teachers, learners proceed to next level. Kenya Certificate of Primary Education (K.C.P.E) is a high stakes exam offered by the Kenya national Examination Council (K.N.E.C) at the end of the 8^{th} year and this test is used to admit learners in secondary school. Teachers use different types of formal assessment (FA) tests which include; final exams, mid-term exams, end of unit tests, quizzes and so on. Teachers also may make decisions to have teacher-made tests, which are set by subject teachers themselves and administered to the classes the teacher teaches.

Also, subject panel can set as a group and administer to learners in different classes, the challenge here is, though these multiple choice items can be scored easily and objectively but are difficult to prepare. Essay tests on the other hand are easy to prepare but difficult to score (Elliott, Kratochwill, Cook & Travers, 2000) this therefore questions the reliability and validity of these teacher-made tests. According to research report by Strengthening Mathematics and Science Education (SMASSE) (1998) reasons for poor performance in mathematics included teachers use of poor testing instruments, teachers use of inappropriate teaching methods, negative attitude towards the subject by students and lack of resources among others.

Commercial tests are also used widely by many mathematics teachers to serve as test tools for end month tests, mid-term tests or end of term exams. This could be according to Sharky and Murnane (2003) teacher's lack of knowledge and skills on

how to design a valid FA test and how to make inferences about students' knowledge and skills from the results of a well-developed assessment. This was not left out teachers in Isinya District of Kajiado County where these commercial tests are highly purchased and used as tools of assessment. Summative Assessment (S.A) tests like the K.C.P.E are typically used to evaluate the effectiveness of instructional programs and services at the end of an academic year or at a pre-determined time.

Table 1.1.: Formative and Summative Tests in Primary Schools in Kenya.

Formative tests	**Summative tests**
Mid-term and end month tests	Final exam
Quizzes and essays	National exams (K.C.P.E, K.C.S.E)
Diagnostic test	Entrance exams

In the United Kingdom the high stakes "Examinations such as GCSE, A levels, Scottish higher, the welsh Baccalaureate or national curriculum tests are administered, in the U.S.A different Examinations are used by different states and there are various examination bodies monitoring these exams, these assessments to list a few include; in California standardized tests and reporting (star) and California High school exit exam (CAHSEE) New Hampshire, New England common Assessment program (NECAP) and in New York they have pegents examinations. Math's and English languages arts performance assessment.

In Africa countries like South Africa (S.A) the education system takes 12 years of formal schooling and National senior certificate, (NSC) examinations commonly known as matrix are done. In Kenya there are two national exams being the K.C.P.E in primary school level and K.C.S.E in O level secondary school. Higher education

institutions like universities and colleges offer examinations according to the courses being taken by learners.

According to Stiggins et a;l (2007) there are two kinds of assessment during instructions; assessment for learning involves use of homework, assignments, quizzes and self-assessment drafts, this kind is child centered and gives learner an opportunity to find information about areas of strength and areas of further learning. Assessment like mid-terms and final exams which are teacher centered and judgmental are meant to inform the final grade of the learner. In Scotland the development of a coherent assessment system, assessment for learning has been hence a government priority since 2001, adopting the Stiggins view of assessment for learning.

The Scottish government no longer collects information on all pupils through national assessments but does monitor achievement through the Scottish survey of achievement sample survey Whetton, C. (2009). Though testing is of great benefit to education, the validity and value of the standardized tests are being debated. Studies done by (cannel, 1987, Linn, crave and sanders, 1989; Shapen 1990) raise questions about whether improvements in test score performance actually signal improvement in learning. As in the case of formative assessments which teachers use as assessments for learning in most of the public and private primary schools in Kenya, leave a lot to be questioned in terms of their reliability and validity.

According to Hogan T.P. 2007, reliability is a measure of degree of consistency with which candidates'' responses to an assessment are judged. To be reliable, then standards should be maintained when making decisions on candidates' performance across all assessors for all candidates undertaking the same assessment task. Assessment decisions are reliable when they are generated by valid assessments

which are produced under conditions of assessment that are consistently applied. The decision should also be taken on the basis of clearly defined standards of performance and the authenticated work of the candidates is being assessed. (SQA, Guide to Assessment, 2009).

Validity is about whether the assessment measures all that it might be felt important to measure, (Edward. Carmine, 1979), (Satheesh Kumar January 2.2008- Educational Journal) also defines validity of test to be extend of a test measuring what it intends to measure, where predictive validity, content validity and construct validity should be clear. The test administrator may also be responsible for preparing the students for the assessment. For students to give their maximum performance on an educational assessment, those who are responsible for administering the assessment must provide them with basic information that they require, including when the assessment will be administered, the content and abilities that will be assessed, what the assessment will emphasize, the standard or level of performance expected, how the assessment performance will be scored, and how the results of the assessment will be used to make decisions, (Mehrens& Lehmann, 1991; Nitko, 2004).

Any one administering a test should be careful about its validity just in case the exam program is ever challenged is a court of law this would be best defense of professional testing. Another important factor to be considered in assessment is time. Mathematics in Kenya is a prerequisite subject to many other later careers in life like Engineering, medicine, and other business courses like accounting, finance and banking, this therefore mostly forms a basis for one to join these courses high scores to the subject are a requirement (university of Nairobi 2008) but the subject has been faced by poor performance which is a concern not only in Kenya but globally. According to a

research conducted by African Population and Health Research Centre, (APHRC, 2010).policy brief no.18 indicated that mathematics is the most poorly done subject at primary level.

The Kenya Certificate of Primary Education (K.C.P.E) 2008 average score for the subject was a D+, according to the Kenya National Examination Council K.C.P.E 2010 analysis report the subject mean in 2009 was 24.78, 2010 was 26.90 2011 was 25.45 and 2012 was 26.87. APHRC (2010) according to their study pointed out that the mathematics teachers did not have mastery of the subject having the lowest to have scored 17% and the highest 94% hence poor scores were associated with teachers. The job of a teacher is to impact knowledge, skills and attitudes and mathematical concepts into the learner, therefore mathematics learning depends on the teacher and not the testing (Onwukapa and Nweka, 2000) to achieve this, the teachers should give assignments and tests to pupils but should later discuss the test results.

An article captured in the Daily Nation 23rd April 2010 entitled" children troop to school, but still illiterate" a study done by Non-Governmental organization "Uwezo" covering 70 districts and 40,386 pupils were interviewed revealed that one out of 10 standard eight pupils would not solve a class two mathematical problem, 30% of class five failed the sum and 20% of class two were able to solve it. This is a wake-up call to mathematics teachers to change the methods of instruction and concentrate on teaching pupils more on how to solve mathematical problems rather than pumping examinations to them which are rarely revised or never at all.

1.2 Context of the study

The researcher in this study has a lot of concern as it Pertains the focus put on test scores rather than the quality of the assessment. A lot of time is devoted in using these commercial exams which sometimes are randomly purchased without scrutinizing the reliability and content validity of the exam. For instance, immediately learners open the school in Isinya sub-county, class eight pupils in the public primary schools sit for an opener exam and the scores attained are compared with previous end term scores. This makes teachers make wrong judgments because these opener exams may have content which learners are expected to cover that term.

The same case as in Texas and Chicago as found out in one study by Jones (2007), he noted that some teachers devoted large amount teaching time to coaching; using exercises similar to those that will appear is the tests. This new behavior pattern has been the focus of studies in the United States (Jones 2007) and the United Kingdom (England) but of which use high-stakes testing. According to parliamentary report on testing and assessment (House of commons,2007) a study by the royal society in 2003 indicated that there were large varieties in the amount of time devoted to tests in the United Kingdom (England).Where there is extensive testing.

The same report calculated that in the spring term, 70% of primary schools spend three hours a week on teaching to the key stage test taken in year six. In Isinya teachers are accountable of the scores attained by candidates in subject scores. For example, after the mock analysis of 2nd term 2013, the results were compiled and released by district committee. The following was Mathematics subject panel comment on the mathematics district mean score, addressing the mathematics teachers.

The mathematics mean score recorded by the public schools of 48.25 as compared to that of private schools 71.45 is miles away and no explanation can be given. Mathematics teachers should therefore seriously think of the right approaches to use and improve this score. Teachers will be accountable for their results.

1.3 Statement of the problem

Pupils' performance in the mathematics subject in Isinya District- Kajiado County has not been satisfactory, both at the county level and national exams. This is an investigation on the reliability and content validity of commercial assessment tests which are widely used in the district to serve as formative assessment tests. In addition it unveils the correlation of performance between the formative and summative assessment tests. In actual sense, there is poor reliability in the commercial tests and hence one cannot compare summative assessment scores and formative assessment scores.

Content validity in the commercial tests has not been given consideration as it should, Kerlinger, (1986) pointed out that content validity regards the representativeness or sampling adequacy of the content of a measuring instrument and its always guided by a judgment's the content of the measure representative of the universe of the concept being measured. Initially, Isinya District was part of Kajiado North District and therefore the young district has got 29 public primary schools and more than 45 private primary schools. Isinya Districts has registered K.C.P.E candidates for three years now - (2011- 2013).

Teachers use varied ways in testing in preparation of the class eight candidates awaiting the (K.C.P.E) homework is given using exercises in pupils text books, structured questions and the commercial tests are used widely as weekly exams, mid-term exams or end term exams. The table below shows the mean scores of the formative assessment tests (Isinya District Mock) which was a commercial exam, to the high stakes exams (K.C.P.E) 2011 and 2012 in mathematics

Table 1.2: Isinya District Mock and K.C.P.E Results.

Year	**Isinya District mock Mean scores**	**K.C.P.E Isinya District Result Mean scores**	**Expected mean**	**Deviation**
2011	66.30	53.91	100	+12.39
2012	68.57	55.82	100	+12.75
2013	68.25	55.667	100	+12.59

Source: DEO Isinya District 2014

From the table above, observations made that the districts mean scores reduced by the formative assessment exam indicated a very high score yet in the high stakes (K.C.P.E) the exams were far below what the district exams indicate. This indicates that the commercial tests used as a mirror to the high stakes exams were not at all a good tool to be used by the mathematics teachers. Mathematics curriculum has been reviewed over time hence mathematics tests should be set in line with the changed curriculum. The curriculum emphasizes mathematical thinking and reasoning, conceptual understanding and problem solving in realistic contexts (curriculum and evaluation standards for school mathematics).

There is a lot of testing done in schools, yet too little use of a good test that is valid and reliable (Musau, 2004).According to Smarter Balanced mathematics item specifications high schools (April, 2012) mathematic items/ tasks usually take the

form of graphs, models, figures e.t.c.so, students must read and examine in order to respond to the item or tasks. Therefore selected response (SR) items include a stimulus and stem followed by three or five options from which a student is directed to choose only one or best answer, by redesigning some SR items, it is often possible to both increase the complexity of the item and yield more useful information regarding the level of understanding about the mathematics that a student's response demonstrates. This is what is adopted by the commercial tests.

The above being the right way however, according to Black and William (1998) most of the tests are designed and developed by unskilled teachers/people hence contains poorly focused test items. This is because their responses require factual knowledge but lacks high order cognitive skills. A study done by Schmidt, et al (2002) found out that teachers in U.S.A follow textbooks which are too wide because publishers produce elementary mathematic textbooks that cover a variety of topics so that they can sell in different states. As a result teachers do not develop in their pupils a deep conceptual understanding of mathematics topics and their application (Schmidt Houang and Cogan, 2002)

According to Professor Kiptoon, former Secretary in the Ministry of Education (MOEST, 2001) claimed that the poor performance in primary mathematics was caused by teachers who lacked the subject knowledge, incompetent and were unskilled, hence the government in 2001 through the Ministry of Education, (M.O.E) introduced distance learning course called School Based Teachers Development. Due to the continuous unsteady scores recorded in the mathematics scores in Isinya District observed from the previous two years there is need to adopt and continuously

use the right tool for testing learners in order to have the reliable and valid results which can be later used for improving mathematics in the district.

1.4 Significance of the study

The study brings out the importance of using reliable test tools in examining the subject of mathematics. The findings of the study will provide useful information for adaptation by mathematics teachers, the policy makers' examination boards, and the K.N.E.C officers is Kenya making the necessary interventions in the area of testing in mathematics. The private firms also involved in the commercial exams setting, could use the findings of the study and embrace the changes to have quality mathematics examinations.

Further research could be undertaken to establish whether commercial examinations firms use trained personnel when setting their examinations hence solve the problem of poor performance in Mathematic subject in public examinations.

1.5 Main objectives

The main objective of the study is to investigate the reliability and content validity of commercial tests and their correlation in pupil performance in mathematics in Isinya District.

1.6 Objectives of the Study

i. To find out the reliability and content validity of mathematics commercial tests used in public primary schools in Isinya District
ii. To investigate the teachers rational for choosing commercial exams

1.7 Research Questions

i. What is the reliability and content validity of mathematics commercial tests used in assessment of pupils in public primary schools in Isinya District?

ii. Why do mathematics teachers opt to choose mathematics commercial tests?

1.8 Research Hypothesis

i. The aspect used in designing the mathematics instrument is reliable.

ii. Teachers rely on commercial tests because they are readily available.

1.9 Basic Assumptions of the Study

The study will be based on the following assumptions: that all mathematics teachers in all schools conducted for the study will be cooperative and give correct information; that the sample to be taken will be a true representation of the target population; that all mathematics teachers have used commercial tests in testing learners.

1.10 Limitation of the Study

According to Best &Kahn, (1998) limitations are conditions beyond the control of the researcher, hence providing restrictions on the conclusions of the study and their application in other areas. The study will be carried out across Isinya District and most of the public schools are found in interior parts of the district which would force the researcher to travel to many kilometers using motor bikes incurring a lot of finances. The researcher will come up with a manageable budget to meet all the expenses and therefore the limitation will not interfere with the study.

1.11 Delimitations of the Study

Only public primary schools in Isinya District which have had candidates for the last three years will be focused on, mathematics teachers who handle class eight and the mathematics panel heads also will be considered much. Isinya District is a young district from the larger Kajiado North District, therefore this district was chosen because apart from mathematics most of the subjects are still below average, and this might be as a result of time taken to embrace change.

1.12 Organization of Study

The study will be organized in five chapters; chapter one contains the background to the study, the statement of problem, purpose of the study, objectives, research questions, significance, basic assumptions, limitations, and delimitation of the study, lastly operational definitions of various terms used in the study are provided. Chapter two is a review of the related literatures derived from relevant studies carried out on standardized tests and the principles of good tests/assessments administered to learners. Also included will be reliability and content validity of assessments, factors which affect both reliability and content validity and how these factors are controlled. The interaction between reliability and contentment validity in relation to performance, repetition socioeconomic status and gender and age will be intensively discussed. Chapter three of the study will contain the introduction, then discuss the research design, target population, sample and sampling techniques, research instruments, pilot testing, instrument reliability, instrument validity, data analysis and the ethical consideration.

Chapter three of the study will contain the introduction of the study, research design; target population, sample and sampling techniques, research instruments will be

clearly explained, pilot testing, instrument reliability, instrument validity, data analysis and the ethical issues will all be discussed.

Chapter four of this research project will have the introduction, data analysis and interpretation of findings which will be presented in tables and figures.

Chapter five of the research project will have the introduction, aim of the study, summary of findings, policy recommendations, suggestion for further research and the conclusion.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter highlights on literature reviews on qualities of a good test, definition of reliability, types of reliability, factors affecting reliability and how they are controlled. Definition of content validity, how it is measured, factors that affect content validity will also be discussed and how they are controlled. The interaction between reliability and validity of a test in regard to performance will be exhaustively discussed. The theoretical framework relating to the study will be discussed and finally the conceptual framework.

2.2 Qualities of a good test/Assessment

According to Guidelines for Good Assessment practice, (2011)), assessment in education and training is about collective evidence of learners' work, so that judgments about learners achievements or non-achievements, can be made and decisions arrived at. These decisions might be for different purposes which include promotion to next grade, placement, improve on a given programme or the quality of the programme and many others.

Therefore, a good test should have the following qualities; fair, valid, reliable and practical. Also according to No Child Left Behind (NCLB) requirements for an effective test should be reliable, valid and unbiased. Where a reliable test must produce consistent results, a valid test must be shown to measure what it is intended to measure and unbiased test should not place student at disadvantage because of gender, ethnicity, language or disability. Kenya Education management (KEMI) diploma in Education Management module 6- pg 69-70 a good test and examination

should have the following characteristics-: Objectivity, Validity, Reliable, Efficiency and fairness/Equity.

2.2.1 Fairness and equity

A test should not in any way hinder or advantage a learner. This means that the test process is clear, transparent and available to all learners (Criteria and Guidelines for Assessment of NQF). According to (Principles for fair assessment practices for education in Canada) assessment methods should be free from bias brought about by student factors extraneous to the purpose of assessment. It should not favour any group of candidates either because of their gender, social-economic, cultural or tribal background.

2.2.2 Validity

According to Kenya Education Management institute (KEMI) module 6, validity is the extent to which the outcome of the test is a fair measure of what was intended to be tested it's the extent to which a test measures what it purports to measure. A valid test measures what it says it is measuring be it knowledge, understanding, subject content, skills, information behaviors and others (Criteria and guidelines for Assessment, Page 15-19)To achieve validity therefore, the assessors should state clearly the outcome (s) is/are being assessed, use an appropriate type or source of evidence, use an appropriate method of assessment and then select an appropriate instrument of assessment.

2.2.3 Reliable

According to Worthen et al, (1993) reliability in Assessment reliability is about consistency. Consistency refers to the same judgments being made in the same or similar contexts each time a particular assessment for specified stated intentions is administered. Therefore, reliability is all about making sure that test results are not influenced by variables like assessors or bias in terms of learner's gender, ethnic origin, sexual oriental, religion, like/dislike and others, also different assessors interpreting unit standards or qualifications inconsistency or applying different standard or even assessor stress and fatigue. To avoid such variance in judgment, result assessment should ensure that each time an assessment is administered; the same or similar conditions prevail.

2.2.4 Objectivity

This refers to freedom from scoring subjectivity. It requires that the task in the test are more definitive such that the reasons for awarding or withholding a score are obvious to both the pupil and the teacher. There is only one acceptable answer for each question and the score to be awarded for that answer is pre-determined. This means that the pupils score is the same even when the script is marked by different examiners or by same examiner at different times. (K.E.M.I, Module 6).

2.2.5 Efficiency or Discrimination.

According to(K.E.M.I Module 6.Diploma in education management) An efficient test should give a fair distinction between those who are able and who are less able. The test should be able to discriminate well between those who have acquired changes, through educational instruction, from those who have not. So to achieve this test should of moderate level of difficult and aim at testing a wide range of skills.

2.3 Reliability

According to Carmines and Zelle, (1979) reliability is the extent to which an instrument yields the same results on repeated trials, hence, the tendency towards consistency found in repeated measurement is referred to as reliability. Worthen et al, (1993) defines reliability to be how consistent the scores are for each individual from one administration of an instrument to another, hence, reliability is a measure of how stable, dependable, trustworthy and consistent a test is in measuring the same thing each time. In this research, the scores attained by learners in the Formative assessment (FA) tests are to be compared to scores attained by same learners in summative assessment (SA) tests (K.C.P.E).The extent into which the two tests agree on the scores is an indication of reliability. In sum, reliability is consistency of measurement (Bollen, 1989) as quoted by Ellen Drost in Educational Research &perspectives, vol.38, no.1.

2.3.1 Measurement of Reliability

Due to some errors which can be obtained in data, it is important to measure reliability. A research by (Ellen Drost, 2011) on validity and reliability in social science research indicated that there are many ways that random errors can influence measurements in test. For example, number of items in the instruments if they are small, how well the students perform on the test will depend to some extent on their luck in knowing the right answers. (Nunnally, 1978) noted that when a student guesses answers on a test, such guessing adds an element of randomness or unreliability to the overall test results.

2.3.1.1 Test- Retest Reliability

Test-retest reliability is one of the measures to measure some factors which affect reliability. Test-retest reliability refers to the temporal stability of a test from one measurement session to another. This involves administering the test to a group of respondents and then administers the same test to the same respondents at a later date. The correlation between scores on identical tests given at different times operationally defines its test-retest reliability Ellen A. Drost (2011). Despite its appeal, the test retest reliability technical has several limitations (Rosenthal Rosnow, 1991) for instance, when the interval between the first and second is too short, respondents might remember what was on first test, hence affecting the answers on the second test by their memory.

Again when the interval between the two tests is too long, the respondents could have been exposed to things which changed their opinions, feelings or attitudes about the behavior under study. The teachers in Isinya sub-county apply this method where the commercial tests administered to the learners are in series of 001, 002, 003,004 …009 or even 010 all are distributed to cover one academic year. Teachers compare the scores of especially the 3rd term tests and use these scores to focus on the scores which would be attained in the summative exam the K.C.P.E at the end of year.

2.3.1.2 Equivalent Forms

Another technique which is used to measure reliability is equivalent forms also referred to as parallel forms. According to Bollen, (1989) in Social Science research as quoted by Ellen A. Drost, alternate forms are used where different measures of behavior (rather than the same measure) are collected at different times. In educational assessment the equivalent forms of same measure are administered to

either the same group or different groups of respondents, the higher the degree or correlation between the two forms the more equivalent they are.

Michael J. miller (Western International University), Miller also puts across that the administration of equivalent forms is seldom implemented as it is difficult to verify that two tests are indeed equivalent that is, have equal means, variances and correlation with other measure.

2.3.1.3 Split-Half

The split-half approach is another method to test internal consistency. In social science it assumes a number of items are available to measure a behavior, Ellen Drost, (2011). Bowling, (2009) defines it as the extent to which the items relating to a particular dimension in an instrument tap only this dimension and no other. According to Oluwatayo J.A (2012), internal consistency demands that the test, be administered once on the intended group of respondents and their scores collected for analysis using the right statistical tools.

In educational research split-half reliability assumes that the items in an instrument can be split into two matched halves in terms of contents and cumulative degree of difficulty. The advantage of split-half method over equivalent forms and test retest is that effect of memory does not operate here and they are usually cheaper and easily obtained than over time data (Bollen, 1989).

2.3.1.4 Interrater Reliability

There is then the interrater reliability which involves two raters/teachers independently observes and record specific behaviors during the same time period. Their judgments, ratings or scoring should apply same standards of assessment of the responses Michael Miller (Western International University). This is calculated using

the spearman- Brown formula. Ellen Drost, (2011) citing Rosenthal and Rosnow, 1991 pg 51-55. What comes up as a problem in interrater reliability is that teachers can get tired or get bored and randomly start assigning scores, therefore it is important to do this and ensure that two different raters scored the scale using the scoring rules hence attaining same result. This type of reliability is measured by computing the correlation coefficient between the scores of two raters for the set of respondents, for example the correlation of at least .9 is pretty high through what is considered to be accepted vary from situation to situation.

2.4 Factors Affecting Reliability

There are various variables that affect the reliability of a test. These include the length of the test, the interval between the test, environmental factors, the test taker and the quality and type of test itself.

2.4.1 Length of the test

According to Mehrens and Lehmann (1991) and Scattler (2001), there are several factors that affect reliability of a test the first one, is the length of the test where Scattler says that the longer the test the more reliable it is, Lucy Jacobs (1991) referred this as item sampling where she said that because a test is only a sample of all possible items the item sample itself can be a source to Scattler,(2001) says, longer tests are typically more reliable because we get better sample of the course and context and student performance. Longer test tends to reduce guessing.

To the contrary L. Jacobs (1991) pointed out that the problem with longer tests would occur if the additional items are of poor quality this would instead induce error and lower reliability, and also there is a point of diminishing returns. Many items risk student fatigue that will lower reliability.

2.4.2 Environmental factors

The aspect of variation with testing situation or the environmental factors is a factor which can affect, Lucy Jacobs (1991) wrote that heat, light, noise, confusing directions and different testing time allowed to different students can affect students' scores. Mehnrens and Lehmann (1991) and Scattler (2001) said that errors in testing situation which include students misunderstanding or misreading test directors, noise level sickness can cause test score to vary. Griswold,(1990) noted if the testing environment is distracting or noisy this would affect the test reliability. so he advises that actins ought to be taken to ensure that the testing environment is comfortable.

2.4.3 Interval between the test

According to (Scattler, 2001) the length of time between test and retest interval can affect reliability due to the effect of memory where the test taker is able to remember the questions easily if the interval is too short hence the shorter the time the higher the reliability, but if the interval between the two administration is wide then the test taker would not be able to remember hence low reliability. The situation in Isinya sub-county is that, the time when the first exams is given to candidates sometimes varies because they can sit for an exam immediately the term begins and after every two weeks they have another exam, but some other time these exams are in two weeks, three weeks or even after a month, this will not give the test administrator a good judgment of the reliability of these commercial exams.

2.4.4 Type of test

Group homogeneity and heterogeneity of the items according to (Mehnrens et. al 1991) and (Scattler, 2001) said that the more heterogeneous the group of students who take the test the more reliable the measure will be, while Lucy Jacobs (2001) pointed out that under scoring the objectivity or the extent to which equally competent scores

obtain the same score is a factor affecting reliability. She noted that, objective test is more reliable because the test scores reflect true differences in achievement among students and not the judgment and opinion of the scorer.

Again easy tests have lower reliability than multiple choice tests because subjectively in scoring lower reliability. Item construction by the test administrator is also a factor which affects reliability. Poorly worded or ambiguous questions are trick questions are threat to reliable measurement. According to Lucy Jacobs (2011) too easy or too difficult test will typically have low reliability. This is because scorers will be clustered together at either higher and or the low end of the scale with small differences among students. Reliability is high when the scores are spread out when the entire scale sharing real differences among students.

2.5 Measures of Reliability

These are concerned with determining the degree of consistency in scores due to random error. Spearman -Brown and Kulder- Richardson provides estimates of the extent to which students would receive similar scores if they were retested with an equivalent form of the test. Spearman- Brown according to Nunnally& Berstain ,(1994)reflects consistency due to item sampling only. Kulder Richardson (K-R 20) measures consistency of responses to all the items within the tests and reflects the error sources. Where there is reliability coefficient of.80 and above this shows a perfect reliability where the coefficient is .60 and less the reliability is suspect. (Nunnally,1994), hence it's important for teachers to strive for at least a reliability of at least.70.

2.5.1 Standard Error of Measurement (SEM)

Standard Error of Measurement (SEM) is a statistic that obtains the confidence internal for many obtained scores. It represents the hypothetical distribution we would have, if someone took a test on different times. It's based on the assumption that any test score contains an error component the SEM is used to estimate a band or interval within which a person's true score would fall incase no error of measurement.

2.5.2 How to Control Factors Affecting Reliability

2.5.2.1 Writing Longer Tests

According to Lucy Jacob (1991) there are various ways to improve reliability of classroom test, one of these ways is writing longer tests, therefore instructors should write as many questions as one thinks the learners can complete in testing time available. (Nunnally, (1978) also earlier had noted that for reliability other reasons in psychometrics the maxim holds that, other things being equal a longer test is a good test Ellen Drost, (2011) points out that the principal method to make test more reliable is to make them longer, thus adding more items.

2.5.2.2 Construction of Test Items

Another way of controlling the factors affecting reliability is to pay more attention to the careful construction to test questions. Lucy Jacobs (1991) points out that each question should be clearly phrased to enable learners know exactly what is required. In agreement to this is Nunnally, (1978) as cited by Ellen Drost (1991) that reliability can be improved by writing items clearly, making test instructions easily understood and training the raters effectively by making rules for scoring as explicit as possible. The test writer should start writing the items well ahead of the time the test is to be given.

A hurriedly written test is likely to be unreliable. Lucy Jacobs (1991) says one should write clear directions and use standard administrative procedures. Kinyua and Okunya,(2014) noted that use of Blooms taxonomy in test item construction and prior training of teachers on test construction to enable them design items that address various cognitive levels of thinking as per Blooms taxonomy affect the reliability of a test (.Justine & John, 1996.) says carefully written tests with an adequate number of items usually produce high reliability. This is because they provide a representative sample of behavior being measured and the scores are apt to be less distorted by chance factors.

2.6 Content Validity

Content validity is part of the three types of validity. Context validity refers to the extent to which an assessment represents all facets of tasks within the domain being assessed. (Melissa Hurt, PhD) context validity answers the question. Does the assessment covered representative sample of the context that should be assessed. Therefore, when examiners are setting an end of year cumulative exam but the test only covers material precompiled in the last three weeks of class, the exam would have low context validity.

According to Yaghmaie F, (2003) context validity can help to ensure construct validity and give confidence to the readers and researchers about instruments, it is can be used to measure the appropriate sampling of the context domain of items in questionnaire. Kerlinger, (1986) as cited by Yaghmaie, (2003) argues that context validity is representative of the context, thus context validity of an instrument depends on the adequacy of a specified domain of context that is sampled.

Bush, (1985) pointed out that context validity refers to the degree that the instrument covers the context that it is supposed to measure. This research is focused on the content validity of the commercial assessment, tests which are used by teachers in assessing the learners. Again these commercial exams are used by the researcher is to predict the scores of the summative exam (K.C.P.E) hence the researcher felt that the commercial tests have low context validity and therefore a need to find better instrument for predicting to outcome of the learners scores or improve on the commercial tests and make them more standardized assessment is now at the intersection of new trends that have been shaping educational policies in OECD countries since the 1980s.

For instance in France the "Mirror effect" theory demised by Thelot (2003, quoted in Pons, 2008) states that standardized assessment should have a "Mirror effect". This means that the assessments should confront players in the education system with the results of their actions but need not necessarily provide them with explanations.

2.6.1 Measurement of Content Validity.

According to lecture by Chris Clause, (Prof. West Virginia University) content validity is often measured by relying on the knowledge of people (experts) who are familiar with the construct being measured. These subject experts are provided with access to the measurement tool and are asked to provide feedback on how will each question measure the construct in question. Their feedback is the analyzed and informed decision can be made about the effectiveness of each question. A method which is widely used to measure content validity was developed by C.H Lawshe, (1975).

It is a method for gauging agreement among rates or judges regarding how essential a particular item is .Lawshe (1975) as cited by the Wikipedia proposed that each of the subject matter experts raters (SMEs) on the judging panel respond to the following question for each item. “Is the skill or knowledge measured by this item ‘essential’ useful, but not essential, ‘or’ not necessary to the performance of the construct?” According to Lawshe, if more than half the panelists indicate that an item is essential, that item has at least some content validity. Lawshe developed a formula termed the context validity ratio CVR=ne-N/2, where CV12=context validity ratio, Ne= number of SME panelists indicating “essential”, N total number of SME panelists. This formula fields values ranging from +1 to -1, positive values indicate that vat least half the SMEs rated the item as essential.

According to Yaghmaie F, (2003) there is no complete objective method for determining the content validity of an instrument nor is any statistical approach, he however says that, content validity in the judgment stage is based on quantitative evidence. So in this stage professional subjective is required to determine the extent to which the scale was designed to measure a trait of interest. According to Oluwatayo James (2012) cited Sireci (1998) that in the statistical method, the most frequently used method, is factor analysis. Factor analysis is used to determine whether items are the instrument fit into conceptual domain.

Another method is the test specification method which provides the organizational framework for the development of the instrument Whiston,(2005) cited by Oluwalayo James,(2012).This is followed by defining the behavioral change, affective or cognitive changes that the research intends to measure. From the above discussion on the measures of content validity, there remain a question to be addressed if at all the

commercial mathematics test developers ever take this into consideration in the setting process of these exams, and if at all this is done, is it done by qualified personnel (experts) to conquer with Sireci,(1998).

2.6.2 Factors Affecting Content Validity

2.6.2.1 Test-Related Factors

According to Deale, (1975) as cited in the chapter 7 (Reliability and validity of assessment methods p.143), he says that, a long test is more reliable and valid because of short test would not adequately cover a years' work. Therefore it is important to have the syllabus sampled. The content tested in any given time should be relevant to what the learner has covered, this may not be the case with most of the mathematics commercial assessment tests used by teachers in Isinya district. Some topics assessed in term one are found in term three work, hence making the tool of low content validity.

According to Maizam Alias, (2005) determining the test objective is the first step in test construction, this is the criterion that will be used in order to judge if the test is valid or not. Therefore test should reflect the skills to be tested which involves looking into consistency between the syllabus content, the test objective and test contents. A study which was conducted by National Assessment of Education Progress (NAEP) Validity Studies (NVS) panel in 2006, one of the questions which it was supposed to answer was, if the NAEP framework was offering reasonable content and skill- based coverage compared to the assessment of states and other nations.

The panel undertook a number of experts' reviews where this question was addressed by a committee or mathematics and mathematic educators to compare the NAEP framework to the standards and test blue prints of selected six states. The findings of

this study showed that although the NAEP mathematics assessment is sufficiently robust to support the main conclusions that had been drawn about United States and state progress in mathematics since 1990, there are gaps among subgroups which existed validity issues uncovered by the study tended to be local in nature –affecting a particular set of items on particulars subscale.

Therefore the conclusion was made that there was need to sharpen the framework by the national assessment governing Board. The advice given was, focus should be on not worrying about leaving thing out; but about targeting the most important things, by reducing number of objectives and also sharpen the language of the objectives to give tests developers a better target rather than using language that tries to include all possibilities. However, this study did not give any recommendations on the importance of having experts; (judges) who can give their suggestions after the scores have been awarded to learners.

2.6.2.2 The criterion to which one compares his/her instrument may not be well enough established.

A research done by George J. SolterJr, (2010) where the focus was on criterion-related validity of the 8th grade assessment in New Jersey which he focused on trying to solve the problem of having the administrators and teachers only using the summative assessments to determine achievement levels but must go beyond meeting the mandates of state regulators to meet academic proficiency. George Jr. (2010) used the grade 8 assessment in either adequate year progress (AYP)/ Grade Eight Proficiency Assessment (GEPA) 8 scale score as depended variable, and high school proficiency assessment scale score of each student in mathematics in a New Jersey B-district factor group school, (DFG).

An approximate of 200 students in the district was chosen where the mathematics scores by grade 8 class were used to predict indicators of grade II assessment scores. The outcome of the study was that the grade 8 assessment scores were strongest predictor for grade II assessment scores. However from the finding of this study, the researcher did not address the use of standardized assessment tests by the grade II class as a predictive measure to their end of year scores. Therefore this gap to be addressed and focus on the importance of using the rights tools or instrument to predict the outcome of grades and scores of a candidate class.

2.6.2.3 Intervening Events

Having seen that content validity is highly based on the agreement of several judges, the content validity can be affected. For example, in an event where 3 to 5 raters are expected to come up with results which conclusions are drawn from, then one of the 3 or 5 raters declines. This automatically affects the content validity Michael Miller,(2011). In simple terms an assessment can be valid yet not reliable or it can be reliable and not valid (chapter 7 reliability and validity of assessment methods p. 144).

2.7 Interaction between Reliability and Validity of a Test.

Some people may think of reliability and validity as two separate concepts in reality reliability and validity are related. According to journal by Scotland Qualification Authority (SQA, 2004), validity and reliability are interdependent. An assessment that produces inconsistent results cannot provide valid information about a candidate's achievement; on the other hand highly consistent results do not necessarily indicate high validity, since the assessment could be inappropriate for the skill being assessed.

2.7.1 Performance

From a research done by George J.Jr (2010) on a criterion related validity study on the grade 8 assessment and the High School Proficiency Assessment (HSPA) mathematical for AB district factor group school in New Jersey, the population involved was 200 students, where the grade 8 student results were used to determine the entry to grade II, and again the administrators use the grade II a results to plan performance objectives as outlines in the New Jersey Quality Single Accountability Continuum (NJQSAC).The researcher found out that the students proceed to grade 12 with the HSPA grades hence having the results meaningless and this automatically affects reliability of the assessment.

From, the finding of this study the researcher also followed two studies for predictive variable of grade and assessment namely Alindish, (2003) who evaluated the predictive variable of grade 8 assessment scores on the Pennsylvania System of School Assessment (PSSA) along with grade-point average (GPA), and course work as predictors of Grade II assessment scores. The outcome was grade 8 assessment scores were the strongest predictive indicator for grade II assessment scores. Mindish, (2003) cited the results from the study of Potaski, (1996) who found the best prediction of student scores on assessment were scores on another assessment.

However the above study did not clearly do a researcher on having concurrent exams of the same grade as a predictor of joining the immediate grade as it is the case in this research. Where the third term scores are used by administrators to predict the scores are used for the summative exam in this case, the K.C.P.E.

2.7.2 Repetition

Due to use of poor assessment instruments which give wrong implications it leads to adverse negative effects to learners, and effects which sometimes lead to demoralizing the individual for life. For example repetition, study by Debard and Kubow, (2002) in a school is Ohio, 83% of primary school pupils and 45% of secondary pupils maintained that they worked hard because of the tests. Also school dropout due to repetition as seen in research by Haney, (2000) testing added stress to learners with some pupils forced to repeat a year and others stigmatized because of having learning difficulties and are unable to pass the tests. Due to high accountability required by administrators therefore, teachers do not give these learners attention because focus is on scores and not on the quality of the assessment tool.

2.7.3 Socioeconomic Factors (SES)

The landmark study by (Colemans, 1966) of equality of educational opportunity as quoted by (June & Stock, 2003) in his research, socioeconomic status has been a strong predictor of learner achievement. Coleman asserted that the influence of student background was greater than anything that goes on within schools. This being an important variable as noted by Coleman, the issue of Socioeconomic Status (SES) and its relationship to student achievement is more complex. First the relationship can be explored on various unit levels, from that of nations and states, districts and schools, and to classes and individual students.(June &Cathy, 2003).

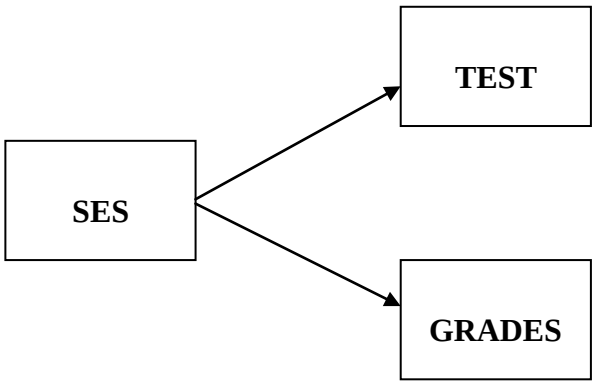
Learners from disadvantaged background with limited resources to prepare the test, branded as failures in schools with mediocre performance results. Jones (2007) these learners also experience more intensive coaching for tests and a stronger narrowing of the curriculum than in schools with privileged pupils. Focus or scores detracts from

environmental studies and general culture which is demanding for pupils whose families are less able to provide support. In support of this a study carried out by Maylone, (2002) as cited by George Jr., (2010) studied the connection between school districts social economic status (SES) and aggregate student scores at the Michigan Educational Assessment program (MEAP).

The SES included free and reduced lunch, amount of state aid, percent poor children, percent of one parent, households mean. Income median income and households with income less than $ 30,000.The results of Maylone (2002) produced results in agreement with Cooley, (1993) that a student living in poverty, with a single parent who is not a high school graduate accounted for 60% of the variance in the average district test scores for Pennsylvania school districts.

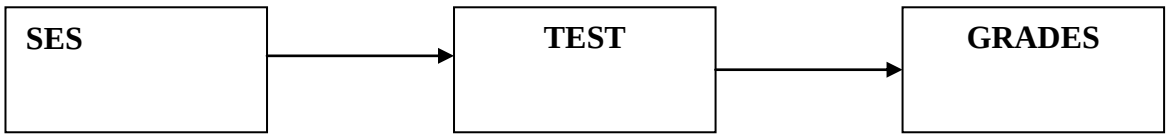
The implicit criticism is that socioeconomic status (SES) has an artificial and irrelevant effect on test scores: High SES leads to high test scores (for example through knowledge of test taking techniques) but not to higher true standing on characteristic the test is intended to measure.(Paul R. Sackett,et al; 2009).There are two conceptual models of relationship between test score and grades which were contrasted by (Paul Sackett et al; 2009) Model I, implicit that SES influence test scores, and SES influences grades, but there is no direct relationship between the characteristics measured by the tests and grades. Any correlation between test scores and grades is an artifact of the common influences of SES on both test scores and grades. If the model is correct. Then the correlation between test scores and grades will drop to zero when statistically controlling SES.

Figure 2. 1. Model 1 SES = Socioeconomic Status



In model II, the SES affects the characteristics measured by test which subsequently affect the grade, but here SES is not posited to have a direct relationship with grades. Its link to grades is a result of the mediating role of test scores.

Figure 2.2 Model II SES =Socioeconomic Status



Both of these models articulated above, indicate SES test relationships where Model I, views this relationship as art factual: controlling SES, the test-grade performance drops to zero or near zero. Again Model I true, continued test would be inappropriate which is opposite of Model II true that test scores contain meaningful information predictive of academic performance and the focus shift to the question of societal consequences of the fact that being higher in SES confers a meaningful advantage.

Having this argument then, there is a call for interventions and address the use of tests. It's also important to differentiate between criticizing tests on the grounds that they are not valid measures of academically relevant skills and not criticizing tests on

the grounds that one is not comfortable with social consequences of using a test, despite its being a valid predictor of academic performance.

2.7.4 Gender and Age

According to a study carried out by (Fennema Carpenter & Levi 1998) Mathematical ways of thinking may differ by gender. They studied 82 children as they progressed from 1st, 2nd to 3rd grades. They identified gender differences in strategy use that was evident from the beginning of the study and persisted through the end. Girls tended to use more modeling or counting strategies. The boys used more abstract strategies such as derived facts or invented algorithms, by the 3rd grade girls used significantly more standard algorithms than did the boys.

A research by (Donahue et al; 1999) as cited by (June & Cathy2003) indicated that some correlation appears to exist between gender and reading achievement. The results of the National Assessment of Educational Progress (NAEP, 1998) reading results by gender rather than race reveled that females out performed males in the 4th,8th&12th grades and they did also the same in 1992 and 1994.At 4th grade levels 1994, the males however made a significant gain while the females remained the same.

On the NAEP (2000) Mathematics assessment (U.S Department of Education 2001b), however a higher percentage for boys performed at or above proficient than girls at 4ht, 8th and 12th grades, with the older two grades being significantly higher. However there was no significant difference by gender at the 4th grade level. In an international comparison of Third International and science (TIMS), study in English-speaking countries,(Webster & Fisher 1999) indicated that in Australia and United States, very little of the student level variance was explained by gender and SES ,although most of the variance was at the student level and not at the class level.

The United States (U.S) Department of Education (2000) analysis of the same data revealed that males out performed females in 3 of the 25 countries at the 4th grade level, in 8 of the 39 countries at the 8th grade level and in 18 out of the 21 countries participating in their final year of secondary school. However in the U.S, males an read questions d females scored similarly at all three levels. A study by (D. H Carol 2009) on socioeconomic data and the achievement data derived from the Mathematical tests applied to children from grade 2 onwards (Statistics Canada 2001) A sample of 6,290 students which involved children and adolescents aged 7-15 attending school, took math's test and had mathematical scores in two sets.

A test with 15 questions was administered in school. For grade 2 students an interviewer read questions and recorded the answer on answer sheet. Test difficulty varied with grade of the students. Different forms depending on the grade level in which a student enrolled were therefore given. The result of this research was that the response rate of the Mathematics test was rather low: 485, 74%, 49% and 81% in grade 1-4 respectively. It was evident that there was interaction of SES with age, and it is not related to response rate and can be less biased if other demographic factors are controlled like family income, parental education and parental occupation.

2.8 Objective 2: Teachers rational for choosing commercial exams

2.8.1 Teachers test construction competency

For teachers to be able to construct their own tests or have the knowledge in choosing the right test instrument they must have the competency in test construction. In Ireland it is clearly stated in the N.C.C.E that it is important to provide support to teachers and schools to enable them use assessment in the most effective way to enhance teaching and learning, to construct and communicate useful and helpful

summarized records to children's progress and attained across a range of curriculum areas.

According to the results a study by Pascal M. Kagete (2013) one of the objectives was the frequency of testing and purpose of testing where the results indicated that, 64% of the teachers were from schools which had a defined number of formal assessments to be given in a given school term, only 6% admitted that they have the freedom to choose the number of assessment tests in a given school term.

The National Council for Curriculum & Assessment (N.C.C.E, 2004) report on the recent developments in assessment, it was noted that in Ireland many teachers construct and administer their own tests, administer standardized tests, and report the result to the parents and others. These teachers also engage in their own informal assessments of pupils and use their findings to inform ongoing teaching and learning activities. This is because in the primary school curriculum in Ireland it is noted that,

> *Assessment assists communication about children's progress and development between teacher and child, between teacher and parent and between teacher and teacher. (Primary school curriculum, 1999, pg. 17).*

Generally the overall idea is to reduce the importance of test scores and academic burden, schools are therefore, to be evaluated based on how much academic burden they put on student. To ensure that teachers use the right assessment tools, the government can develop a school policy on assessment and be well defined in the education act of Kenya.

Assessment and evaluation in Kenya primary schools is more based on measuring the examination outcome rather than measuring the learners' abilities. School based

assessment need to be strengthened so that regular and cumulative assessment to the forum of competence assessment tests (C.A.Ts) is put in place. The current education system examination based and that the assessment has little regard to molding good citizens and for self-reliance (Education reforms-recommendations by Task Force (TF) appointed by education minister Professor Sam Ongeri, January 2011) according to the TF it was recommended that there is need therefore to introduce competency based curriculum. The TF noted that assessment is not seen as part of the teaching and learning process but as a sieve to determine those who can move to higher education where the limited available spaces dictates the teaching /learning process towards examination as opposed to competences applicable to life.

2.8.2 Time factor

Testing apart from exerting a positive influence on student learning, it may slow the learning and instructional process, distort curricula and interfere with valuable instructional time (Bracey, 1989, Williams 1989; stake 1988) this criticizes the commercial examinations which are all printed in series- series 1-10 or 1-9 all of which many primary schools sit for, this interferes with time for instructions because teachers concentrate on prior preparations before the exams and revision of the same exams. According to the results of Pascal M. Kagete (2013) one of the objectives was the frequency of testing and purpose of testing. Where the results indicated 54% of the teachers were from schools which had a defined number of formal assessments to be given in a given school term and only 6% of the teachers admitted that they had the freedom to choose the number of assessment tests in a given school term.

Black and William (1998) synthesized over 250 studies linking assessment and learning, and found that the intentional use of assessment in the classroom to promote learning improved student achievement. Increasing the amount of time on assessment,

however, does not necessarily enhance learning. Rather, when teachers use classroom assessment to become aware of the knowledge, skills, and beliefs that their students bring to a learning task, use this knowledge as a starting point for new instruction, and monitor students' changing perceptions as instruction proceeds, classroom assessment promotes learning.

To the contrary of what the United States and some western countries are doing, China seems to be taking a new turn on testing. In a recent document sent to all provincial education authorities (July 19th 2013), the ministry of education unveiled guidelines and new framework for evaluating schools. These were known as the new ten commandments of education reform- no standardized tests no written homework, no tracking.. Some arguments for this change was that current evaluation style hamper children development as a whole person, stunt their health growth and limit opportunities to cultivate social responsibilities, creative spirit and practical abilities in students.

2.9 Theoretical Framework

This research will adopt the classical test, theory (CTT) which is a body of related psychometric theory that predicts outcome of psychological testing such as the difficulty of item or the ability of test takers classical test theory was codified by Norick (1966).CTT has been the most widely used theory in area of educational testing (Greg Pope, 2009 p.2).In classical test theory the observed scores obtained by test takers in assessment is composed of a theoretical measurable "the score" and error. This means that,

Observed assessment score= True score (Exam score) + measurement error, (Edward A Carmines, 1979).

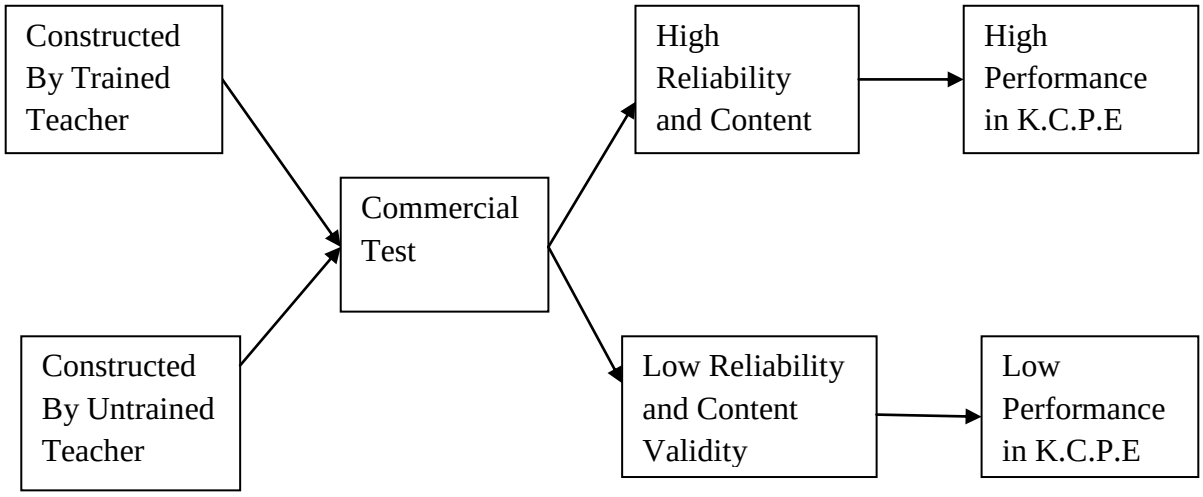
Observed assessment score= Trust exam score (can't be measured) + measurement error (can't be estimating).

Measurement error can be estimated and relates to reliability. High assessment score reliability means less error measurement. Classical Test Theory (CTT) provides formative for item analysis analytics. This is done for the purposes of finding out whether the questions in a given assessment test are performing a manner that is psychometrically appropriate and defensible. This helps in assessing the content validity as for as a given test is concerned. It is also helpful in evaluating the psychometric performance of questions and incase the items need to be improved, sent to the scrap heap, or left as they are because they meet all the criteria for bring in an assessment.(Nunnally, 1978) cited by Edward Carmines, (1979).

2.10 Conceptual framework

Figure 2.3 Conceptual framework

Shows the variables to be studied.



The above conceptual framework shows the relationship between the independent and the dependent variables. The independent variables can be used to improve the pupils scores that is by having qualified teachers to assess the credibility of the mathematics

assessment instruments. The use of poor quality test instrument may negatively affect the pupil score. This will therefore, provide wrong feedback to both the pupil and the teacher leading to wrong judgments. Good assessment tools leads to clear judgments on how to better learners results and give a good 'mirror' to focus on the summative assessment tests.

The dependent variables are those that the research measures in order to come up article change that can be done. Use of good assessment test fools which are reliable and have the right content (content validity) in Isinya sub-county will be able to bring out a good reflection of the pupils performance in both internal and external examinations.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter discusses the research methodology that will be used in the study. This includes the research design, target population, sample size, sampling techniques, research instruments, validity and reliability of instrument, data analysis and data collection procedures.

3.2 Research Design

Descriptive survey was employed to investigate the quality of commercial mathematics tests in relation to pupil's performance. Secondary data will be also used. This is the use of existing sources of information (Cozby, 2004) which include the statistical records in the D.E.O Isinya, KNEC and various sampled school. The survey research design will involve selection of a sample of teachers in public primary schools of Isinya sub-county teaching mathematics to the questions in the questionnaire.

3.3 Target Population

Mcmillan and Schumacher, (2010) state that target population is a group of elements or cases, whether individuals, or objects, or events, that conform to specific criteria and to which the results of the research can be generalized. The target population in this research was the class eight candidates term 3 scores in the commercial exams in relation to the K.C.P.E scores in the years 2011, 2012 and 2013.

Respondents were teachers of mathematics or mathematics panel heads; this is because these teachers are involved in actual teaching and guiding the learning of

mathematics in their schools. They are responsible for planning and implementing the process of testing mathematics in schools.

3.4 Sampling and Sampling Techniques

A sample is a smaller group obtained from accessible population. In this study a sample was selected to be a representative of whole population with salient characteristics. Also sampling refers to taking a portion of a population or universe as representative of that population or universe (Kerlinger, 2006). The sampling techniques are the methods employed in selecting a representation portion from each of the population relevant to the study.

According to (Kothari, 2004) stratified random sampling is accurate, easily accessible and divisible into relevant strata also enhances better comparison. Therefore, this type of sampling will be highly dependent upon. The strata was 2 divisions in Isinya district, (Kitengela and Isinya divisions). Purposive sampling technique was also used, this was to collect various commercial assessment tests ensuring that various kinds or tests were captured using likeart scale each item will be analyzed on its validity each with content validity.

Table 3.1: Sample per Division in Isinya District

Name of division	**Total No. public primary schools**	**No. of public primary schools for sampling**
Isinya	16	8
Kitengela	13	7
Total number of schools	**29**	**15**

Source: D.E.O Isinya , 2014

3.5 Research Instruments

One type of the instrument was used to collect data. This was self-administered questionnaires. The self-administered questionnaire as defined by Bernard, (2006) is a questionnaire that a respondent completes his/her own either on paper or any other writing material provided, by answering questions designed to obtain answers pertinent to research hypothesis. This instrument was considered appropriate since all respondents were expected to have sufficient literacy level to enable them read, understand and answer the given question required.

This instrument was applicable because it was easy to disperse the questionnaire to the respondents. A further justification for choice of this instrument is that it is less expensive as compared to personal interview or telephone interview. Questionnaire instrument as shown in Appendix 1 were administered to the class 8 mathematics teachers currently teaching the class or have previously handled the a candidate class, the mathematics panel heads will also respond to each item in the questionnaire. The questioner contained 20 items related to the objectives of the study; this helped in comparing the reliability and content validity of the formative and summative exams. Also several formative and summative items were analyzed to investigate their reliability and content validity.

3.5.1 Commercial Assessment Past Papers (Mathematics test for standard 8)

This was used to compare the scores in school based FA test and the SA tests. The reliability and content validity were evaluated using a set of sampled commercial exams for each item.

3.5.2 K.C.P.E Mathematics Past Papers

This was the mathematics papers done in the years 2012 and 2013.

3.6. Pilot Testing

Pilot testing is done in any study in research as it helps in making the research instrument effective in capturing the information required(.Mugenda and Mugenda, 1999).The researcher will undertake this in 4 public primary schools in Isinya district and the sample of 2 per division. The respondents, who were mathematics teachers, were given at least three days to fill in the questionnaires. After this the data was collected and analyzed. In this case, respondents did not raise an issue of concern as relating to the language used or any other issue, so the researcher did not modify the language or rephrase the questions. The instrument was ready for use to meet the objectives of the study.

3.7 Instrument Reliability

Instrument reliability refers to the degree for which a test consistently measures whatever it is intended to measure. The more reliable an instrument is, the more confidence we can have meaning that the same results will be obtained in case the research was to be re-administered to the same respondents (Gay et al 2006)To ensure that the instrument of the data collection was reliable, equivalent forms were used during the piloting stage where two different but alternative forms of questionnaires were administered to the respondents in the pilot schools at the same time.

The question items were different on the forms but constructed to sample the same content. The scores for the two groups of responses were collected to determine the agreement among the responses: then the Person product moment correlation coefficient (r) was used in order to obtain correlation coefficient of the two scores.

3.7.1 Instrument Validity

Validity refers to the accuracy of an assessment or study. If the results or the study can be interpreted and generalized to other population (Conen, 1988), the instrument was tested so as to check its content validity and face validity. To ensure validity of questionnaire some question were based on respondent's attitude and opinions and multiple choice questions with adequate opinions were used. These questionnaires were piloted and answers provided were analyzed to determine their validity.

3.8 Data Analysis

The data analysis included sorting, editing, coding and processing the data (Borg &Gall ,1996) this was done by Ms Excel and the Statistical Package for Social Sciences (SPSS) version 12 renamed predictive analysis software (PASW) 2009. Both qualitative and quantitative approaches were used. Quantitative approach was done using descriptive survey and was be analyzed through content analysis basing on the respondents information on the commercial tests and Fleiss Kappa method was used to make the conclusion. This was to avoid the field data errors, Qualitative data was analyzed and presented in frequency tables. Nine different raters were used to judge the content validity of three different questionnaires .Their responses were in the scale of 1-5 with 1 being lowest, 2-low, 3-moderate, 4- high and 5- highest.

Comparison between formative assessments (FA) with standardized national examination (K.C.P.E) was done. This means that statistical analysis based on the frequency at which the Mathematics Formative Assessment (FA) tests resembled the K.C.P.E Mathematics papers done in 2011 – 2013.

Piloting stage where two different but alternative forms of questionnaires was administered to the respondents in the pilot schools at the same time. The question items were different on the two forms but constructed to sample the same content.

The three sets of data from the two questionnaires were used to calculate reliability coefficient using the Fleiss kappa method

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Where: *k*= Fleiss kappa Coefficient

$1 - \bar{P}_e$= Degree of agreement that is attainable above chance, and,

$\bar{P} - \bar{P}_e$ = Degree of agreement actually achieved above chance.

If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

3.9 Ethical Issues

The research was purely conducted on voluntary participation. This means that the respondents were not be forced or enticed to participants in the research. Permission was sought from individual institutions participating in the study. Again the respondents were informed the purpose of the research and likelihood of the study becoming a reference document in the institutions. The researcher assured the respondents of their confidentiality and for this purpose the study did not reveal the identity of the participating colleges.

CHAPTER FOUR

DATA ANALYSIS, RESULTS AND DISCUSSION

4.1 Introduction

This chapter deals with the data analysis and interpretation of the findings. The results have been grouped under major headings. The results have been presented in both tables and figures showing mean, standard deviation of the scores and the correlation between the years.

4.2 Demographic Information of Teachers

4.2.1: Teachers' Demographic Information: School A

This entire section involves the biographical information of the respondent and is divided into two sections: Section I and Section II which explains biographical information and content validity of question paper on the scale of 1-5 which is: 1 lowest, 2- low, 3-moderate, 4-high, 5- highest respectively.

Gender

The table 4.2.1 shows that the respondent for Paper 006 where male and female representing 66.7 % and 33.3% respectively. This shows that the male teachers are actively involved in teaching and evaluating mathematics more than the female.

Table 4.2.1: Gender

GENDER					
SEX		Frequency	Percent	Valid Percent	Cumulative Percent
	Male	2	66.7	66.7	66.7
	Female	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.2.2 shows the age in years of the mathematics teachers which shows majority are between 31-40 years representing 66.7%, 41-50 Years representing 33.3%. This

Proofs that mathematics is usually taught by young professionals whose ages are 31-40 years.

Table 4.2.2: Age

AGE					
YEARS		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 30 Yrs	0	0	0	0
	31-40 Yrs	2	66.7	66.7	66.7
	41-50 Yrs	1	33.3	33.3	100.0
	Above 50	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.2.3 shows that 33.3 % and 66.7 % of the respondents are degree and Certificate holders are respectively. None of them had master degrees or any other qualification. It shows that certificate teachers take part in teaching mathematics.

Table 4.2.3: Education Level

EDUCATION LEVEL					
LEVEL		Frequency	Percent	Valid Percent	Cumulative Percent
	Master	0	0	0	0
	B.ED	1	33.3	33.3	33.3
	P1	2	66.7	66.7	100.0
	Other	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.2.4 shows that mathematics teachers are equally distributed in Kajiado County with working experience of 6-10 years, 11-15 years and 16-20 years representing 33.3% each. Less than 5 years and above 21 years was represented by 0% which means mathematics is taught by energetic young and expertise teachers.

Table 4.2.4: Years in Teaching profession

YEARS IN TEACHING					
YEARS	Years	Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
	6-10 Yrs	1	33.3	33.3	33.3
	11-15 Yrs	1	33.3	33.3	66.7
	16-20 Yrs	1	33.3	33.3	100.0
	Above 20	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.2.5 proves that mathematics is taught by expertise teachers of 11-15 years representing 66.6 % and 6-10 years representing 33.3%. None of the teachers taught mathematics with less than 5 years expertise.

Table 4.2.5: Years in Teaching Mathematics

YEARS IN TEACHING MATHS					
YEARS	Years	Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
	6-10 Yrs	1	33.3	33.3	33.3
	11-15 Yrs	2	66.7	66.7	100.0
	16-20 Yrs	0	0	0	100.0
	Over 20 Yrs	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.2.6 Shows that 66.7 % of the teachers have not been trained in test development hence lack knowledge of evaluating the commercial papers. 33.3 % of the teachers have attendant test development test.

Table 4.2.6: Test Development Training

TEST DEVELOPMENT TRAINING					
		Frequency	Percent	Valid Percent	Cumulative Percent
YEARS	YES	1	33.3	33.3	33.3
	NO	2	66.7	66.7	100.0
	Total	3	100.0	100.0	

Source: Research Findings

4.3 School B

4.3.1: Teachers' Demographic Information (paper 008)

The table 4.3.1 shows that the respondent for Paper 008 where male and female representing 66.7 % and 33.3% respectively. This shows that the male teachers are actively involved in teaching and evaluating mathematics more than the female.

Table 4.3.1: Gender

GENDER					
GENDER		Frequency	Percent	Valid Percent	Cumulative Percent
	Male	2	66.7	66.7	66.7
	Female	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.3.2 shows the age in years of the mathematics teachers which shows majority are between 31-40 years representing 66.7%,less than 30 years representing 33.3 %, 41-50 Years representing 0 %. This Proofs that mathematics is usually taught by young professionals whose ages are less than 30 years and 31-40 years

Table 4.3.2: Age

AGE					
		Frequency	Percent	Valid Percent	Cumulative Percent
YEARS	Less 30Yrs	1	33.3	33.3	33.3
	31-40 Yrs	2	66.6	66.6	100.0
	41-50 Yrs	0	0	0	100.0
	Above 50 Yrs	0	0	0	100.0

Source: Research Findings

Table 4.3.3 shows that 33.3 % and 66.7 % of the respondents are degree and Certificate holders are respectively. None of them had master degrees and any other qualification. It shows that certificate teachers take part in teaching mathematics.

Table 4.3.3: Education Level

EDUCATION LEVEL					
LEVEL		Frequency	Percent	Valid Percent	Cumulative Percent
	Masters	0	0	0	0
	B.ED	1	33.3	33.3	33.3
	P1	2	66.7	66.7	100.0
	Other	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.3.4 shows that mathematics teachers are equally distributed in Kajiado County with working experience of 6-10 years, 11-15 years and 16-20 years representing 33.3% each. Less than 5 years and above 21 years was represented by 0% which means mathematics is taught by energetic young and expertise teachers.

Table 4.3.4: Years in teaching profession

YEARS IN TEACHING					
		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
YEARS	6-10 Yrs	1	33.3	33.3	33.3
	11-15 YRS	1	33.3	33.3	66.6
	16-20 Yrs	1	33.3	33.3	100.0
	Over 20 Yrs	0	0	0	100.0

Source: Research Findings

Table 4.3.5 proves that mathematics is taught by expertise teachers of 6-10 years, 11-15 years and 16-20 years representing 33.3% each. None of the teachers taught mathematics with less than 5 years expertise.

Table 4.3.5: Years in Teaching Mathematics

YEARS TEACHING MATHS					
YEARS		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
	6-10 Yrs	1	33.3	33.3	0
	11-15 Yrs	1	33.3	66.7	100.0
	16-20 Yrs	1	33.3	100.0	100.0
	Over 20 Yrs	0	0	0	100.0

Source: Research Findings

Table 4.4.6 Shows that 66.7 % of the teachers have being trained in test development hence have knowledge of evaluating the commercial papers. 33.3 % of the teachers have never attendant test development training hence may lack knowledge of evaluating commercial papers.

Table 4.3.6: Test Development Training

TEST DEVELOPMENT TRAINING					
TEST		Frequency	Percent	Valid Percent	Cumulative Percent
	YES	2	66.7	66.7	66.7
	NO	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

4.4 School C

4.4.1 Demographical Information for teachers Analyzing (Paper 009)

The table 4.4.1 shows that the respondent for Paper 009 where male and female representing 66.7 % and 33.3% respectively. This shows that the male teachers are actively involved in teaching and evaluating mathematics more than the female.

Table 4.4.1: Gender

GENDER					
		Frequency	Percent	Valid Percent	Cumulative Percent
Gender	Male	2	66.7	66.7	66.7
	Female	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.4.2 shows the age in years of the mathematics teachers, which shows majority are between the ages of 31-40 years representing 66.7%, 41-50 Years representing 33.3%. This Proofs that mathematics is usually taught by young professionals whose ages are between 31-40 years.

Table 4.4.2: Age

AGE					
Age		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 30 Yrs	0	0	0	0
	31-40 Yrs	2	66.7	66.7	66.7
	41-50 Yrs	1	33.3	33.3	100.0
	Above 50	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.4.3 shows that 66.7 % and 33.3% of the respondents are degree holders and Certificate holders are respectively. None of them had master degrees and any other qualification. It shows that certificate teachers take part in teaching mathematics.

Table 4.4.3: Education Level

EDUCATION LEVEL					
Level		Frequency	Percent	Valid Percent	Cumulative Percent
	Master	0	0	0	0
	B.ED	2	66.7	66.7	66.7
	P1	1	33.3	33.3	100.0
	Other	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.4.4 shows that mathematics teachers are not equally distributed in School C in Kajiado County with working experience of 6-10 years, 11-15 years,16-20 years and 21 years above representing 0%, 66.7%, 0% 33.3% each. Less than 5 years and above 21 years was represented by 0% which means mathematics is taught by energetic young and expertise teachers.

Table 4.4.4: Years in Teaching Profession

YEARS IN TEACHING.					
Years		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
	6-10 Yrs	0	0	0	0
	11-15 Yrs	2	66.7	66.7	66.7
	16-20 Yrs	0	0	0	66.7
	Above 21Yrs	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Table 4.4.5 below proves that mathematics is taught by expertise teachers of 6-10 years, 11-15 years and 16-20 years representing33.3 %. None of the teachers taught mathematics with less than 5 years expertise and above 20 years.

Table 4.4.5: Years in Teaching Mathematics

YEARS IN TEACHING MATHS					
Years		Frequency	Percent	Valid Percent	Cumulative Percent
	Less 5 Yrs	0	0	0	0
	6-10 Yrs	1	33.3	33.3	33.3
	11-15 Yrs	1	33.3	33.3	66.7
	16-20 Yrs	1	33.3	33.3	100.0
	Above 20 Yrs	0	0	0	100.0
	Total	3	100.0	100.0	

Source: Research Findings

Tables 4.4.6 below, shows that 66.7 % of the teachers have been trained in test development hence have knowledge of evaluating the commercial papers. 33.3 % of the teachers have never attendant test development test.

Table 4.4.6: Test in Development Training

TEST DEVELOPMENT TRAINING					
		Frequency	Percent	Valid Percent	Cumulative Percent
	YES	2	66.7	66.7	66.7
	NO	1	33.3	33.3	100.0
	Total	3	100.0	100.0	

Source: Research Findings

4.5 OBJECTIVE I. Research Findings –Reliability and Content Validity of Mathematics Commercial Test used in Public Primary Schools.

4.5.1 Reliability of commercial exams:

This section shows the research findings of class 8 Mathematics mean scores performance in percentage of three schools A, B and C taking maximum of four exams and minimum of three exams per term which translates to nine or twelve exams per year.

4.5.1 Mathematics Mean Scores School A : Paper 009

The table 4.5 shows the mean scores of 2013 and 2012 of 41.25 and 46.21 of school A respectively with standard deviation of 3.16 and 5.01 respectively. In

Table 4.5.1 Mathematics mean scores School A. paper 009

Paper 009 2013/2012		
TEST	**MEAN SCORE 2013**	**MEAN SCORE 2012**
1st	34.53	44.29
2nd	39.60	40.25
3rd	42.00	38.61
4th	40.94	54.94
5th	42.78	43.16
6th	44.96	47.18
7th	41.29	40.96
8th	45.00	52.26
9th	40.16	50.09
10th	-	45.81
11th	-	47.34
12th	-	49.68
TOTAL	**371.26**	**554.57**
MEAN	**41.25**	**46.21**
STD	**3.16**	**5.01**
K.C.P.E	**48.56**	**44.67**

Source: Research Findings

Table 4.5.2 proves of correlation between performance 2013 and 2012. The correlation coefficient r=1.000, p =.000.The correlation is significant at the 0.01 level (2-tailed).There were no mean scores for the 10th, 11th and 12th the school did not have these exams.

Table 4.5.2: Correlations of 2013 and 2012 mean scores.

Correlations			
		2013	2012
2013	Pearson Correlation	1	1.000**
	Sig. (2-tailed)		.000
	N	13	10
2012	Pearson Correlation	1.000**	1
	Sig. (2-tailed)	.000	
	N	10	10
** correlation is significant at the 0.01 level (2-tailed)			

Source; Research Findings

Table 4.5.3 shows the mathematics mean scores of school B with mean of 54.90 and 49.03 in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 49.50 with a difference of 5.40% from the commercial papers and 2012 mean was 49.12% with difference of 0.08%,from data collection the mean score for the 11th exam was not given and therefore the total given was for 10 exams.

Table 4.5.3: Mathematics mean scores School B :

Paper 009 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	56.15	47.34
2nd	54.59	45.81
3rd	59.79	50.09
4th	40.94	52.26
5th	54.23	54.94
6th	51.70	49.96
7th	56.51	61.51
8th	61.45	44.29
9th	59.25	43.16
10th	56.15	40.96
11th	53.15	-
TOTAL	**603.91**	**490.30**
MEAN	**54.90**	**49.03**
STD	**5.48**	**6.12**
K.C.P.E	**49.50**	**49.12**

Source: Research Findings

Table 4.5.4 proves of no correlation between performance 2013 and 2012. The correlation coefficient r= -0.298, p=0.402 hence the correlation is not significant at the 0.01 level (2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.5.4: Correlation of 2013 and 2012 mean scores.

Correlations			
		2013	2012
2013	Pearson Correlation	1	-0.298
	Sig. (2-tailed)		0.402
	N	11	10
2012	Pearson Correlation	-0.298	1
	Sig. (2-tailed)	0.402	
	N	10	10
Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Findings

Table4.5.5 shows the mathematics mean scores of school C with mean of 52.47 and 50.99 in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 46.62 with a difference of 5.85% from the commercial papers. In 2012 K.C.P.E mean was 46.67 with a difference of 4.32 %. Standard deviation of 8.08 and 8.52 in 2013 and 2012 respectively. The school did not have the 12th exam hence there was no mean score indicated.

Table 4.5.5: Mathematics mean scores School C: Paper 009

Paper 009 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	38.96	46.23
2nd	46.74	42.85
3rd	45.68	48.10
4th	59.78	45.78
5th	59.92	47.70
6th	56.72	42.47
7th	60.72	50.55
8th	58.59	47.68
9th	45.61	47.45
10th	44.72	60.72
11th	59.72	63.57
12th	-	68.72
TOTAL	**577.16**	**611.82**
MEAN	**52.47**	**50.99**
STD	**8.08**	**8.52**
K.C.P.E	**46.62**	**46.67**

Source: Research Findings

Table 4.5..6 proves of no correlation between performance 2013 and 2012.The correlation coefficient is r= 0.088 and p= 0.79.The correlation is not significant at the 0.0 level (2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.5.6: Correlation of 2013 and 2012 mean scores.

Correlations			
		2013	2012
2013	Pearson Correlation	1	0.088
	Sig. (2-tailed)		0.797
	N	11	11
2012	Pearson Correlation	0.088	1
	Sig. (2-tailed)	0.797	
	N	11	12
Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Findings

4.6 Mathematics mean scores school A: Paper 008

Table 4.6.1 shows the mathematics mean scores of school A, with mean of 58.76 and 52.57 in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 51.45% with a difference of 7.21% from the commercial papers and 2012 mean of 46.45% with a difference of 6.12% Standard deviation of 3.74 and 4.77 on 2013 and 2012 respectively. In 2013 the school did not have the 11th exam hence no mean score was recorded.

Table 4.6.1 Mathematics mean scores school A

PAPER 008 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	57.45	46.71
2nd	60.12	52.51
3rd	55.12	54.51
4th	61.45	59.42
5th	65.42	49.15
6th	59.15	46.51
7th	61.42	51.65
8th	52.12	51.55
9th	59.12	54.47
10th	56.24	50.15
11th	-	61.65
TOTAL	**587.61**	**578.28**
MEAN	**58.76**	**52.57**
STD	**3.74**	**4.77**
K.C.P.E	**51.45**	**46.45**

Source Research Findings

Table 4.6.2 proves of no correlation between performance 2013 and 2012.The correlation coefficient r= 0.038 and p= 0.918. The correlation is not significant at the 0.01 level (2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.6.2: Correlations of 2013 and 2012 mean scores

Correlations			
2013		2013	2012
	Pearson Correlation	1	.038
	Sig. (2-tailed)		.918
2012	N	10	10
	Pearson Correlation	.038	1
	Sig. (2-tailed)	.918	
	N	10	11
Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Finding

Table 4.6.3 shows the mean scores of paper 008 in school B with mean of 53.20% and 52.34% in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 49.52% with a difference of 3.268% from the commercial papers and 2012 mean of 44.96% with a difference of 7.38%. Standard deviation of 6.39 and 6.88 on 2013 and 2012 respectively. In 2013 the school did not have the 11th exam hence no mean score was recorded.

Table 4.6.3: Mathematics mean scores school B Paper 008.

PAPER 008 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	42.89	51.01
2nd	49.24	46.34
3rd	47.62	45.25
4th	50.67	49.95
5th	56.50	50.49
6th	52.12	46.25
7th	65.71	69.42
8th	54.24	55.35
9th	59.35	57.50
10th	53.69	54.51
11th	-	49.68
TOTAL	**532.03**	**575.75**
MEAN	**53.20**	**52.34**
STD	**6.39**	**6.88**
K.C.P.E	**49.52**	**44.96**

Source: Research Findings

Table 4.6.4 proves of correlation between performance 2013 and 2012.The correlation coefficient r= 0.806, p=0.005.The correlation is significant at the 0.01 level (2-tailed). This shows that there is relationship between commercial exams of 2013 and 2012.

Table 4.6.4: Correlation of 2013 and 2012 mean scores

Correlations			
		2013	2013
	Pearson Correlation	1	.806**
	Sig. (2-tailed)		.005
2012	N	10	10
	Pearson Correlation	.806**	1
	Sig. (2-tailed)	.005	
	N	10	11
**. Correlation is significant at the 0.01 level (2-tailed).			

Source: Research Findings

Table 4.6.5 below shows the means of paper 008 in school C with mean of 53.06% and 58.95% in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 45.76% with a difference of 7.3% from the commercial papers and 2012 mean of 50.76% with a difference of 8.19%. Standard deviation of 8.51 and 8.44 on 2013 and 2012 respectively. School C did not have the 12th exam in the year 2013 hence there was no mean score.

Table 4.6.5 Mathematics mean scores school C: Paper 008.

PAPER 008 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	40.76	45.55
2nd	55.76	55.76
3rd	54.68	50.76
4th	52.34	45.76
5th	50.68	60.76
6th	44.78	67.76
7th	45.68	65.76
8th	48.66	60.76
9th	55.78	55.56
10th	68.76	60.78
11th	65.76	68.50
12th	-	69.70
TOTAL	**583.64**	**707.41**
MEAN	**53.06**	**58.95**
STD	**8.51**	**8.44**
K.C.P.E	**45.76**	**50.76**

Source: Research Findings

Table 4.6.6 proves of no correlation between performance 2013 and 2012. The correlation coefficient r= 0.229 and p= 0.499.The Correlation is not significant at the 0.01 level (2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.6.6: Correlation of 2013 and 2012 mean scores

Correlations			
2013		2013	2012
	Pearson Correlation	1	.229
	Sig. (2-tailed)		.499
2012	N	11	11
	Pearson Correlation	.229	1
	Sig. (2-tailed)	.499	
	N	11	12
** Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Findings

4.7 Mathematics mean scores school A: (paper 006)

Table 4.7.1 shows the mean scores of school of paper 006 in school A with mean of 53.28% and 54.66 % in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 47.24% with a difference of 6.04 % from the commercial papers and 2012 mean of 49.61% with a difference of 5.05%. Standard deviation of 3.69 and 2.93 on 2013 and 2012 respectively

Table 4.7.1 Mathematic mean scores school A: Paper 006

PAPER 006 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	49.68	51.10
2nd	50.16	54.15
3rd	49.12	55.19
4th	50.45	56.15
5th	55.46	51.20
6th	53.67	55.16
7th	54.10	51.69
8th	57.41	59.12
9th	59.46	58.15
TOTAL	**479.51**	**491.91**
MEAN	**53.28**	**54.66**
STD	**3.69**	**2.93**
K.C.P.E	**47.24**	**49.61**

Source: Research Findings

Table 4.7.2 proves of no correlation between performance 2013 and 2012.The correlation coefficient r= 0.442 and p= 0.233. The correlation is not significant at the level 0.01(2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.7.2: Correlation of 2013 and 2012 mean scores

Correlations			
2013	2013	2012	
	Pearson Correlation	1	.442
	Sig. (2-tailed)		.233
2012	N	9	9
	Pearson Correlation	.442	1
	Sig. (2-tailed)	.233	
	N	9	9
**Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Findings

Table 4.7.3 shows the means of paper 006 in school B with mean of 53.26% and 56.04 % in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 49.76% with a difference of 3.47 % from the commercial papers and 2012 mean of 6.28 % with a difference of 7.38%. Standard deviation of 11.03 and 8.49 on 2013 and 2012 respectively

Table 4.7.3: Mathematics mean scores school B: Paper 006.

PAPER 006 2013/2012		
TEST	MEAN SCORE 2013	MEAN SCORE 2012
1st	38.76	39.76
2nd	38.97	46.80
3rd	36.76	49.70
4th	54.68	52.60
5th	45.76	50.60
6th	57.80	58.70
7th	59.80	61.00
8th	59.76	59.75
9th	67.78	56.87
10th	60.87	60.20
11th	50.46	67.70
12th	67.76	68.78
TOTAL	**639.16**	**672.46**
MEAN	**53.26**	**56.04**
STD	**11.03**	**8.49**
K.C.P.E	**49.76**	**49.80**

Source: Research Findings

Table 4.7.4 proves of correlation between performance 2013 and 2012.The correlation coefficient r= 0.750, p=0.005.The correlation is at the 0.01 level (2-tailed). This shows that there is relationship between commercial exams of 2013 and 2012.

Table 4.7.4: Correlation of 2013 and 2012 mean scores.

Correlations			
2013	2013	2012	
	Pearson Correlation	1	0.750**
	Sig. (2-tailed)		0.005
2012	N	12	12
	Pearson Correlation	0.750**	1
	Sig. (2-tailed)	0.005	
	N	12	12
**. Correlation is significant at the 0.01 level (2-tailed).			

Source: Research Findings

Table 4.7.5 shows the means of paper 006 in school C with mean of 50.33% and 53.25 % in 2013 and 2012 respectively. In 2013 K.C.P.E mean was 47.38 % with a difference of 2.95 % from the commercial papers and 2012 mean of 48.89 % with a difference of 4.36 %. Standard deviation of 9.9 and 6.32 on 2013 and 2012 respectively. The 12th exam in 2013 was not done and therefore no mean score was indicated.

Table 4.7.5: mathematics mean scores school C: Paper 006.

PAPER 006 2013/2012		
TEST	**MEAN SCORE 2013**	**MEAN SCORE 2012**
1st	36.70	47.28
2nd	39.78	45.10
3rd	45.76	55.20
4th	42.78	56.70
5th	45.05	48.20
6th	50.00	52.05
7th	50.52	50.58
8th	50.68	54.68
9th	62.00	45.68
10th	64.78	56.89
11th	65.58	60.89
12th	-	65.76
TOTAL	**553.63**	**639.01**
MEAN	**50.33**	**53.25**
STD	**9.90**	**6.32**
K.C.P.E	**47.38**	**48.89**

Source: Research Findings

Table 4.7.6 proves of no correlation between performance 2013 and 2012.The correlation coefficient r= 0.468, p= 0.147. The correlation is not significant at the 0.01 level (2-tailed). This shows that there is no relationship between commercial exams of 2013 and 2012.

Table 4.7.6: Correlation of 2013 and 2012 mean scores.

Correlations			
2013		2013	2012
	Pearson Correlation	1	.468
	Sig. (2-tailed)		.147
2012	N	11	11
	Pearson Correlation	.468	1
	Sig. (2-tailed)	.147	
	N	11	12
**Correlation is not significant at the 0.01 level (2-tailed).			

Source: Research Findings

4.8.Analysis of the Content Validity of Commercial exams: Using Fleiss Kappa method:

4.8.1 Content validity of Paper 006.

Let N be the total number of subjects, let n be the number of ratings per subject, and let k be the number of categories into which assignments are made. The subjects are indexed by i = 1 ...N and the categories are indexed by j = 1, ... k. Let nij represent the number of raters who assigned the i-th subject to the j-th category.

N=50, n=3 and κ =5

= 1/50 (18.333) = 0.3667

= 0.0676^2 + 0.220^2 + 0.392^2 +0.273^2 +0.472^2

= 0.2844

= (0.3667 - 0.2844) / (1-0.2844)

κ = 0.1150

Interpretation and conclusion

< 0 Lowest

0.01-0.25 Low

0.26-0.50 Moderate

0.51-0.75 High

0.76-1.00 Highest

Paper 006 was rated 0.1150 hence the rating was rated low.

4.8.2 Content Validity of Paper 008

Let N be the total number of subjects, let n be the number of ratings per subject, and let κ be the number of categories into which assignments are made. The subjects are indexed by i = 1 ...N and the categories are indexed by j = 1 ... κ. Let nij represent the number of raters who assigned the i-th subject to the j-th category.

N=50, n=3 and κ =5

= 1/50 (21.333) = 0.4267

= 0.073^2+0.267^2 +0.453^2 +0.200^2 +0.007^2

= 0.32

= (0.4267 - 0.32) / (1-0.32)

κ = 0.157

Interpretation and conclusion

< 0 Lowest

0.01-0.25 Low

0.26-0.50 Moderate

0.51-0.75 High

0.76-1.00 Highest

Paper 008 was rated 0.157 which is between 0.01-0.25 hence the rating was rated low.

4.8.3 Content Validity of Paper 009.

Let N be the total number of subjects, let n be the number of ratings per subject, and let k be the number of categories into which assignments are made. The subjects are

indexed by i = 1, ...N and the categories are indexed by j = 1, ... κ. Let nij represt the number of raters who assigned the i-th subject to the j-th category.

N=50, n=3 and κ =5

= 1/50(19.000) = 0.38

= 0.060^2 +0.213^2 + 0.367^2+0.267^2 +0.093^2

= 0.26

= (0.38-0.26)/ (1-0.26)

κ = 0.162

Interpretation and conclusion

< 0	Lowest
0.01-0.25	Low
0.26-0.50	Moderate
0.51-0.75	High
0.76-1.00	Highest

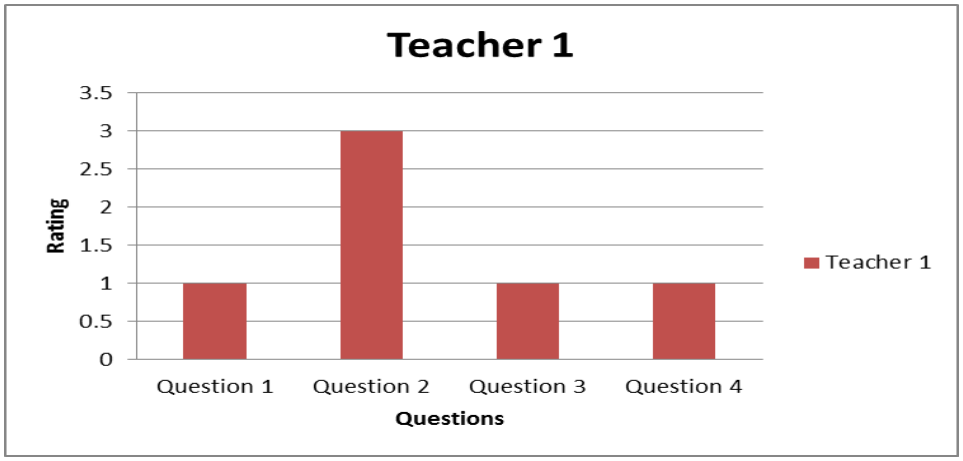
Paper 006 was rated 0.162 which is between 0.01-0.25 hence the rating was rated low.

4.9: OBJECTIVE 2: Research Findings-Teachers Rational for Choosing Commercial Exams

Teacher 1

Figure 4.9.1 shows that teacher 1 rated the source of mathematics exams are set by mathematics teacher with 59%-69% credibility and the source is reliable. The strategic plan was concluded to be minimizing the number of exams in a year; this is to ensure that the mean scores are satisfactory.

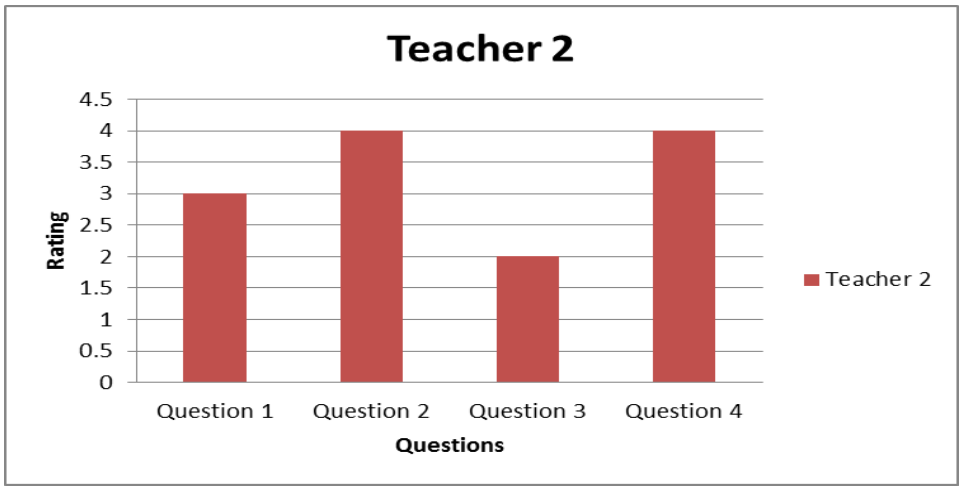
Figure 4.9.1: Teacher 1: Rational for choosing commercial exams.



Source: Research Findings

Figure 4.9.2 shows that teacher rated the source of mathematics exams are borrowed from other schools with 49%-59% credibility and the source is not reliable. The strategic plan was concluded to be change the exam source, this is to he ensure that the mean scores are satisfactory.

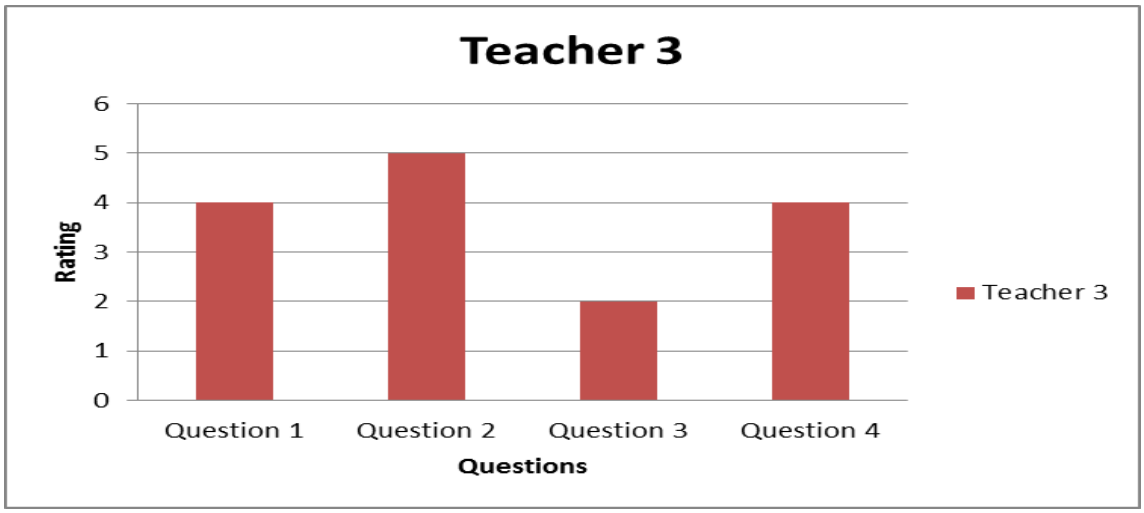
Figure 4.9.2: Teacher 2: Rational for choosing commercial exams.



Source: Research Findings

Figure 4.9.3 shows that teacher rated the source of mathematics exams as purchased from commercial dealers with below 48% credibility and the source is unreliable. The strategic plan recommended was to change the source of internal exams since the mean scores where unsatisfactory.

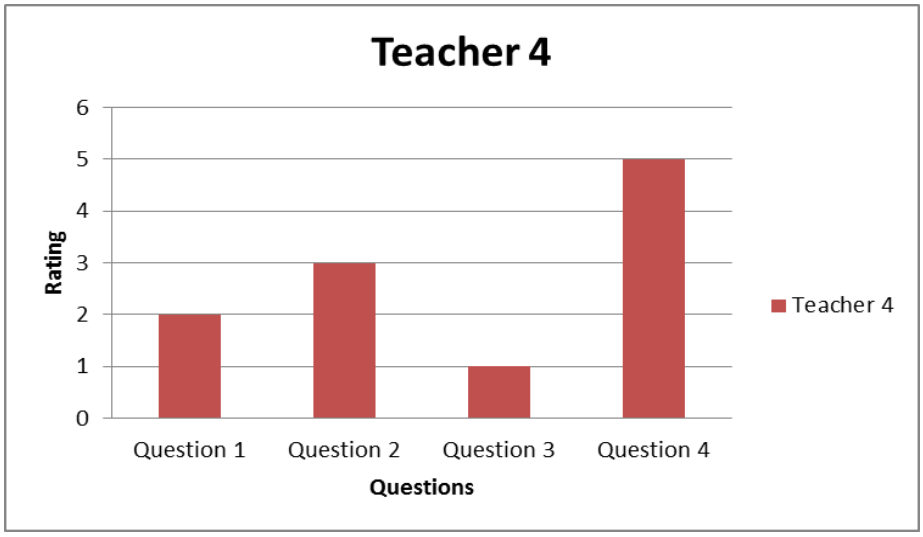
Figure 4.9.3: Teacher 3: Rational for choosing commercial exams.



Source: Research Findings

Figure 4.9.4 shows that teacher rated the source of mathematics exams as set by the mathematics panel with 59%-69% credibility and the source is reliable. The strategic plan was not recommended since the mean scores where satisfactory.

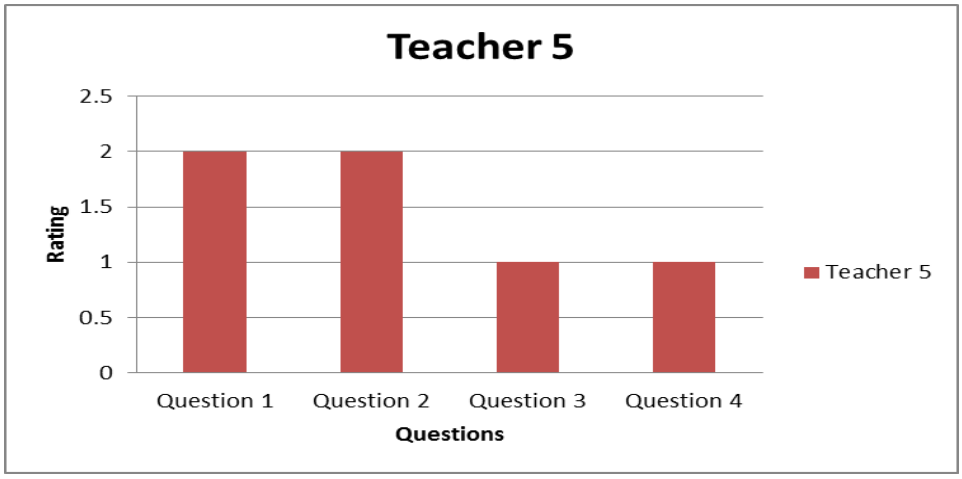
Figure 4.9.4: Teacher 4 Rational for choosing commercial exams.



Source: Research Findings

Figure 4.9.5 shows that teacher rated the source of mathematics exams are set by mathematics panel with 69%-79% credibility and the source is reliable. The strategic plan was concluded to minimize the number of exam, this is to he ensure that the mean scores are unsatisfactory

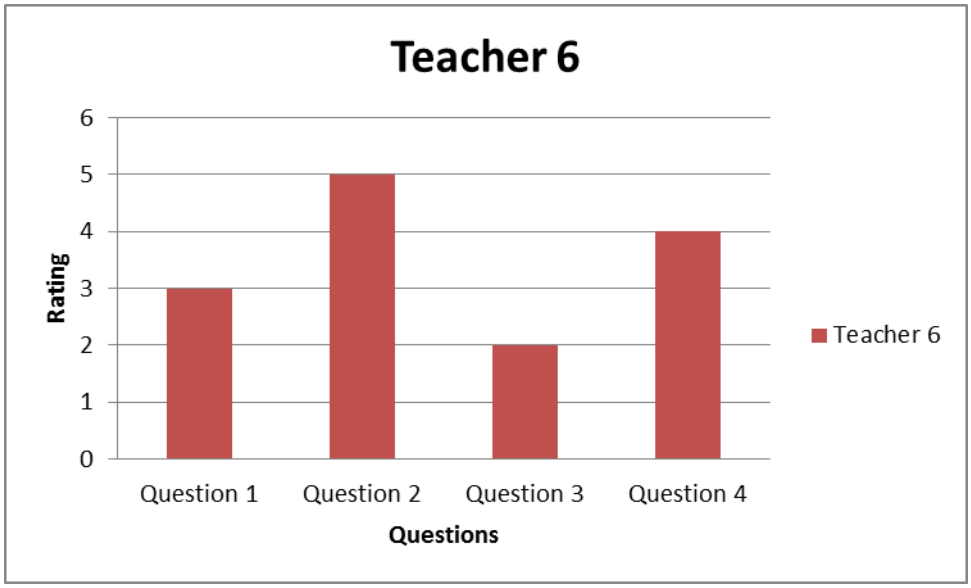
Figure 4.9.5: Teacher 5: Rational for choosing commercial exams



Source: Research Findings

Figure 4.9.6 shows that teacher rated the source of mathematics exams are borrowed from other schools with below 49% credibility and the source is unreliable. The strategic plan was concluded to change the source of exam, this is to he ensure that the mean scores are satisfactory.

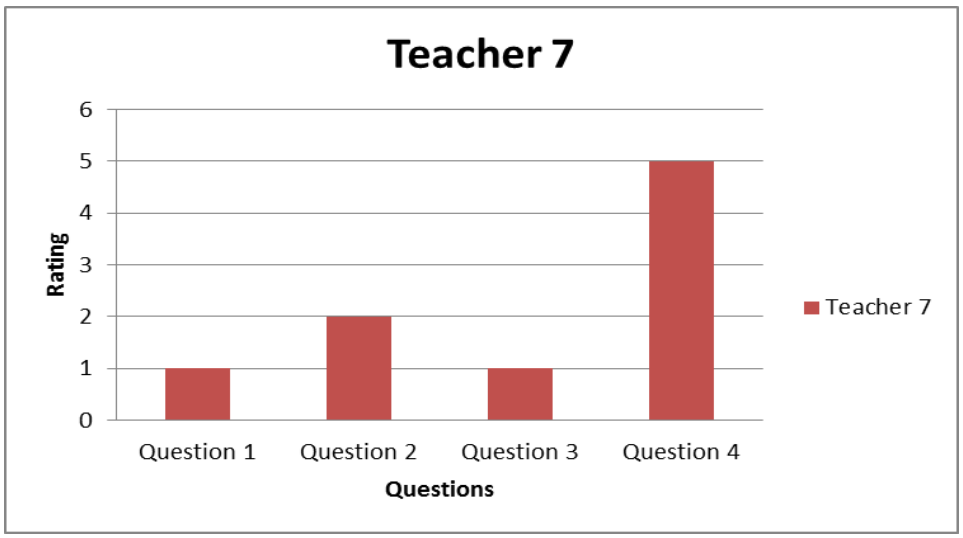
Figure 4.9.6: Teacher 6: Rational for choosing commercial exams



Source: Research Findings

Figure 4.9.7 shows that teacher rated the source of mathematics exams are set by mathematics teacher with below 69%-79% credibility and the source is reliable. The strategic plan was not concluded hence the mean scores where satisfactory.

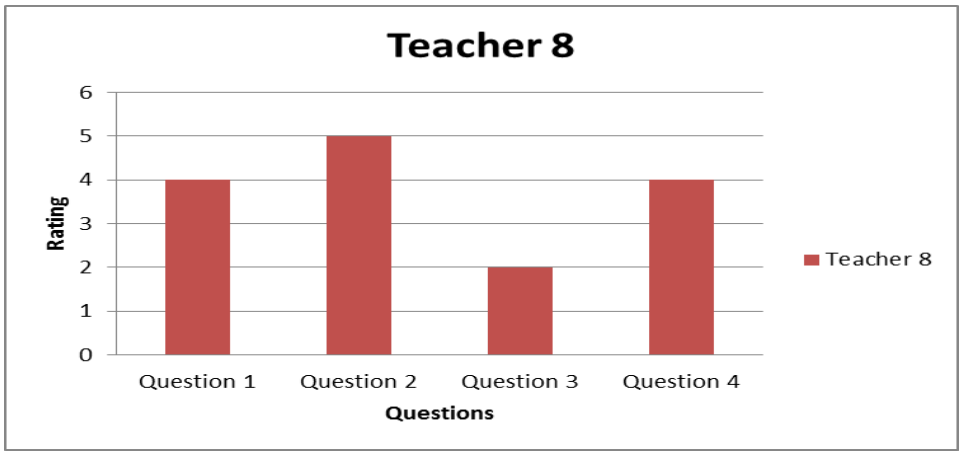
Figure 4.9.7: Teacher 7: Rational for choosing commercial exams



Source: Research Findings

Figure 4.9.8 shows that teacher rated the source of mathematics exams are purchasing from commercial dealers with below 49% credibility and the source is not reliable. The strategic plan was concluded to be change the exam source, this is to ensure that the mean scores are satisfactory

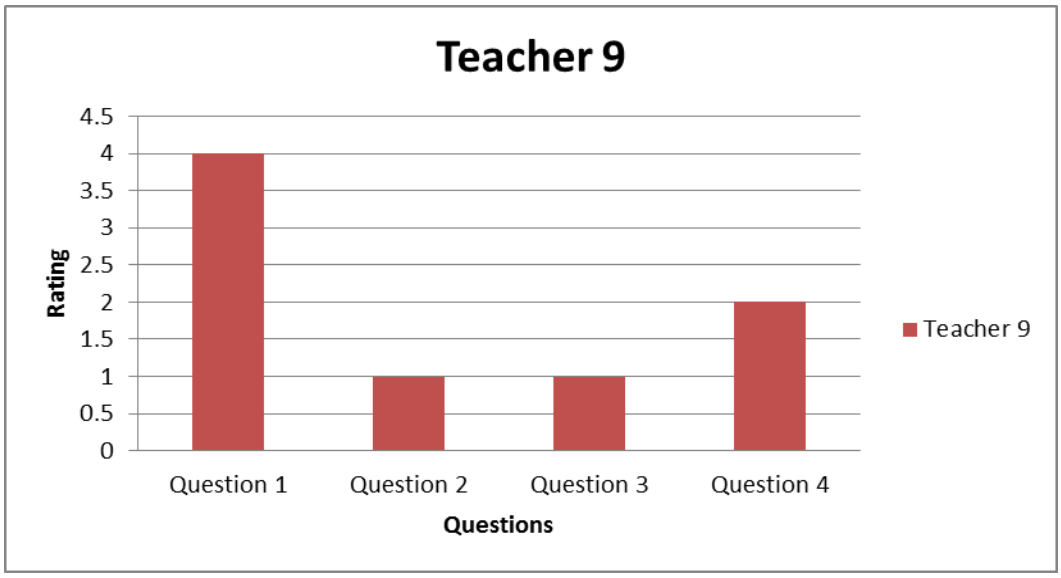
Figure 4.9.8: Teacher 8: Rational for choosing commercial exams



Source: Research Findings

Figure 4.9.9 shows that teacher rated the source of mathematics exams are purchasing from commercial dealers with below above 80% credibility and the source is reliable. The strategic plan was concluded to increase the number of exams, this is to ensure that the mean scores are satisfactory.

Figure 4.9.9: Teacher 9: Rational for choosing commercial exams



Source: Research Findings

4.10 Conclusion

From the findings of this study, it can be concluded that the reliability and content validity of commercial exams is low, and also the lack of test construction knowledge by the teachers has contributed to the use of commercial exams in assessing the pupils.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter gives a summary of the study, draws conclusions from the study and also makes recommendations as to how to address the critical issues that emerge from the study.

5.2 Aim of the study

The study was to determine the reliability and content validity of commercial tests and their correlation to pupils' performance in mathematics: the case of public primary schools, Isinya District, Kajiado County. The research was conducted using three sets of commercial mathematics papers -0006, 008 and 009, which were evaluated in three schools in Isinya Sub-County. Nine Mathematics teachers were involved in the analysis of the three set of papers and also indicated their respective mathematics mean scores for the years 2013 and 2012 in all the commercial exams done for 1st, 2nd and 3rd term and also the K.C.P.E mean score respectively.

5.3 Summary of the findings

The objective of the study was to determine the reliability and content validity of mathematics commercial tests used in public primary schools in Isinya District. The research aim was attained and concluded that the commercial exams are not reliable and the content validity is questionable, where according to research findings there was a mean score difference of 6.2 % with the K.C.P.E exam.

From the general findings of the study in addressing objective one on whether the commercial exams are reliable and have the content validity, respondent rating using Fleiss Kappa, all the three papers where rated low with k=0.115, 0.157 and 0.162 for

paper 006, 008 and 009 respectively. This proves that the commercial papers used in schools are not set to the required standard as compared to K.C.P.E. Due to the sub-standard of commercial papers the mean scores are relatively higher than K.C.P.E with an average mean difference of 6.2 %.

Basing on the literature review (Sharky and Murnane 2003), teachers lack of knowledge and skills on how to design a valid formative assessment test, and how to make inferences about students' knowledge and skills from the results of a well-developed assessment, is a factor highly contributing to use of poor assessment instruments by teachers. From the research finding only 33.3% of the mathematics teachers have the test construction knowledge. It was also observed that teachers use the commercial exams because of time factor because there is a defined number of formal assessments to be given in a given school term (Pascal M. Kagete, 2013). Hence teachers use the Mathematics commercial exams in testing their learners.

This study was able to come up with a proper link between the formative exams (commercial exams) as used by different schools as the real exams were analyzed in comparison to the real summative exams (K.C.P.E.) done in the years 2013 and 2012, which indicated the average mean difference of 6.2%.However the number of raters was small where a total of 9 raters were used, with 3 raters analyzing a set of the exams. For better results more raters like 10 could be more applicable and more commercial exams.

5.4 Policy Recommendations

From the research is evident that commercial papers lack reliability and content validity within the syllabus. The ministry of education should put in place measures and guidelines of exam content within exam publishers. However, it can provide

exams to schools through the county administration that will be using qualified personnel in setting exams.

The research shows that only 66.7% of the mathematics teachers lack knowledge on exam content and validity evaluation only 33.3 % of the teachers received training on test development. The school administrators and ministry should provide frequent trainings on test development and evaluation.

5.5 Suggestions for Further Research

Research is recommended to establish the extent to which Commercial papers affect the actual K.C.P.E performance in all schools in Isinya Sub- County.

A further research may be done to unveil other reasons that hinder public schools from purchasing standard set exams.

In future, a similar study can be done on reliability and content validity of commercial tests and their correlation to pupils' performance in other subjects.

A research can be done relationship between reliability and content validity of commercial tests and their correlation to pupils' performance in mathematics: the case study of urban setup.

5.6 Conclusion of the study

The general performance in Mathematics in the formative tests (commercial papers) is higher than that of the Summative exam (K.C.P.E) where the scores are not closely related. This is in spite that the teachers use these papers with the aim of trying to make the learners put into practice what they have been taught, the results after a lot of work is done is quite unsatisfactory.

The study also concluded that teachers use a lot of time during these examinations administration. This is evident from the research findings that almost all the schools administer over 10 exams, this means that the learners are not given enough time to internalize what has been taught and also take their time in self-evaluation. Few teachers have the knowledge of test construction and therefore just go ahead to purchase mathematics commercial exam without even scrutinizing their reliability and content validity. This therefore shows that teachers use assessment instruments that are of low quality, leading to unintended feedback which give faulty information and consequently lead to inappropriate action.

REFERENCES

Black, P. & William D. (1998).assessment & Classroom Learning. Assessment in Education: Principles, Policy and Practice, Carefax, Oxfordshire.

B. Bell, B.Cowie. The Characteristics of Formative Assessment in Science Education. Article. Revised September, 2000.

CAESL:http//www.caesl.org.

DH Caro et al (2009) Socioeconomic status and Academic Achievement. Trajectories From Childhood to Adolescence. A Canadian Journal of Education 32, 3(2009): 558 590.

Edward G. Carmine (1979). Assessment & Evaluation.

Elliott, S.N. Kratochwill, T.R. Cook, J. L. & Travers, J.F. (2000). Educational Psychology: Effective Learning, 3rd edition. Boston: McGraw Hill.

Edward H. Haertel, (2013) Reliability and Validity of Inferences about teachers based on Student test scores.

Eggen, P.P. Kalichack. D. (2004). Educational Psychology: windows on Classroom, 6th edition. New Jersey: Pearson.

Gallagher, J. D. 1998. Classroom Assessment for Teachers. New jersey. Prentice Hall.

Hogan, T.P. (2007). Educational Assessment: A Practical Introduction. Danvers, Wiley. Kenya Education Management (KEMI) Module 6: page 66-68

http://en.wikipedia.org/wiki/Classical_test_theory

James W.S. & James H. (1997). Understanding and Improving Classroom Mathematics. An overview of TIMMS Phi Delta Kappan .

June Thomas & Cathy Stockton (2003) Socioeconomic Status, Race, Gender & Retention: Impact on student achievement.

Kinyua and Okunya, (2014) Validity and reliability of teacher made tests; case study of year II. Physics in Nyahururu District of Kenya. African Research Journal Vol(2) Page 61 – 71 May 2014

Kothari CR (1990). Research Methodology Methods Techniques, 2nd edition, New Age International Limited, N.W. Delthi

Miller (1995) Coefficient Alpha: A basic introduction from the perspectives of classified test theory & structural equation modeling.

Musau Susan M. (2004). Factors Influencing Pupils' performance in K.C.P.E in Central Division, Machakos District, Unpublished Thesis, UON.

Mhairi McAlphine, Blue paper 1, February 2002 Principles of Assessment, edited by CAA Centre, University of Inton

National Council for curriculum and assessment-NCCA (2004) Assessment in primary schools in Ireland

No Child Left behind Act of 2001, L. No. 107-110, 115 Stat. 1425 (2002).

Oluwatayo, James Ayodele (Ph. D) Validity and Reliability Issues in Educational Research. Journal of Educational and Social Research Vol 2, (2) May 2012)

Pascal M. Kagete (2013). 186 classroom "Assessment for leaning" in secondary schools in Kenya.

Paul R Sacket et (2009). Research Report No. 2009.1 Socioeconomic status and the Relationship between the SAT and Freshman GPA: An Analysis of Data from 41 Colleges and Universities.

Pearson Assessment report, 3003 Fundamentals of Standardized testing

Professor John Polesel, Ms Nicky Dulfer, Dr. Malcom Turnbel (2012). THE EXPERIENCEOF EDUCATION; The impacts of high stakes testing on school students and their families.

Quality Education for Development Education Reforms- Recommended by taskforce (TF)on the Re-Alignment of the Education sector to vision 2030 and the constitution of Kenya 2010 (Jan 2011)

Rosenthal, R. and Rosnow, R.L (1991) Essentials of Behavioral Research: Methods and Data Analysis Second Edition. McGraw – Hill Publishing Company, PP 46 – 65

Rubio, Berg Weger, Tebri Lee and Rauch, 2003 Objectifying Content Validity: Conducting a Content Validity Study in Social Work Research. Social Work Research. 27 (2):94 – 104

Sharkey &Murnane (2004) Leaning from Student Assessment Results: Lessons for New York State.

Shiundu (1987), Learning of mathematics in Public Primary Schools in Bomet Central Division, Bomet County.

Sireci, S.G. (1998). The construct of content validity. Social indicators research. 45 (1 – 3): 83 - 117

Smarter Balanced mathematics items Specification in High Schools (April, 2012).

Solter, George J. Jr (2010) "A criterion related validity study of the Eighth Assessment and High School proficiency Assessment in Mathematics for a B District group school in New Jersey

Stiggins R.J.C (2007). Evaluation Classroom Assesses Training in Teacher Education of Educational Measurement, Issues and Practice 18.

Stiggins R.J.(2001). Student Involvement Classroom Assessment, 3rd Edition.

wikipedia.org/wiki/content validity modified on 7 August 2014

www.michaelmillerend.wm/res500_lectuenotes.reliability.and.validity.pdf

^ Traub, R. (1997). Classical Test Theory in Historical Perspective .*Educational Measurement: Issues and practice 16* (4), 8-14. doi:doi:10.1111/j. 1745-3992. 1997.tb00603.x

^ **a b** Hambleton, R., Swaminathan, H., Rogers, H. (1991). *Fundamentals of Item Response Theory.*Newbury Park, California: Sage Publications, Inc.

---------2004. Classroom Assessment and evaluation <http://www.bced.gov.bc.ca/classroom_assessment/classes.htm>

APPENDIX I: Fleiss Kappa Paper 006

Q	R1	R2	R3	R4	R5		Pi
	1	**2**	**3**	**4**	**5**	**TOTAL**	
Q1	0	1	1	0	1	3	-
Q2	1	0	0	1	1	3	-
Q3	0	1	0	2	0	3	0.333
Q4	0	0	3	0	0	3	1.000
Q5	0	1	1	1	0	3	-
Q6	0	0	1	2	0	3	0.333
Q7	0	2	0	1	0	3	0.333
Q8	0	1	1	1	0	3	-
Q9	0	2	1	0	0	3	0.333
Q10	0	1	1	1	0	3	-
Q11	1	0	1	1	0	3	-
Q12	2	0	0	1	0	3	0.333
Q13	1	1	1	0	0	3	-
Q14	0	0	2	0	1	3	0.333
Q15	0	1	1	1	0	3	-
Q16	0	1	1	1	0	3	-
Q17	0	0	2	1	0	3	0.333
Q18	0	1	0	2	0	3	0.333
Q19	0	0	2	0	1	3	0.333
Q20	0	0	2	1	0	3	0.333
Q21	1	0	2	0	0	3	0.333
Q22	0	1	1	1	0	3	-
Q23	0	0	2	1	0	3	0.333
Q24	0	0	1	2	0	3	0.333
Q25	0	1	0	2	0	3	0.333
Q26	0	0	3	0	0	3	1.000
Q27	0	0	0	3	0	3	1.000
Q28	0	1	2	0	0	3	0.333
Q29	0	0	3	0	0	3	1.000
Q30	0	0	2	1	0	3	0.333
Q31	0	0	0	3	0	3	1.000
Q32	0	1	0	1	1	3	-
Q33	0	2	0	1	0	3	0.333
Q34	0	0	3	0	0	3	1.000
Q35	0	1	2	0	0	3	0.333
Q36	0	0	1	2	0	3	0.333
Q37	0	0	2	1	0	3	0.333
Q38	0	1	2	0	0	3	0.333
Q39	0	0	2	1	0	3	0.333

Q40	0	0	3	0	0	3	1.000
Q41	0	0	2	0	1	3	0.333
Q42	0	2	1	0	0	3	0.333
Q43	0	2	1	0	0	3	0.333
Q44	2	0	0	1	0	3	0.333
Q45	0	1	2	0	0	3	0.333
Q46	0	1	0	2	0	3	0.333
Q47	0	3	0	0	0	3	1.000
Q48	0	2	1	0	0	3	0.333
Q49	2	1	0	0	0	3	0.333
Q50	0	0	0	2	1	3	0.333
Total	10	33	59	41	7	150	18.333
P1	0.067	0.22	0.393	0.273	0.047	1.000	

Source: Research Findings

Appendix II: Fleiss Kappa Paper 008

Q	R1	R2	R3	R4	R5		Pi
	1	**2**	**3**	**4**	**5**	**TOTAL**	
Q1	3	0	0	0	0	3	1.000
Q2	0	1	2	0	0	3	0.333
Q3	0	0	2	1	0	3	0.333
Q4	2	1	0	0	0	3	0.333
Q5	0	0	2	1	0	3	0.333
Q6	0	2	1	0	0	3	0.333
Q7	1	2	0	0	0	3	0.333
Q8	0	0	2	1	0	3	0.333
Q9	1	2	0	0	0	3	0.333
Q10	0	2	1	0	0	3	0.333
Q11	0	0	3	0	0	3	1.000
Q12	0	2	1	0	0	3	0.333
Q13	0	1	2	0	0	3	0.333
Q14	0	1	2	0	0	3	0.333
Q15	0	0	2	1	0	3	0.333
Q16	0	0	1	2	0	3	0.333
Q17	0	0	2	1	0	3	0.333
Q18	0	0	1	2	0	3	0.333
Q19	0	0	1	2	0	3	0.333
Q20	0	1	2	0	0	3	0.333
Q21	0	2	0	1	0	3	0.333
Q22	0	1	1	1	0	3	-
Q23	0	1	2	0	0	3	0.333
Q24	0	2	1	0	0	3	0.333
Q25	0	1	2	0	0	3	0.333
Q26	0	0	3	0	0	3	1.000
Q27	0	2	1	0	0	3	0.333
Q28	0	1	1	1	0	3	-
Q29	0	0	3	0	0	3	1.000
Q30	0	0	2	1	0	3	0.333
Q31	0	0	3	0	0	3	1.000
Q32	0	2	1	0	0	3	0.333
Q33	0	0	1	2	0	3	0.333
Q34	0	0	1	2	0	3	0.333
Q35	1	2	0	0	0	3	0.333
Q36	1	2	0	0	0	3	0.333
Q37	0	2	1	0	0	3	0.333
Q38	0	0	3	0	0	3	1.000
Q39	0	1	2	0	0	3	0.333

Q40	0	0	2	1	0	3	0.333
Q41	0	0	2	1	0	3	0.333
Q42	0	0	1	2	0	3	0.333
Q43	0	1	0	2	0	3	0.333
Q44	0	0	3	0	0	3	1.000
Q45	2	1	0	0	0	3	0.333
Q46	0	0	2	1	0	3	0.333
Q47	0	0	1	2	0	3	0.333
Q48	0	3	0	0	0	3	1.000
Q49	0	1	2	0	0	3	0.333
Q50	0	0	0	2	1	3	0.333
Total	**11**	**40**	**68**	**30**	**1**	**150**	**21.333**
P1	**0.073**	**0.267**	**0.453**	**0.2**	**0.007**	**1**	

Source: Research Findings

APPENDIX III: Fleiss Kappa Paper 009

Q	R1	R2	R3	R4	R5		Pi
	1	**2**	**3**	**4**	**5**	**TOTAL**	
Q1	0	1	0	0	2	3	0.333
Q2	0	0	1	2	0	3	0.333
Q3	1	0	0	1	1	3	0.000
Q4	0	1	0	2	0	3	0.333
Q5	0	1	2	0	0	3	0.333
Q6	0	0	2	1	0	3	0.333
Q7	0	0	2	1	0	3	0.333
Q8	0	0	3	0	0	3	1.000
Q9	0	1	0	0	2	3	0.333
Q10	1	0	0	1	1	3	0.000
Q11	0	1	0	1	1	3	0.000
Q12	0	0	2	1	0	3	0.333
Q13	0	1	0	2	0	3	0.333
Q14	0	1	0	0	2	3	0.333
Q15	0	0	0	1	2	3	0.333
Q16	0	1	0	1	1	3	0.000
Q17	0	0	2	1	0	3	0.333
Q18	0	1	1	1	0	3	0.000
Q19	0	0	2	1	0	3	0.333
Q20	0	0	1	2	0	3	0.333
Q21	0	0	1	2	0	3	0.333
Q22	1	1	0	1	0	3	0.000
Q23	0	0	3	0	0	3	1.000
Q24	0	1	2	0	0	3	0.333
Q25	0	0	3	0	0	3	1.000
Q26	0	1	2	0	0	3	0.333
Q27	0	0	0	3	0	3	1.000
Q28	0	3	0	0	0	3	1.000
Q29	0	2	1	0	0	3	0.333
Q30	0	1	2	0	0	3	0.333
Q31	0	0	2	1	0	3	0.333
Q32	0	1	2	0	0	3	0.333
Q33	1	0	0	2	0	3	0.333
Q34	0	1	0	2	0	3	0.333
Q35	1	0	2	0	0	3	0.333
Q36	0	0	2	1	0	3	0.333
Q37	0	1	2	0	0	3	0.333
Q38	0	1	2	0	0	3	0.333
Q39	0	2	0	1	0	3	0.333

Q40	2	0	0	1	0	3	0.333
Q41	0	0	1	2	0	3	0.333
Q42	0	2	1	0	0	3	0.333
Q43	1	1	1	0	0	3	0.000
Q44	0	1	2	0	0	3	0.333
Q45	0	0	3	0	0	3	1.000
Q46	1	2	0	0	0	3	0.333
Q47	0	0	0	3	0	3	1.000
Q48	0	2	1	0	0	3	0.333
Q49	0	0	1	2	0	3	0.333
Q50	0	0	1	0	2	3	0.333
Total	**9**	**32**	**55**	**40**	**14**	**150**	**19.00**
P1	**0.06**	**0.213**	**0.367**	**0.267**	**0.093**	**1**	

Source: Research Findings

APPENDIX IV:

QUESTIONNAIRE I

SECTION A: Part 1: Biographical Information

1. What is your gender? Male [ ] Female[ ]
2. What is your age bracket?
    a) Less than 30 [ ]
    b) 31-40 [ ]
    c) 41-50 [ ]
    d) 50 years and above [ ]
3. What is your highest education level?
    a) Masters [ ]
    b) B.Ed. [ ]
    c) PI [ ]
    d) Any other (specify) [ ]
4. For how long have you been in the teaching profession?
    a) Less than 5 years [ ]
    b) 6-10 years [ ]
    c) 11-15 years [ ]
    d) 16-20 years [ ]
    e) Over 20 years [ ]
5. For how long have you been teaching mathematics
    a) Less than 5 years [ ]
    b) 6-10 years [ ]
    c) 11-15 years [ ]
    d) 16-20 years [ ]
    e) Over 20 years [ ]
6. Have you received any training on test development? Yes [ ] No [ ]

Part 2

1. What is the source of your internal exams?

a) Set as a teacher.

b) Set as a subject panel.

c) Borrow from other schools.

d) Purchasing the exams.

2. Given the credibility of 100%. What score can you give to the source of your internal exams you have been using?

a) Above 80%

b) 69-79%

c) 59-68%

d) 49-58

e) Below 48%

3. Comparing your K.C.P.E mean score with the scores of the internal exams, do you think the source of internal exams is reliable?

a) Yes

b) No

4. What is your strategic plan to make sure that the upcoming exams mean scores are satisfactory?

a) Minimize number of exams.

b) Increase number of exams.

c) Weak learners repeat classes.

d) Change source of internal exams.

e) No action taken.

SECTION B

1. Please judge the validity of the questions in the following question papers on a scale of 1-5 with 1 being the lowest, 2 being low, 3 being moderate, 4 being high and 5 the highest.

Question no	1	2	3	4	5
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					
11.					
12.					
13.					
14.					
15.					
16.					
17.					
18.					
19.					
20.					
21.					
22.					
23.					
24.					
25.					
26.					

27.					
28.					
29.					
30.					
31.					
32.					
33.					
34.					
35.					
36.					
37.					
38.					
39.					
40.					
41.					
42.					
43.					
44.					
45.					
46.					
47.					
48.					
49.					
50.					

QUESTIONNAIRE II: SECTION A

Questionnaire for Mathematics Teachers in Isinya District

Dear Sir/Madam,

You are invited to participate in the above mentioned research project. The survey should only take 15 – 30 minutes to complete. To ensure confidentiality of all responses, you are not obliged to provide your name. The information you give in response to this survey will be purely used for academic purpose.

Please fill in the blank spaces.

A. How has been your class 8 mathematics mean scores in the year 2013

First term	**1st exam**	**2nd exam**	**3rd exam**	**Others**
	_____	_____	_____	_____
Second term	**1st exam**	**2nd exam**	**3rd exam**	
	_____	_____	_____	_____
Third term	**1st exam**	**2nd exam**	**3rd exam**	
	_____	_____	_____	_____

What was your K.C.P.E mathematics mean score for the year 2013

_____

QUESTIONNAIRE II: SECTION B

Questionnaire for Mathematics Teachers in Isinya District

Dear Sir/Madam,

You are invited to participate in the above mentioned research project. The survey should only take 15 – 30 minutes to complete. To ensure confidentiality of all responses, you are not obliged to provide your name. The information you give in response to this survey will be purely used for academic purpose.

Please fill in the blank spaces.

B. How has been your class 8 mathematics mean scores in the year 2012

First term	**1st exam**	**2nd exam**	**3rd exam**	**Others**
	_____	_____	_____	_____
Second term	**1st exam**	**2nd exam**	**3rd exam**	
	_____	_____	_____	_____
Third term	**1st exam**	**2nd exam**	**3rd exam**	
	_____	_____	_____	_____

What was your K.C.P.E mathematics mean score for the year 2012

_____

APPENDIX V
Letter of introduction to the respondents

Lilly Sangale

University of Nairobi

Department of Psychology (Measurement and Evaluation)

P.O Box 447

Kitengela

The respondent,

RE: RELIABILITY AND CONTENT VALIDITY OF COMMERCIAL TESTS AND THEIR CORRELATION TO PUPILS' PERFORMANCE IN MATHEMATICS: THE CASE OF PUBLIC PRIMARY SCHOOLS, ISINYA DISTRICT, KAJIADO COUNTY.

I am a student at the University of Nairobi pursuing a master of education degree in **Measurement and Evaluation** conducting a research to the above topic. I am kindly requesting you to respond to the questionnaires as honestly as possible.

The questionnaires are meant for this research only and the response given will be treated with utmost confidentiality. To ensure this, no name of the respondent or school will be written on the questionnaire. I look forward to your honest participation. Thank you in anticipation.

Yours faithfully,

LILY SANGALE

Cell phone: 0722437687

Email: lillysankalek@gmail.com