

**RELATIONSHIP BETWEEN TYPE OF ASSESSMENT PROCEDURE OF
AGRICULTURE PROJECT AND THE RELIABILITY OF STUDENT
SCORES IN AGRICULTURE IN MATUNGU SUB-COUNTY**

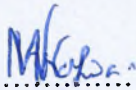
**BY
MILDRED WERE
E58/84934/2016**

**A PROJECT SUBMITTED IN PARTIAL FULFILLMENT FOR THE
DEGREE OF MASTER OF EDUCATION IN MEASUREMENT &
EVALUATION IN THE DEPARTMENT OF PSYCHOLOGY IN THE
UNIVERSITY OF NAIROBI**

2020

DECLARATION

This research project proposal is my original work and has not been presented for a degree in any other university.

Signature 

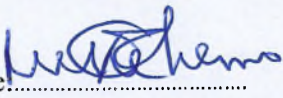
Date 16/11/2020

MILDRED WERE

E58/84934/2016

SUPERVISOR'S APPROVAL

This research project has been submitted for examination with my approval as university supervisor.

Signature 

Date: 17/11/2020

DR. LUKE ODIEMO

Senior Lecturer

Department of Psychology

University of Nairobi

DEDICATION

This project is dedicated to my late dad Patrick Were Chesino, mum Flora Auma Were and my son Austin Amani.

ACKNOWLEDGEMENT

To the Almighty who miraculously reigned on the COVID – 19 situation enabling re-opening of schools which provided an opportunity for data collection.

My sincere thanks and appreciation goes to my supervisor Dr. Luke Odiemo for his expert and subtle guidance when I seemed confused, for the detailed feedback on work submitted and most of all encouragement, motivation and patience. Your supervision was exceptional! Dr. Karen Odhiambo; course Coordinator; your guidance at the initial stage of undertaking this research inspired me. Thank you.

Special appreciation goes to the Sub-county Director of Education; Matungu Sub-county, the school principals, the coordinator of joint examinations in the sub-county and subject teachers who played a crucial role in availing the required data.

I acknowledge my fellow colleagues and friends whose moral support came in handy throughout this study. Much appreciation to Shadrack who dedicatedly conducted data analysis and research assistants who assisted in data collection. My son and family on whom I selfishly imposed long hours of absence, thank you for sacrificing and trusting in me.

ABBREVIATIONS

CA	Continuous Assessment
CBC	Competency Based Curriculum
CTT	Classical Test Theory
EAC	East African Community
EAEC	East African Examination Council
EFL	English as First Language
ESL	English as a Second Language
GCSE	General Certificate of Secondary Education
JSC	Junior Secondary Certificate
KCPE	Kenya Certificate of Primary Education
KCSE	Kenya Certificate of Secondary Education
KICD	Kenya Institute of Curriculum Development
KNEC	Kenya National Examinations Council
ME	Measurement Error
SEM	Standard Error of Measurement
MOE	Ministry of Education
PA	Performance Assessment
SBA	School Based Assessment
SSC	Senior Secondary Certificate
TA	Teacher Assessment
TS	True Score
WAEC	West African Examinations Council

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
ABBREVIATIONS	v
TABLE OF CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	x
ABSTRACT	xi
CHAPTER: ONE INTRODUCTION	1
1.1 Background of the study.....	1
1.2 Statement of the problem.....	5
1.3 Research Questions.....	7
1.4 Objectives.....	7
1.4.1 Overall Objective.....	7
1.4.2 Specific Objectives.....	7
1.5 Hypothesis.....	8
1.6 Justification of the Study.....	8
1.7 Significance of the Study.....	9
1.8 Scope of the Study.....	9
1.9 Research Limitations and Delimitations.....	11
1.10. Research Assumptions.....	12
1.11. Definition of Terms.....	12
CHAPTER: TWO LITERATURE REVIEW	14
2.0 Introduction.....	14
2.1 The Education System in Kenya.....	14
2.1.1 KNEC Examination Body.....	15
2.1.2 The Agriculture Curriculum in Kenya.....	16
2.1.3 International Approaches in Assessment of Agriculture.....	17
2.1.4 Assessment of KCSE Agriculture Project.....	20
2.2 Reliability.....	22
2.3 Forms of Reliability.....	25
2.3.1. Test Stability.....	25
2.3.1.1 Test-retest.....	25
2.3.1.2 Alternate-form Reliability (Equivalence).....	26
2.3.2 Internal Consistency.....	27
2.3.2.1 Split-half Reliability.....	27
2.3.2.2 Inter-rater Reliability (assessed by Kappa).....	27
2.4 Factors that Affect Reliability of Agriculture Project Scores.....	30

2.4.1 Test Taker Characteristics	30
2.4.1.1 Gender.....	30
2.4.1.2 Test Wiseness of the Student	32
2.4.1.3 Family Background.....	32
2.4.1.4 Learners with Disability	33
2.4.2 Characteristics of the Assessor	34
2.4.2.1 Prior Experience in Assessment.....	34
2.4.2.2 Prior Training in Assessment	35
2.4.2.3 Teacher Qualification.....	35
2.4.2.4 Gender	36
2.4.2.5 Age.....	36
2.4.3 Testing Environment	37
2.4.4 Test Characteristics of Secondary School Agriculture Project	37
2.4.4.1 Test Length	38
2.5 Theoretical Framework.....	38
2.5.1 Classical Test Theory	38
2.6 Conceptual Framework.....	40
CHAPTER: THREE: RESEARCH METHODOLOGY	42
3.0 Introduction.....	42
3.1 Research Design.....	42
3.2 Study Site	42
3.3 Target Population.....	43
3.4 Sample Size.....	43
3.5 Sampling Procedure	44
3.6 Research Instruments	45
3.7 Data Collection Procedure	46
3.7.1 Quantitative Data.....	46
3.8 Validity and Reliability.....	47
3.9 Data Analysis	48
3.10 Ethical Considerations	49
CHAPTER FOUR: DATA ANALYSIS AND RESULTS.....	50
4.0 Introduction.....	50
4.1 Summary Statistics.....	50
4.2 Establishing the average of scores of the teacher and inter-rater score and determining their reliability in relation to theory examination.....	58
4.3 Establishing the inter-rater concordance of the scores generated in agriculture project	60
4.4 Correlating subject teacher scores and scores generated by the inter-raters.....	61
4.5 Establishing reliability coefficient of the teacher versus the inter-rater	61
4.6 Establishing the strength of association between teacher score, inter-rater score, and the theory examination score	62

CHAPTER FIVE: DISCUSSION, CONCLUSION AND RECOMMENDATIONS.....	63
5.0 Introduction.....	64
5.1 Internal and External Validity.....	64
5.2 Summary of Findings.....	65
5.3 Discussion of the Findings.....	66
5.3.1 Establishing the average of scores of the teacher and inter-rater score and determining their reliability in relation to theory examination.....	66
5.3.2 Establishing the inter-rater concordance of the scores generated in agriculture project.....	67
5.3.3 Correlating subject teacher scores and scores generated by the inter-raters	68
5.3.4 Establishing reliability coefficient of the teacher versus the inter-rater.....	68
5.3.5 Establishing the Strength of association between subject teacher, inter-rater and theory examination score.....	69
5.4 Conclusion	69
5.5 Recommendations.....	70
5.6 Recommendations for Further Research.....	72
REFERENCES.....	73
APPENDICES	81
Appendix I: Sample Size Form.....	81
Appendix II Score Sheet for Collection of Student’s Scores.....	82
Appendix III Agriculture Teacher Questionnaire	83
Appendix IV Inter-Rater Questionnaire.....	84
Appendix V: Agriculture Project 2020 Assessment Sheet	85
Appendix VI: Sample Question Paper for Sub-County Joint Exams Agriculture Paper 1	86
Appendix VII: Letter of Introduction – UON.....	91
Appendix VIII:Nacosti Research License	92
Appendix IX:Matungu Sub-County Permission Letter	93

LIST OF TABLES

Table 1: Sample size per cluster	44
Table 2: Distribution of the gender of the respondents	50
Table 3: Distribution of the age of the respondents	51
Table 4: Summary statistics for inter-raters	52
Table 5: Summary statistics for subject teachers	53
Table 6: Proportions on whether respondent attended training on assessment of agriculture projects	53
Table 7: Number of students taught	54
Table 8: Time spent on agriculture projects	55
Table 9: Records and pictorial evidences	56
Table 10: Marking scheme provided by KNEC	57
Table 11: Summary statistics on the continuous variables	57
Table 12: Distribution of the proportions of the independent variable in the dependent variable	58
Table 13: Paired t-test on reliability of subject teacher scores and theory scores	59
Table 14: Paired t-test on reliability of inter-rater scores and theory scores	59
Table 15: Regression results on effect of subject teacher and inter-rater scores on theory	60
Table 16: Fleiss' Kappa inter-rater concordance	60
Table 17: Correlating subject teacher scores and scores generated by the inter-raters	61
Table 18: Pearson moment coefficient between subject teacher scores and inter-rater scores	61
Table 19: The strength of association between teacher score and theory examination score	62
Table 20: The strength of association between average inter-rater score and theory examination score	62

LIST OF FIGURES

Figure 1: Structure of organization of Education and training in Kenya.....	15
Figure 2: Conceptual frame work	41
Figure 2: Distribution of gender of respondents	51
Figure 3: Distribution of the age of the respondents.....	52
Figure 4: Whether respondent attended training on assessment of agriculture projects.....	54
Figure 5: Number of students taught	55
Figure 6: Time spent on agriculture projects	56
Figure7: Whether respondents keep records and pictorials	56
Figure 8: Whether respondents use KNEC marking scheme.....	57
Figure 9: Distribution of the proportions of the independent variables in the theory examination scores.....	59

LIST OF IMAGES

Image 1: Kakamega County Image: 2 Matungu Sub-County.....	43
--	----

ABSTRACT

A number of study findings have indicated teacher biases in performance based assessment. This causes inconsistencies in the scores assigned by teachers in such assessments posing serious reliability concerns. The current study sought to establish the relationship between type of assessment procedure of agriculture project (assessment by subject teacher or inter-rater) and reliability of students' scores in theory examinations in Matungu Sub-county. Correlation design was employed and survey was conducted to collect both quantitative data from a clustered sample of 12 schools implying 12 subject teachers of agriculture and 2 inter-raters who were purposively sampled. A total of 380 agriculture project work samples for all registered students across the sampled schools were awarded a score by the subject teacher and another set of score by the two inter-raters. The teachers and inter-raters completed a self administered questionnaire to collect data on their demographic factors. The subject teacher, inter-rater and theory examination scores for each student were captured/entered into the students' score sheet for collection of students' scores. Data on teachers' demographic factors was analyzed using descriptive statistics. The measures of central tendencies and measures of dispersion for the continuous confounding variables were also presented. The data on scores was analyzed using descriptive statistics and STATA version 16 software for paired t-tests, regression analysis, Fleiss Kappa inter-rater concordance, correlation coefficients, Pearson's moment coefficients and chi square test. The study revealed a statistically significant relationship between inter-rater scores and theory examination scores ($\beta=.5237$, $t=4.14$, $p=0.000$). While the relationship between subject teacher and theory examinations was ($\beta=.4280$, $t=3.18$, $p=0.002$). The regression analysis on the influence of teacher and inter-rater scores yielded results of $R^2 = 0.3271$. The inter-rater concordance of the scores generated by the inter-rater established a Fleiss Kappa concordance of .40004 implying moderate agreement between the two inter-raters' scores. The Pearson's moment coefficient between the subject teacher and inter-rater score was .8535 indicating a strong positive correlation between the variables. Chi square test and Cramer's V statistics on the strength of association between subject teacher, inter-rater and theory examination scores yielded results that indicated statistically significant ($\chi^2=1300$, $p=0.007$, $\Phi=0.3577$) while that between subject teacher and theory examination was statistically insignificant at ($\chi^2=1540$, $p=0.285$, $\Phi=0.3437$). The statistically significant association between the inter-rater and theory scores revealed that the scores were highly reliable.

CHAPTER ONE

INTRODUCTION

The chapter provides study background encompassing reliability of performance assessments, teacher biases in the assessment and multiple marking as an alternative assessment procedure for students' projects, problem statement, study objectives, research questions, hypothesis, justification, study significance, scope, limitations of the study and finally definition of terms.

1.1 Background of the study.

There is a growing concern among stakeholders on reliability of students' scores in internal assessment of agriculture project in comparison to the students' performance in written agriculture examinations. Mwanyumba and Mutwiri (2009) discovered that School Based Assessment (SBA) marks from teachers were unreliable, where teachers tended to bend the assessment criteria and at times submitting fake marks, there is lack of uniform facilities making assessment rather subjective than being objective, and teachers were noted to be dishonest. Mwanyumba and Mutwiri note that all the above led to inaccurate scores which failed to correlate with attainment in the same theory tests in the ultimate examination. SBA marks are thus scaled down (moderated) using theory papers.

Teacher assessments (TA) are recurrently blamed due to the fact that they are liable to biasness, in relation to issues like gender and student aptitude as renowned by (Hoge and Butcher 1984). Spear (1984) conducted research which indicated that transcribed work may be assessed differently by teachers based on their knowledge of student's gender. They consider the outcome to be high especially when the assessment is conducted one on one. In an inclusive combination assessment and testing research review, Wood (1991) cited various studies of different kinds of biasness in (TAs). Considerable Studies have shown biases of scorers to be as a result of how they view student working patterns in respect to their societal affair.

Other problems associated with performance assessments like in the case of agriculture project may include; ambiguous or inaccurate performance indicators,

unsuitable scoring processes, inferior exercises, and training of raters (Stiggins, 1994). A number of researchers argue that tests that evaluate performance of learners on a given task are more valid. (Darling-Hammond, 2006; Darling-Hammond & Snyder, 2000) compared two conventional ways of teaching capability. Further it stresses that raters have to score subjectively. Additional related matter of concern not adequately considered is how reliable or valid assessment performance is interpreted, measured, and disclosed.

The objective of SBA is to verify how accurate and reliable are the outcomes of student performance at the end of cycle examination, assessment of affective, psychomotor and cognitive domains of the learner, and to develop effective and productive learning habits in the learner (Bello & Tijani, 2003). Ideally, the scores generated by the subject teacher in the assessment of agriculture project should be comparable to students' score in theory agriculture examination. As a matter of fact, when humans are used in the evaluation process, usually the concern is whether the results are reliable. It is for this reason that Bull and Kimball (2000) argue that people are well known for their disparity. Fisher, Brooks and Lewis (2002) support this view and argue that fitness for purpose is the core of all testing work and SBA assessments are subjective thereby more prone to reliability issues.

Reliability can be perceived as an indication that administered tests are free from inaccuracies. Ebel and Frisbie (1991) define reliability as the accurateness or dependability of a measure. It is evident that when the random error is sized to a minimal level, score preciseness and reproducibility can be generalized to supplementary evaluation tests and related tests. Reliability, therefore, is viewed as how dependable a measure is. It can be deduced that it is the estimate of scores accuracy and dependability. In other words, the degree to which a score measures the behaviour being assessed rather than other factors that cause score variation. In this case, a score is thought to be reliable if we would get duplicate results if the test were done on different occurrences. This implies that it should be possible then for student scores in agriculture project to be correlated positively to the theory agriculture examinations. Although Greenberg (1992) asserts that no test score is completely reliable because every testing setting differs. Sources of error deep-rooted in any measurement setting comprise of deviations in the behavior of test

taker (like illness and lack of sleep), uncertainty in the administration of the test (inadequate space) and disparity in rater scoring behaviour (tolerance or cruelty). Where such factors are under control the scores should be positively correlated.

According to Brown and Coughlin (2007), it is the ability of one assessment tool to forecast subsequent performance either in the same activity (like an accomplishment in college) or on another assessment of the similar construct. The predictive reliability of survey instruments and psychometrics tests is considered by scholars to be a measure of agreement between outcomes gathered from more direct and unbiased measurements. The predictive reliability generally gauged by the correlation coefficient between two sets of measurements obtained by same target population.

Daniels and Schouten (1970) and Owoyemi, (2000) note that a prediction on the subsequent exam could be shaped with fair success on grounds of the results of previous examinations. However, there are divergent views on predictability of some examinations. The West African Examination Council (WAEC 1990, 1993) discovered that the scores awarded by teachers in SBA were higher than what the students scored in external examinations, a clear implication that teachers were too generous in awarding scores. SBA scores also appeared to be clustered together indicating an effort by the teacher to ensure each candidate was close to the maximum score in the class. This rendered the SBA scores so unreliable that the WAEC reduced the weighting of SBA from 40% to 30% (Bello), besides moderating them before incorporating them into the final grading in an effort to improve their validity and reliability.

Andala, Digolo, & Kamande (2014) did a research directed towards determining reliability of mock examinations in terms of quality assurance factors and the ability of mock examinations to predict candidates' results in the KCSE examinations. The study population was all the secondary schools in the country. The researchers employed survey research method and used questionnaires and unstructured interview guide to collect data. Stratified random sampling was used to obtain 65 schools per category which represented all respondents. Data was analyzed through SPSS while descriptive statistics was also employed in data analysis. Correlation

coefficients were calculated in investigating the connection amid mock and final results. Among other findings, Andala et al. (2014) study showed strong positive linear correlation between the mock and KCSE examination. Both the Pearson correlation and the spearman's rho correlation gave a high positive correlation of .949 and .942 respectively significant at 0.01%. As a result of the high positive correlation, the study concluded that mock examinations were reliable. The study also recommended the need for harmonization of structures of setting, moderation and invigilation a move they argue will make it more rigorous.

A number of researchers have had contradictory findings on the relationship between assessment procedure and the reliability of theory examination. Othuon and Kishor, (1994) carried out an investigation on this and found CPE marks have a reasonable affirmative linear connection with the CSE grades. Other states in Nigeria, performance in JSC examinations were established to be undoubtedly correlated to performance in SSC examinations, (Adeyemi, 2001). Through the study it was established that the rule of choosing students intended to join secondary school is in line with supposition that students who excel in the selected exams have a high probability of achievement in secondary schools. In their study, there was no variance from school to school. In other words, future performance in the Kenya national examination is purely predictable from performance of teacher made examinations.

In another study, Shohamy, Gordon, and Kramer (1992) carried out a study to compare the holistic scores awarded by trained English Second Language (ESL) teachers with untrained English L1n laypeople. Shohamy et al. established that there were no differences between the groups regarding the scores awarded to letters written by secondary level English Foreign Language (EFL) learners. Moreover, other groups of laypeople and ESL teachers were trained as raters, results showed that training remarkably increased inter-rater reliability. Although the degree of improvement was the same for both groups.

The study suggests that training had a compelling influence on marking. Although not any influence was established in markers history. This scenario was constant throughout the marking process. The research implies that raters are in a position to

rate reliably despite a teacher's experience or training. It was also evident that it was enhanced as a result of more markers being involved.

“the practical implication of this finding is that decision makers, in selecting raters, should be less concerned about their background, since that variable seems not to increase reliability. More emphasis, however, should be put into intensive training sessions to prepare raters for their task.” (p. 31)

Findings in the various studies indicate that, the assessment procedure approves or taints the evaluation process, the end results endowed to students and its consistency. The question that arises is whether the agriculture project scores are reliable. How do the scores assigned by the subject teacher in the project work compare with those of standardized test score? The most direct comparison between performance in agriculture project and achievement is to compare student score in the project to achievement in theory examination which are standardized. Ideally there would be a strong correlation between students score in the theory examination and in the project work. As highlighted earlier, recent studies have shown that scores in agriculture project are not consistent with the same student scores on standardized tests in this case the agriculture theory examinations paper one and two.

Consistency is a necessary factor in establishing validity and reliability (Cresswell, 2005). Inconsistencies in the assessment of agriculture project should cause alarm. Inconsistencies can indicate the application of inappropriate assessment procedure or could be a result of teacher own subjectivity in marking the project work. The current study intends to examine the consistency of project scores and gain an understanding of the alternative by incorporating the use of inter-raters in the assessment of the agriculture project. Inter-rater concordance will be based on to establish whether the scores will have a stronger correlation to students' theory examination.

1.2 Statement of the Problem

The project component that is agriculture paper three (443/3) is a paper that contributes to the final student score in KCSE and is assessed at the school level by the subject teacher. Reports from KNEC reveal that moderation has had to be carried out to moderate the scores assigned by the teachers in the project work. This

has been as a result of the inconsistencies in the marks attained by students in the project work as compared to marks attained by the same student in the theory papers which are externally assessed. Such inconsistencies impact negatively on the assessment practices of the project work.

Following the inconsistencies, the KNEC carried out training and workshops for the teachers involved in the assessment of subjects with a project component in 2019. The training and workshops were designed to arm teachers with the applicable information and skills in the assessment of projects so as to improve their skills in assessment and uphold reliability of agriculture project scores.

According to Sawilosky (2000) reliability is a necessary condition and forerunner to validity. This statement implies that tools used can be more valid than they are reliable. The argument is; it is not possible for one to determine whether a mechanism measures the construct intended if it produces totally unpredictable results. Validity is made possible by reliability hence predictability; an evaluation mechanism that yields unreliable outcome cannot deliver applicable information concerning what is being measured. The marks must be alleged to be fair, predictable in dissimilar settings, patterns, or dissimilar raters of similar performance or instrument will have negligible use.

Numerous marking has the possibility of improving marking dependability for particular questions especially those having some degree of biasness in marking. As cited by Tisi, Whitehouse, Maughan and Burdett (2013), there is also a hypothetical review that consolidates double marking to yield an ending mark acknowledging genuine disparities in judgement can happen amongst scorers.

Therefore, the current study focuses on the relationship between the assessment process (by the subject teacher or inter-raters) and reliability of students' score in theory examination. As noted by Sawilosky (2000) when scores are not in line with the testing procedure, the scores are treated as having been determined by random errors of measurement. The relationship between scores and other variables will be a weak one and will not be a precise reflection of scores uses and analysis that are crucial for validity. The argument in the current study is that perhaps with the use of

multiple raters and establishing inter-rater concordance of the raters and comparing this to the subject teacher score it is possible to improve comparability between school-based assessment scores and students' final score in agriculture thereby enhancing the reliability of the scores. Specifically, the study seeks to establish the relationship between assessment procedure (either by inter-raters or subject teacher) and reliability of students score in agriculture.

1.3 Research Questions

The questions that guided the study are:

- (i) What is the average score of the teacher and inter-rater assessment procedures and its reliability to theory examination?
- (ii) What is the degree of concordance of scores generated by inter-raters?
- (iii) Is there any correlation between the scores generated by the teacher and inter-rater scores?
- (iv) What is the reliability coefficient of the teacher versus the inter-rater scores?
- (v) What is the degree of association between teacher scores, inter-rater scores and the scores in theory examination?

1.4 Objectives

1.4.1 Overall Objective

Overall objective of the study was to examine the relationship between type of assessment procedure of agriculture project and reliability of students' scores in theory examinations in Matungu Sub-County.

1.4.2 Specific Objectives

The specific objectives of the study are to:

- i. Establish the average of scores of the teacher and inter-ratter score and determine their reliability in relation to theory examination.
- ii. Establish the inter-rater concordance of the scores generated in agriculture project
- iii. Correlate subject teacher scores and scores generated by the inter-raters
- iv. Establish reliability coefficient of the teacher versus the inter-rater – Pearson moment coefficient

- v. Establish the strength of association between teacher score, inter-rater score and the theory examination score.

1.5 Hypothesis

The study tested the following hypothesis:

Ho.: There is no association between, the agriculture project scores generated by subject teacher and inter-rater in relation to theory examination score.

1.6 Justification of the Study

Fearnley (2005) spelt out that reliability is the degree of compliance with an independent score of a high-ranking examiner. Yet, several writers contend that increasing the number of markers offers a further accurate rating of candidates' true marks than specific mark awarded by the various scorers (Brooks, 2004). Head; Lucas; Wood and Quinn (as cited in Tisi et al., 2013) empirical and theoretical views assert that multiple marking is more reliable than an individual score. Plinner (1969) as cited in Tisi et al. (2013) showed statistically that so long as there was a fair similarity amongst the markers, the mean of the numerous marks is a credible interpretation of marking teams' agreement. Further, it is noted that reliability proportionally increases with team size. Analysis done on increasing the number of markers established a considerable increase in reliability was as a result increasing the size of the marking teams.

Therefore, it is evident that the aspect of inter-raters can improve reliability of PA by dealing with the challenge of biases introduced by single assessor in this case the subject teacher. A number of researches conducted in Kenya on agriculture project have focused on challenges on implementation of agriculture practical. The researcher could not find any study that had been conducted on the relationship of assessment procedure of agriculture project and reliability of students' score in theory examinations. The current study aims at filling the gap by obtaining empirical data that will bring out the correlation of assessment procedure of agriculture project and reliability of students' score in agriculture theory examination.

1.7 Significance of the Study

SB performance evaluation has been condemned greatly generally due to the absence of reliability. (Chong, 2009). Assessment processes are essential for achieving quality output and undertakings having temporary information that are not easily appraised by paper-based tests (Black, 1995; William & Black, 1996). Scoring of performance evaluation possess challenges since it frequently encompasses verdict. The subjectivity of performance evaluation is incredibly decreased by use of multiple raters (Airasian, 2005; Airasian & Russell, 2008; Thorndike & Thorndike-Christ, 2010), with studies indicating that the same test could be scored adversely by various teachers while still the same scorer could award marks contrarily at varying times (Rennert-Ariev, 2005). Lack of consistency between scores and the scoring procedure means that scores are rather determined by random error of measurement. Therefore, such scores will not precisely reflect score application and analysis that are crucial for validity. Following this discussion, Rudner and Boston (1994) contend that multiple ratings increase repeatability of results, in as much as multiple test items can increase the reliability of standardized tests.

Additionally, improved reliability can be of importance to educational policymakers in that they will be in a position to adopt the use of inter-raters in the assessment of project work to improve on the reliability of agriculture scores based on the findings. The Quality Assurance and Standards department and the KNEC- can use the results to explore the possibility of incorporating inter-raters in the assessment of projects to ensure the reliability of scores obtained is achieved. Finally, the study is expected to be of importance for it may set the base for more researchers and educators who would be interested in the same area given that very few studies in Kenya have researched on the issue of reliability of SBA in secondary schools in Kenya across other subjects.

1.8 Scope of the Study

The research aims at examining relationship between the process of evaluating projects and reliability of student score in agriculture theory examinations in Matungu Sub-county.

Assessment procedure of projects is the independent variable which is manifested by two attributes as either assessment by the subject teacher or assessment by inter-raters. To have a measure for this variable, the agriculture project for sampled schools will have an independent score assigned by the subject teacher and another set of scores assigned by inter-raters.

Reliability of the student score in agriculture examination is the dependent variable. For this variable to be quantifiable, the study based on the standardized externally assessed mock theory examinations. The mock examinations in agriculture which are externally set and marked are a perfect example of common standard. They measure the learner's performance using identical test. While, learners' evaluation in project work lacks standardization and mostly depends on rater's judgment on what counts as an average national accomplishment (Oberholzer, 1998). Correlation inquiry will be done in examining the direction and how powerful the affiliation is among the variables. The study will perform reliability tests to establish reliability coefficients of the variables.

Matungu Sub-county has been chosen considering the area has been one of the leading areas in sugarcane production in the country. With collapse of the giant Mumias sugar factory, the area residents are learning to diversify and venture into other forms of agricultural production being the backbone of the area's economy. In the area, only 2 out of the 42 schools do not offer agriculture as an examinable subject at KCSE. The rest offer the subject. (Matungu Sub-county Education Office, 2020). The fact that the subject is prevalent in the area makes it suitable for the study.

Following the challenges that have always been observed in the area of assessment of practical examinations by subject teachers, KNEC carried out intensive training through workshops in the year 2019 with a bid to improve assessment in this area. One of the measures put in place to curb the haphazard award of scores to candidates even in the absence of proof of project work was the introduction of two critical milestones. It will be interesting to establish if this had any positive impact on improving reliability in assessment of agriculture project hence the choice of the

year 2020. If the training yielded results, the researcher envisages that students score in the project will be highly correlated to their final score in the theory examination.

1.9 Research Limitations and Delimitations

The study main data was the students' raw scores in agriculture project and theory examinations. Examinations being an important part of students' learning cycle and considering the confidentiality and seriousness with which teachers treat students' scores, there were instances where teachers were not willing to conceal such information to the researcher. This was resolved by presentation of research permit together with the permission letter by the Sub-county Director of Education which won the confidence of the teachers. The researcher also reassured the participants that the scores were to be used strictly for academic purposes and they were assured of confidentiality of the scores. Similarly, because the raw marks in theory examination is for the sub-county it hinders national generalization of the results making the results only generalizable to the sub-county.

The main concern of this study is subjective assessments. It was expected that inter-raters would be objective in their assessment. As suggested by Armes (2016), such a study will be of no significant use to objective tests because it largely aims at increasing the reliability of raters by reducing their inherent subjectivity. In view of this, multiple-choice questions are not prone to any subjectivity in their marking therefore teachers that mostly deal with multiple-choice tests will find the study to be of minimal use. In addition, the study is not generalizable to other subjects with a project component.

The Hawthorne effect is understood as the difference in feeling/conduct by contestants which may arise simply as result of taking part in the study (Drew, Hardman and Hosp, 2008). The researcher controlled for Hawthorne effect by concealing the correlational intend of the study. The researcher assumed that intentional conformity of teachers and inter-rater rating in the project scores was therefore checked.

The researcher could not find any published instrument initially developed for the purpose of measuring reliability of agriculture assessment. The researcher developed

an instrument to capture the students' scores and a questionnaire for collection of biographical data of the respondents.

1.10. Research Assumptions

It was assumed that the raters will be objective in their scoring of the agriculture project and free from bias. It was assumed that they made use of the specifically designed marking tools and assessment criteria to enhance objectivity (sample of the marking scheme is provided under appendices). This implies that raters had the ability to use knowledge and expertise in an excellent way probable to judge student learning and not allow subjectivity where their personal perceptions and bias distort making sound judgements about student achievement.

Finally, the marks for agriculture theory examinations to be used are the definite marks assigned before they undergo standardization. The mock marks retrieved are assumed to be reliable because the education boards in the various sub-counties adopt credible systems and procedures in the setting of questions and evaluation of the examination (a sample of the agriculture mock examination for the sub-county is attached).

1.11. Definition of Terms

Agriculture:	This is a subject taught in Kenyan secondary schools. It is among the optional subjects in the applied science category.
Assessment procedure:	Refers to the procedure employed in assessing project work. Assessment as carried out either by the subject teacher or by the-inter-raters.
Agriculture project:	Refers to a task that is set up by the KNEC which can be on either crop production or animal husbandry.
Inter-rater concordance:	Consistency among observational ratings provided by multiple raters. For this study, agreement between the 2 inter-raters.
Inter-rater:	Different raters scoring work of the same student to establish reproducibility of the score.

Inter-rater score:	Score assigned to the student in the project work by the inter-rater based on the marking scheme.
Measurement error	It is an observational error, denotes the difference between students' agriculture project score and the true score the student would attain in the event of no errors.
Project Score:	The score assigned to the students' project work either by the inter-rater or the subject teacher based on the marking scheme.
Reliability.	Ebel and Frisbie (1991) denote reliability as the magnitude of scores dependability and without mistakes. Dependability or reproducibility of the scores that is free from error. The assessment procedure will be deemed as reliable if the scores are reproducible and are error free.
Theory Score.	The results of the student in the externally assessed mock agriculture theory paper 1 and 2 examination.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

Considering research questions, literature review was guided by the basic interest on reliability of agriculture project marks. A review of studies on reliability is undertaken. The review of literature is organized into the four sections: section 2.1 reviews the education system in Kenya, 2.2 reliability, 2.3 types of reliability 2.4 factors that can affect reliability of agriculture projects.

2.1 The Education System in Kenya

As noted in an MoE (2014) article, Kenyan 8-4-4 education system was adopted in 1985. The system entails that a learner undertakes 8, 4 and 4 years of study in primary, secondary and university level respectively. Children aged 3-5 years may undertake either 1 or 2 years of pre-primary education for them to transit to primary school although this is not compulsory. Since 2003, free public primary education was compulsory with science, languages, mathematics, crafts, history, religious studies and geography embedded into the curriculum. (MoE, 2014). Since its inception, the education system has employed objective centered syllabus, where importance is on end of cycle cumulative assessment. In this regard, the learners take the KCPE and KCSE on completion of primary and secondary school cycle correspondingly.

As a result of reliance on summative evaluation, there has been a drive for educational reform prompting the new competence-based structure of education which was agreed to substitute 8-4-4 curriculum. The 2-6-3-3 system comprises 2 years of Kindergarten targeting age of between 4-5 years; the following 6 years are subdivided equally for lower and upper primary. The age range for this bracket is 6-11 years; finally, the learners' transit to secondary school which also comprise of 3 years of junior and senior secondary involving learners between the ages of 12-17 years. The new CBC system deviates from the old system that overly was examination oriented and intends to cultivate each student's prospective by certifying that all students achieve the major capabilities as elaborated in basic education framework. Unlike the 8-4-4 system which relied heavily on assessment

of learning, the CBC will rely on assessment for learning. The new system is being rolled out in phases having begun with the lower grades. According to the policy document, it is hoped that the 8-4-4 system will be completely phased out by 2026. Figure 1 below illustrates the arrangement of the current 8-4-4 education System in Kenya.

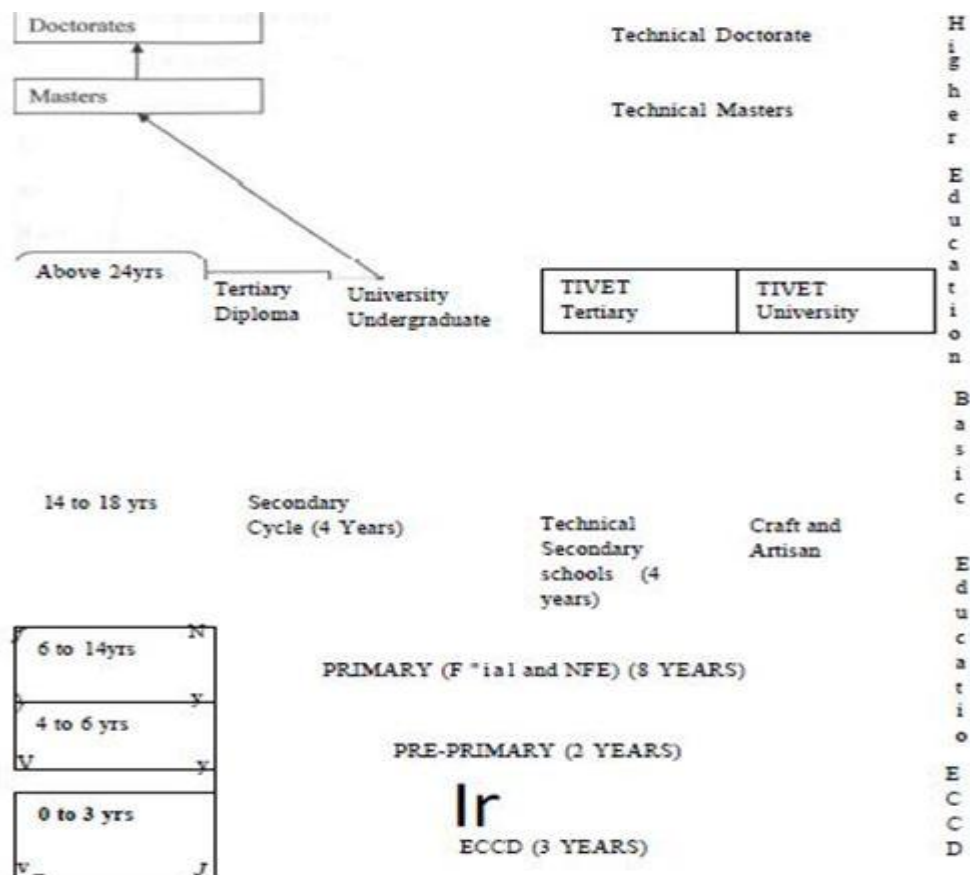


Figure 1: Structure of organization of Education and training in Kenya

Source: Ministry of Education

2.1.1 KNEC Examination Body

The Kenya National Examinations Council (KNEC) was established by an Act of parliament in 1980 upon disintegration of the East African Community (EAC). It then assumed the functions of the then East African Examinations Council (EAEC). Among other things, the Council is decreed to set and sustain examination standards and to regulate both school and post-school examinations. (<http://www.ac.ke/examinations>).

In discharging its mandate, KNEC offers school examinations among others. (<http://www.ac.ke/examinations>). School examinations are offered to students at the end of the primary school cycle as earlier mentioned KCPE and end of four years of secondary education KCSE. The KCSE school curriculum has categorized secondary school subjects into five major clusters. In the Kenyan secondary schools, candidates are expected to register for at least 7 subjects. The following is a list of the subject clusters with their KCSE codes:

- Group I (All compulsory) Languages (English (101), Kiswahili (102))
- Group II Sciences (Mathematics option A (121) and B (122), , Biology (231), Physics(232), Biology (233), Biology for the Blind (236) and General Science (237) goes hand in hand with Mathematics option B.
- Group III Humanities (History and Government (311), Geography (312), Christian Religious Education (313), Islamic Religious Education (314) and Hindu Religious Education (315))
- Group IV Home Science (441), Art and Design (442), ,Agriculture (443), Woodwork (444), Metal work (445), Building and Construction (446), Power Mechanics (447), Electricity (448), Drawing and Design (449), Aviation Technology (450) and Computer Studies(451)
- Group V: French (501), German (502), Arabic (503), Kenya Sign Language (504), Music (511) and Business Studies (565) (<http://www.ac.ke/examinations>).

All subjects in group 1-3 are theory-based, while all subjects in group 4 and 5, except Business Studies (565) are assessed via a project, oral or aural component besides the theory. Agriculture (443) falls within the group 4 subjects, with 443/1 & 2 being theory papers while 443/3 is project-based. Other subjects also assessed via project are home science, art and design, woodwork, electricity, Kenya Sign Language (KSL), drawing and design and Building and Construction.

2.1.2 The Agriculture Curriculum in Kenya.

The KICD is the institution charged with the development of the curriculum to be used in Kenyan schools. Agriculture education comprises of crop production, animal husbandry, soil and water conservation, and various agricultural aspects. Agriculture examination is offered in three papers; Paper 1, 2 and 3 which test students'

proficiency in comprehending agricultural foundations, perceptions and processes spelt out in the curriculum. Immense knowledge and skills are so as to highlight contrasting aptitude of the students. The papers are in the following format:

- Paper 1 (443/1): Paper 1 is theory based and comprises of overall agriculture, Crop Production, Agricultural Economics, Soil and Water Conservation with three section A, B and C, marked out of 30, 20 and 40 respectively totaling to 90 marks.
- Paper 2 (443/2): Is theory based and comprises of Livestock Production, Farm Power, Machinery, Structures, tools and Equipment. It is divided into three segments A, B and C marked out of 30, 20 and 40 respectively totaling to 90 marks.
- Paper 3 (443/3): Is a project paper with two questions (A and B). Project A and B were on establishment of tree nursery and goat rearing for 2020. Candidates select and follow only one of the two projects, as already stated in consultation with their teachers and considering the availability of resources. The two project questions total to 100 marks. (2018 KCSE Examination Report, Volume 2).

2.1.3 International Approaches in Assessment of Agriculture

According to (Chong, 2009) SBA has been criticized because it particularly lacks reliability. This has led to other authors emphasizing that it is essential to harmonize teachers' talented verdict and countrywide testing for countrywide evaluation schemes to be broad, rigorous, purposeful and consequently, enhance teaching and learning (Queensland Studies Authority, 2009; Pellegrino, Chudowsky & Glaser, 2001).

As noted by Hulela (2017) since the inception of agriculture into the curriculum in Botswana in the mid 1970s, assessment via the project component has been the norm. He notes that instruction of agriculture involves hands on, field trips, simulations, writing activities and role-play with an aim of enabling learners to use such competencies attained through learning. Masole (2011) notes that assessment of practical agriculture in Botswana is marred with numerous challenges ranging

from absence of standards, unqualified teachers and resources, poor administrative support, time and financial constraints and performance standards.

Assessment of agriculture in JSS is categorized into 3 units by Hulela (2011). The first paper is an objective test with 40 items. The second paper comprises of semi-structured items, completion and essay question items. Third and last is a collection of various practical agriculture pieces of work carried out by the learners. This is the continuous assessment (CA) which as noted by Chan (2004) can be outlined by various forms, carried out for a long period. Evaluation encompasses accumulated marks of practical and investigative agriculture projects. These are students' projects. As noted by Talbert, Vaughn, Croom, and Lee (2007) students are trained and equipped with skills and knowledge in the vocational area. The practical component carried out by students has an implication on final assessment scores. It is paramount to note that the third constituent evaluation that is school-based applied instruction on rearing and production of animals and crops respectively. The projects are executed by students in school and supervised by subject teachers as required by the curriculum. It is assumed that finally students will be evaluated on all the units. A notion which has changed since the author notes that reducing the number of projects from six to one has led to assessing single component of applied agriculture events in the schools.

Conferring to Leepile (2009), assessment of practical tests in Botswana, is the responsibility of the teacher alone. However, in the case of the project the author notes that it is marked by classroom teacher then exposed to external moderation at the expiration of the course. Leepile notes that moderation investigates scores of the end produce. Moderator's scores are heftier than the teachers. Further, the use of moderators is a practice argued to increase the consistency of performance evaluation. The practice incline to lessen the validity of the evaluation procedure since the moderator has no slight understanding of who did and how it was done (Tindal & Haladyna, 2002), in this case, she/he scores the work relying exclusively on the evaluation conditions. It is important to mention that this practice is also witnessed in the Kenyan system where the project scores from agriculture project paper 443/3 are subjected to moderation for a final score to be computed.

Therefore, in Botswana, Performance evaluation is deliberated by at least (20%) of the three agriculture papers done despite SBA being acknowledged as a major education transformation (Airasian & Russell, 2008; Haynes, 2000; McMillan, 2004; Popham, 2005). Even though it is noted that efforts are made to moderate the marks, handling of performance assessment is still indicated to be characterized by diverse issues (Baku, 2008; Ravoice & Pongi, 2000; van den Merwe, 2000; Yadidi & Banda, 2008). Leepile (2009) carried out routine spot-checks to establish how performance assessment, was handled. According him, there were issues among them marks that were not retrievable meaning lack of proof of the marks assigned. Such issues were also notable in the Kenyan context. Even though only few schools were visited countrywide, via a triangulation method of data collection, standardized conventions for marking project, and training workshops, issues were considered broad.

Broadfoot (1994) alleges that teacher training to obtain the suitable skills is crucial as the society has certainty in trained teachers to convey professional and proper evaluation (Maxwell, 2004). Germany and Australia emphasize teacher skill growth to evaluate (Broadfoot, 1994; Queensland Studies Authority, 1998), bearing in mind that teachers are primarily responsible for the evaluation procedures. It is noted that even if the assessment is for certification and selection purposes, there is very little external intervention or moderation (Gasemann, 1993).

The Queensland Study Authority (2009) clearly indicates that moderation is crucial because it ensures outcomes that can be compared at the same time improves teachers' assessment abilities through the application of known standards consistently by the individuals involved. According to Klenowski and Wyatt-Smith (2008): "Moderation can no longer be considered an optional extra and requires system-level support especially if, as intended, the standards are linked to system-wide efforts to improve student learning." Queensland practiced inter-rater reliability of the moderation system where teachers and schools were responsible for learner's achievement, evaluation and reporting. Additionally, Bennett and Taylor (2004) posit that a system that review moderation of teachers' judgments by professional cooperation benefited teachers and learners' evaluation hence such technique has more than just a feature of quality assurance.

2.1.4 Assessment of KCSE Agriculture Project

Agriculture project is a school-based assessment undertaken by secondary school students who choose agriculture at form four as required by the curriculum. The project component was introduced in KCSE examination in 1989 to enhance a linkage between theoretical knowledge learnt in class and real-life agriculture experience (Nyang'au, Kibett & Ngesa 2011).

Porter and Jenelik (2011) argue that a reform in education that centres on homogeneous tests are vague and restricts rationalization in success to memorizing knowledge. This hinders inventiveness and expressive intellect which are essential skills for the 21st century (Zhao, 2009). He further contends that assessments that depend on the application of knowledge are reliable channels of evaluating a student's understanding.

In the Kenyan system, KNEC, issues instructions for implementation of project/practical based subjects for the KCSE via circular uploaded on their website. Through a circular, the procedure on downloading each year's project/practical is elaborated through the KNEC official portal.

As noted earlier the group IV subjects which are practical projects comprising of Art and Design, agriculture, wood works, metal work, building and computer studies (paper 442/3, 443/3, 444/2, 445/2, 46/2 and 451/3 respectively) are usually commenced in February each year. In a circular to schools by KNEC, in a bid to ensure that the projects/ practical are conducted as per the outlined timelines, two milestones of assessing group IV subjects were put in place and must be completed between the 31st day of march and 15th day of July each year. (KNEC circular, 2019)

According to KNEC, the subject teachers are expected to key the candidates' scores on an online platform of the practical project. The online platform has a portfolio where picture evidences are uploaded at each evaluation stage. The stages are well elaborated in the KNEC circular send to schools. This is a deviation from the past practice during which at the end of the practical project duration the subject teachers would uploaded the scores for the students. This was usually done in September/October each year. The Pictorial evidence comprises of candidates photographs and practical picture submitted in form of .gif or .jpg format. The

schools can obtain these photograph formats using a digital camera. This must clearly indicate the name of the candidate and KCSE index number. (KNEC, circular on projects).

According to the KNEC, in the year 2019, countrywide workshops and training for teachers managing group IV projects were carried out with the main objective of enhancing a valid and reliable KCSE Group IV practical project evaluation. In a report by KNEC, they introduced the two stages due to:

1. KCSE Candidates delay in commencement of the project
2. In Computer Studies it was established that students obtained projects by buying
3. Failure to carry out the projects within the stipulated specifications
4. Lack of records to proof candidates final project work
5. Lack of assessment records for the projects.
6. Lack of coherence between scores and quality of project work accomplished
7. Collective execution of the project instead of each candidate carrying out the project on their own.

Through the country wide workshops; teachers voiced their issues regarding the project-based examinations which ranged from the duration of the project paper, they also cited that though the projects required a lot of input, they finally contributed very little to the end practical scores and absence of reward for teachers evaluating the practical projects.

Students are expected to carry out the project independently. While the teachers' role is to mainly evaluate each student work throughout the practical project enactment. Teachers' assessment should be done class grounds to enable equal distribution of marks from the bottommost, middle and high performers. While this is the ideal case, there are instances where students' marks have been inflated which disadvantages students during standardization (KNEC, 2013).

2.2 Reliability

When developing assessment tools and conducting assessment, assessors need to confirm that the qualities of good evaluations are encountered. Principles of assessments require that among other qualities, an assessment is valid, reliable and fair. The current study is concerned with the relationship between type of assessment procedure of agriculture project and the reliability of student scores in agriculture in Matungu sub-county. Therefore, the study reviews reliability as one of the principles of a good test.

Reliability refers to assessment consistency and whether or not it will always deliver the same result. The following example helps in explaining assessment reliability (Sieborger & Macintosh, 1998, p. 12):

- *Reliable test is repeatable under similar settings, hence yielding the same results. For instance, when a student gave a talk and after some months, he gives the same talk again precisely, he ought to get the same outcome if the assessment is reliable;*

Reliability might be regarded as a broad word denoting different forms of score stability (Andala et al., 2014). The reliability of an assessment tool is an indicator of the stability of the tool in giving the same test scores over time (test re-test reliability), the stability of item scores in the test (internal consistency), or the stability of item ratings by different raters (inter-rater reliability) • The comparison of teacher rankings is crucial to guarantee reliable assessment. (Singh, 2004).

Broadfoot (1994) established that performance evaluation is renowned to rank in all facets of validity. Even though PAs portray remarkable facets of validity, they show serious issues with reliability and hence have been highly criticized on this basis.

The allegation of rooted validity of PA was opposed by Cizek (1991) and Mehrens (1992), who dispute and, in their views, it only covers outside validity, hence referring to how the exam seems considering the fact reliability can be easily assessed and appraised than validity. Reliability is frequently emphasized, in place of validity (Raffan, 2000). However, Woods (1991) and William (1992) maintain a

conciliation viewpoint of an agreement amid valid and reliable systems of national examination.

According to Wells and Wollack (2003) ultimate importance is given to reliable test as consideration for two main factors. Firstly, they note that reliability presents a measure of the degree to which an examinees mark mirrors random measurement error which is as a result of a number of causes. Secondly, it's also crucial to be concerned with reliability because it usually precedes test validity (Wells and Wollack, 2003). In other words, when examination marks cannot be typically allocated, it is futile to imagine the marks precisely measure the intended construct. Validity points out the magnitude of the assumptions completed from an exam (for example whether the learners are aware of the factual importance or not) is rationalized and authentic. Conclusively, validity is the psychometric property about which most researchers and educators are concerned. Nonetheless, for one to correctly assess validity for specific use of a test it can be strenuous and tedious process. In this case, reliability analysis is usually seen as the initial step in the test validation process. In view of this, if a test is shown to lack reliability, there will be no need to investigate its validity because apparently it will not be valid. On the other hand, a test with fair reliability is worthy validation study.

It is well known that the notion of reliability and validity are not autonomous in their conventional application. According to Harlen (2004) the two are usually tied together in a manner that makes reliability to precede validity. The argument is that test that lacks reliability consequently will not be highly valid; whenever there is ambiguity concerning the certainty of the test which is determined by various diverse dynamics, then the degree of measuring expectations must also be vague. In this view therefore, it inclines in attempts to raise leading to close specification and adoption to the processes hence minimum mistakes. This leads to adopting lower scope of evidence which lowers validity.

Focus of this study is reliability of agriculture project scores in KCSE examination which is derived by aggregating project scores and the scores obtained from theory examinations. There are concerns about the factors that relate to the marking procedure that impact reliability of the outcome. These range from; marking

scheme, varying markers individual behavior and procedures. It is expected that a candidate who mastered the theory content was better placed to apply it than otherwise. The KNEC policy stipulates that teachers should assess their class in a manner that follows the normal distribution where the scores are evenly distributed from lowest, average and finally highest performers". Inflating project scores disadvantages students when standardization is done (KNEC, 2013) If the agriculture project scores are reliable then there should be a relationship between a student's score in agriculture practical and the theory examination. In other words, it would be expected that students who scored highly in the project would also post excellent results overall meaning even with moderation/standardization, the students' score would not be much affected.

Reliability is always articulated arithmetically as a coefficient. High coefficient implies highly reliable with minimal mistakes while a low coefficient implies low reliability with a maximal mistake. A coefficient of 1.00 implies that the marks entirely mirror each other in both internal and external assessment. Nonetheless, it is crucial to note that no test is entirely reliable (Wikimode Foundation, 2006).

Reliability is a crucial however not a satisfactory condition for validity. This implies that valid results are undoubtedly reliable but reliable results are not automatically valid. CTT posit that maximal validity of tests is the square-root of the reliability (Magnussum, 1976). Meadows & Bellington (2005) indicates that it is frequently mentioned that validity has much significance than reliability. They allege that it's not important to measure something reliable except if an individual is aware of the concept being measured. Similarly, reliability is a prior requirement for a valid measurement. It will be imperative for an assessment to be valid when students score differs entirely or it's dependent upon scorer. Therefore, if results are not very reliable, they influence the scope of validity in contrast to validity, reliability is a statistical concept. Reliability is conveyed as a reliability coefficient through an average error of measurement mode. (Meadows and Billington, 2005)

Apart from satisfying achievements, the other greatest familiar use of tests is to predict future performance. According to Meadows and Bellington (2005) a test is useful if there is a correlation between the scores on the test (the predictor) and the

scores on whatever one is trying to predict (the criterion). In this case, the scores of the learner in agriculture project which is the (predictor) should be predictive of the candidates' performance in the theory examination.

2.3 Forms of Reliability

Meadows and Bellington (2005) posit that there are definite statistics generally employed in the reliability estimation of test scores for a collection of candidates which include test-retest, split-half, measures of internal consistency, alternate form and inter-rater reliability. These statistics are anchored on correlation coefficient. There are two broad categories of reliability that is stability and internal consistency. Stability refers to the capability of measures to stay unchanged for a duration even with unrestrained testing situations or participants. It shows the degree in which individuals scores are expected to adjust from one test to another. (Allen & Yen, 1979). An entirely reliable measure yields precisely similar scores repeatedly. When testing stability either test –retest or parallel form reliability can be used.

2.3.1. Test Stability.

As mentioned above, methods used in conducting test stability include test-retest and parallel form reliability as discussed below.

2.3.1.1 Test-retest

This form of reliability is attained by conducting identical test two times while correlating the marks. Upon administration of the test twice, the results obtained are correlated. The reliability coefficient is attained by repeating similar measure twice hence the name test-retest reliability (Graziano & Raulin, 2006). This measures the internal consistency (Allen & Yen, 1979). In the event that the reliability coefficient is great, for example, $r = 0.98$, can be concluded that both tools of measurement are free from error. Coefficient yields above 0.7, are treated as fair, and coefficient results of above 0.8, are treated as overly acceptable (Sim & Wright, 2005; Madan & Kensinger, 2017). As Wiliam (2000) posits, when a learner sits for an exam severally and also in the event that no learning had taken place, the candidates' score will not be exclusive on both occasions. The candidate's attentiveness might differ, marker might be extra or less tolerant, change of handwriting as well expression of answers maybe clarified further for the examiner to understand.

Test-retest reliability coefficient is an operational measure of score repeatability since results are explicit measures of consistency from one testing occasion to another. Due to its issues and constraints, it's not favorable practice. The requirement that the test is administered twice to the same candidates makes it expensive. To add on that, administration of the test twice consumes more time. For this reason, if the interval between the two tests is of a short duration, the students might be exceedingly steady in remembering certain questions and answers. While, a long duration is compounded with acquiring knowledge and development implying changes in the candidates.

In a study by Beck, Steer and Carbin. (1996) 26 outpatient respondents on two different therapy periods which were parted weekly yielded .93 correlation thus shows high reliability test-retest of depression record. These findings support the argument above implying that the patients were able to recall the questions and answers owing to the short duration between the administrations of the two tests. It can be concluded that perhaps if the duration was longer, the correlation coefficient would have been somewhat different.

2.3.1.2 Alternate-form Reliability (Equivalence)

This form of reliability is attained by overseeing diverse forms of evaluation instrument to similar group of individuals. The attainment from the two reliability forms can be correlated to assess the outcome for consistency through alternative way. According to DeVellis (2006), if the two versions are correlated highly then they are parallel-form reliability. Researcher presents an example basing on the intensities of worker satisfaction in an organization which can be done through surveys, detailed dialogues and focus groups, and the outcome is correlated highly. In this case, we may be sure that the measures are reasonably reliable (Yarnold, 2014). Alternate forms are related based on content and difficulty. As noted by Yarnold, (2014), the correlation of two sets of marks for same candidates provides another measure of consistency forming an extension of split half reliability. Satterly (1994) points out that although alternate form reliability is favoured by statisticians the one-off nature of almost all UK examinations hinders this.

2.3.2 Internal Consistency

This form of reliability is used in assessing the extent to which diverse test objects that inquire similar hypothesis yield related outcome. Cortina (1993) presents the two main ways of characterizing internal reliability which include inter-rater consistency and Split-half reliability.

2.3.2.1 Split-half Reliability

This type of reliability measures internal consistency magnitude by examining one half of the outcome of a set of scaled items against the other half (Ganesh, 2009). This involves single administration of the test making it more useful in lengthy tests. Once the results are obtained, the scores of half of a test are compared with the outcome from the other half. Ganesh, (2009) suggests two ways through which a test can be split in half. Among them, is by the first half and second half, or by dividing the test separately into even and odd numbers. If the halved tests produce similar outcome, means that there is internal, reliability in the test hence it's a rapid and stress-free method of establishing reliability. As noted by Chakrabartty (2013), this form of testing reliability is efficient when dealing bulky questionnaires in where all questions measure same hypothesis.

Its feebleness is seen when the result of the coefficient varies depending on the test (this problem occurs when objects are intended to be differentially difficult). Again, where speed is a factor it becomes inappropriate (That is, where the scores of candidates depend on the number of objects they reached within the allocated period).

2.3.2.2 Inter-rater Reliability (assessed by Kappa statistics)

Keyton et al. (2004) state that inter-rater type of reliability measures the magnitude in which data is gathered in a steady fashion using two or more ratings. Using this form of reliability, it is possible to provide the rating correlation attained while using a tool used by numerous markers. Livingstone (2018) notes that inter-rater reliability refers to the constancy of marks assigned by diverse markers on similar response. For this to be attained, they insist that there should be no discussion between the raters considering that reliability of their scores is being tested. As already noted, reliability is influenced by correlation from two and above individual markers. It is important to test for these forms of reliability because markers may oftenly

understand answers variedly. Livingstone suggests that raters may differ on ascertaining responses or material illustrate acquaintance within the construct or skillfulness in question. According to Tavakol and Dennic (2011), Cronbach alpha (α) is the most commonly used internal consistency measure inferred as the average coefficient of all split half. Alpha values greater than 0.7 are usually reflected as reasonable and satisfactory, greater than 0.8 reflected as moderately good, and greater than 0.9 reflect excellent consistency (Cronbach, 1951). Acceptable range of alpha in social sciences are projected from 0.7 - 0.8 (Nunnally & Bernstein, 1994).

According to an online article affirming validity and reliability of edTPA (2019) evaluation of inter-rater consistency using the Kappa-n statistics is emphasized. The article further states that Kappa statistics is extensively used in the field making it relevant in this context. Hsu and Field (2003) have demonstrated that if it is expected that scorers assign scores to categories at random but in line with normal base rates Kappa-n should precisely mirror the agreement attribute to scorers. Inter-raters in this study are trained and qualified therefore; there will be reasonable base agreements. This explains the choice of Kappa-n agreement statistic. While Gitomer, Martiniz, Battery and Hyland (2019) argue that Kappa-n statistics can sometimes appear inflated, they note that it is only in the event where scores entirely differ with base rates. However, due to training and qualification of the assessors, this condition is only theoretical and doesn't arise operationally.

Ayodele (1989) notes that inter-rater reliability was used to determine the scale to which different examiners provided rational measure of the same phenomena. According to Rudner and Boston (1994), use of more than one scorer increases reliability in a similar way that multiple test increase reliability of standardized tests. Considering that people are well known for deviations, we must ponder on how we can regulate reliability in observations of two examiners. It is crucial to determine rater reliability without considering the background. Abiri (2006) presents two main methods to measure inter-rater reliability.

Tuckman (1985) emphasizes that Inter-rater reliability is among preeminent methods of measuring reliability in the event of a single observation is by inter-rater reliability. As earlier noted, use of more than one rater increases reliability since

each marker's mistakes incline towards compensating for others mistakes (Thorndike & Thorndike-Christ, 2010). However, John (2015) posit that basing on the need for multiple raters or viewers, rating correlation of similar single viewer repetitively on diverse occasion, allows test-retest method to be used where there is one rater and no need for training. For this study, inter-rater reliability has been chosen with a view of training more raters to assess the agriculture project. Research notes that interrater reliability comes into play when scoring a test involves significant subjective judgement like in the case of the agriculture project. Marking of PA as already mentioned has an array of issues majorly due to the need to make judgement.

According to Thorndike and Thondike-Christ (2008), biasness in scoring of performance-based assessment can be notably decreased by employing more than one rater (Thorndike & Thorndike-Christ, 2010). He further claims that when a single rater is used, it is noted that one rater may award consistently higher scores while other awards lower scores for the same PA. This is witnessed with proof that same test could be scored by different markers, still the similar teacher would score answers in other ways on various occasions (Rennert-Ariev, 2005). To improve reliability, Nitko (2004) suggests that it is important to use scoring criteria.

Greator (2005), in his study to investigate whether the steadiness of evaluator verdicts is affected by evidence, collected assessors' analytic decisions on two fictitious borderline portfolios as well as the assessor's comments, devised by an assessor to be 'just competent' or 'not yet competent'. The researcher carried out an analysis of variance (ANOVA) to identify any interactions between the diverse kinds of proof, evaluator observation, observer testimony, individual statement and writings underpin knowledgable questions and answers. The study found most disagreement in judgements of witness testimony. The authors note a number of limitations in his study, including a reduced sample size (two fictitious portfolios containing 177 decisions were reviewed by 15 and 12 assessors respectively) which is generally related to statistical significance level. Despite this, the study offers a useful method for researching assessor decisions based on portfolio assessments.

2.4 Factors that Affect Reliability of Agriculture Project Scores

The following variables affecting reliability will be discussed: characteristics of test takers, evaluators, environment and characteristics of test itself.

2.4.1 Test Taker Characteristics

These will be looked at in terms of: gender, test wiseness of the test taker and family background.

2.4.1.1 Gender

Among the first studies that encouraged research into biasness in gender in assessment is one carried out by Deaux and Taynor (1973). In their study, they required undergraduate students to appraise the dialogue performance of candidates on study programs abroad. Two proficient interviewees, each from the opposite gender and similarly two others who were less proficient were appraised. In the study, the proficient male was arbitrated to be more knowledgeable compared the female colleague. On the other hand, the less proficient was arbitrated as worse performer compared to least proficient female. Contrary to this, in the event of rigorously pro-female bias as evidenced by Jacobson and Effertz (1974) female leaders were ranked as more proficient than males, despite same performance.

In Goddard-Spear's (1984) study, science teachers appraised a section of work on distillation. He had his sample as half girls and half boys. An equal number of boys and girls originally wrote their specific scripts. For this study, the scripts were randomly allocated a gender. Meaning, scripts written by a boy would be marked as female and vice-versa. The study established that scripts that were allegedly inscribed by male students hence were rated highly. Those perceived to have been written by female were scored lowly. It can be seen that this study revealed that gender bias was in play. It did not matter whether the script was competently written, if it was perceived to be of the correct gender specifically male for this study, it was automatically rated highly.

In a related study, Newstead and Dennis (1990) established that in the case where projects were to be remarked, the second examiner was more lenient in marking men's projects as compared to women. In the event of discord amongst examiners,

males most probably had their scores raised compared to their female counterparts. Though the author noted that the variations had an insignificant effect. It can be concluded that maybe the aforementioned cases used comparatively subjective rating scales rather than detailed marking scheme.

Scottish Examination Board (1992) explored examiner approaches in two subject English and History which were not sciences. The study employed precise experimental scheme. Examiners were issued with writings that differed in relation to centre records, handwritten script showing students gender and ethnic group. The findings established a compelling effect in English scripts. It was noted that printed scripts were marked lower; they attained lower mean mark as compared to the hand written ones. This was alluded to candidates' failure to use the spell check faculty meaning the scripts had errors in spelling as compared to handwritten scripts. On the other hand, in history, scripts associated with female candidates' attained higher marks than those associated with their male counterparts. This was attributed to the fact that girls attained higher scores as a result of being better at essay writing such as the case of history which is oftenly assessed by essay.

Greatorax and Bell's (2002a and b) study shows expansion of gender biasness. They explored not just the differences in the response of male and female assessors to learner's script of the opposite gender, but rather the affiliation amid self-perception of maleness or felinity of raters and scores for both male and female test takers. The study used students' scores (GCSE subjects) from English, history, design and technology as well as assessors. All assessors completed Bem Sex Role inventory which shows the magnitude of gender based response. The study revealed two compelling finding which both were related to English. Firstly, it was established that there was biasness in favor of the female by 0.5. Secondly, senior examiners appeared more considerate in marking implying that the examiners position also played a significant role.

Greatorax and Bell recommended that analysis of question papers should be done in favor of male/female students. Additionally, differences in strictness and tolerance in marking are other facets rather than examiner or candidatures gender. The study

also established that the largest source of deviation was candidates' attainment. This is the ideal way a test should be.

2.4.1.2 Test Wiseness of the Student

Experience in test taking influences reliability of test score. It is notable that when students take sophisticated tests it improves test consistency. However, when in a group, all students have varying degrees of test wiseness. This aspect leads to greater measurement errors.

In testing the hypothesis that multiple-choice tests reward the test wise students Alker, Carlson and Hermann (1969) found out that test wiseness was positively and significantly correlated with both multiple choice questions performance and ability to recognize item ambiguity. The same findings were observed by Rowley (1974) and Ebel (1972). It can be concluded that students who practice agriculture from their home background will be better placed to perform well in project work as compared to their counterparts who don't.

According to Thorndike (1951) and Stanley (1971) faulty or flawed test items used in tests provide cues which introduce variance in the results other than item content or random error. These cues may increase or depress the test takers test score and reliability

2.4.1.3 Family Background

According to Eshiwani (1983) the environment in which students live, influences and shapes their aspirations, self-esteem and motivation. He states that these aspects can either enhance or hinder their educational performance. Similarly, Shittu (2004) in the findings of his study argues that gross deprivation of social and economic wants of children by parents often results to reduced academic performance. This premise is supported by a study by Dermis (2006) which showed that congested accommodation results to reduced space where organized learning can take place leading to declining performance of Somali students in UK.

Hoxbyn Robertson, Symons and Chee (2003) argue that the level of interest in students and the position of the parent in the society sometimes influence scholars' interest in studying vocational subjects. According to them, students whose parents

were educated were less likely to study vocational or technical subjects. The researchers also revealed that a family where a student has been growing since birth exerts insightful effect on the students' career. Meaning that the family was one of the determinants of a child's career. This argument has an implication on agriculture being a vocational subject, reliability of students scores can be greatly influenced in the event that they were forced into taking up this subject

A study by Nzomo, Kariuki and Guantai (2001) shows positive correlation between the social economic status of standard six pupils and the level of their learning achievement in Kenya. From the results, it was also observed that an improvement in social economic status of the pupil led to an increased learning average score due to parents' capability to cater for students learning necessities and resources that are necessary in cultivating performance.

A study by Cecilia and Patrick (2006) in South Africa on social experience affecting the educational accomplishment of learners found out that the family structure and living space affect the academic performance of the learner either positively or negatively. This premise is supported by a study by Dermis (2006) which inclined Somali students' performance in UK to congested accommodation hence no organized way of learning could take place. This view is supported by the Kenya Economic Survey (2008) which indicated that 46.8% of Kenyans live below poverty line, this is a vibrant indicator that most parents are not able to provide for school basic needs, and therefore a conducive environment for the student to study is not availed, eventually affecting their performance.

On the other hand, a study by Nderitu (1999) found no correlation between social economic background and performance of the learner, but agree that students from poor background are regularly sent home from school due to non payment of fees. All these factors the affect students' performance can greatly affect reliability of the results.

2.4.1.4 Learners with Disability

In Elbert and Bagget's (2003) study which was carried out in Pennsylvania, they required agriculture teachers to show alleged proficiency with impaired students.

The respondents in the study were of the view that although currently perceived level was low, their desired proficiency was higher.

It was also noted that working with disabled students made educators feel less knowledgeable. Teachers advised that the growing participation in agriculture education course by disabled students calls for further training to arm them with relevant skills on handling this category of learners. Elbert and Bagget (2003) showed influence of competency on performance in Pennsylvania. The study established that entire agriculture education population would gain by relevant professional development on how to handle special needs candidates, in turn such students would have meaningful experience and make immense contribution in the society.

From the findings of the study, a number of agriculture educators felt less knowledgeable while working with disabled students. This renders such a group of students to be inadequately served by their teachers in turn impacting on the reliability of the scores obtained by such student in agriculture project work.

2.4.2 Characteristics of the Assessor

These will be discussed under: prior experience in assessment, prior training in assessment, teacher qualification, gender and age.

2.4.2.1 Prior Experience in Assessment

In a study conducted by Ruth and Murphy (1988) on differences in marking of essay between trainee teachers and experienced teachers, it was noted that trainees' teachers marked essays more severely as compared to that of experienced examiners. The study also revealed that markers background informed distinctively the various points of view for judging the essays. Sentiments that were echoed by Weigle (1999) in her findings less an experienced assessor were very strict compared to skilled assessors. In her study, training is seen as solution to strictness in marking by inexperienced examiners. This is so because the study reported that before training, inexperienced examiners were significantly strict as compared to their experienced counterparts and this varied with the essay topic.

Shohamy et al. (1992) carried out a study on marker consistency in the evaluation of EFL. The study used markers who were professional, experienced teachers at EFL, or lay people. Half Four groups of examiners that is; trained professionals, untrained professionals, trained lay and untrained lay, had relatively high inter-rater reliability been achieved. This had no relation to their training. For trained raters the reliability coefficients were higher than for untrained ones. The study showed that training appeared to have compelling impact on marking this was not the case for markers background. The findings were replicated in all the three marking procedures employed. The study found that raters could rate reliably despite their background and training. Nonetheless, reliability increased remarkably when raters were procedurally trained.

Even though the studies discussed show a relationship between examiner experience and strictness, the relationship between examiners experience and marking consistency is more inconclusive.

2.4.2.2 Prior Training in Assessment

The studies of Good (1988a) carried out a study on TA in foreign languages. The study was concerned with the variations in marks given by teachers and moderators in French oral examinations of the General Certificate of Secondary Education (GCSE). As part of a study of different ways of grading marks from differentiated examination, teachers received training on administration and scoring of oral examinations in French. A random sample of recorded examination was derived for re-scoring by moderators from the population of candidates involved (177 at 'general' level and 122 at 'extended' level) who were unfamiliar of the marks awarded by the teachers. The teachers' marks were awarded generously as compared to moderators. The correlations show an insignificant difference in deviation of the teacher and moderator and they both complied on the rank order of candidates. Greater extended level correlations were lower than the general level correlations.

2.4.2.3 Teacher Qualification

According to Makau and Sommerset, (1990) academic and; professional qualifications of teachers are compelling aspects that impact performance. According to Kiragu and Okunya (2012) higher level of education consequently

increases reliability and validity. The rise in reliability was connected to the fact that the teachers become more competent in test construction. Meadows and Billington (2007) established that past research typically failed to separately account for the impact of markers' subject knowledge, teaching experience and marking experience on marking consistency.

In view of this, Meadows and Billington (2007) in a study on marking reliability they pursued separation of the effects. Their study established that both subject knowledge and some experience of teaching showed enhanced marking reliability in GCSE English. Nonetheless, comprehensive investigation indicated that the impact of marker characteristics relied on the item that was being marked. Post Graduate Certificate in Education (PGCE) English students marked short answer items reliably as experienced examiners. On the other hand, items that needed longer answers were marked most reliably by the competent examiners.

2.4.2.4 Gender

Greatorex and Bell's (2002a and b) study shows expansion of gender biasness. They explored not just the differences in the response of male and female assessors to learner's script of the opposite gender, but rather the affiliation amid self-perception of maleness or felinity of raters and scores for both male and female test takers. The study used students' scores (GCSE subjects) from English, history, design and technology as well as assessors. All assessors completed Bem Sex Role inventory which shows the magnitude of gender based response. The study revealed two compelling findings which both were related to English. Firstly, it was established that there was biasness in favor of the female by 0.5. Secondly, senior examiners appeared more considerate in marking implying that the examiners position also played a significant role. Greatorex and Bell recommended that analysis of question papers should be done in favor of male/female students. Additionally, differences in strictness and tolerance in marking are other facets rather than examiner or candidates gender. The study also established that the largest source of deviation was candidates' attainment. This is the ideal way a test should be.

2.4.2.5 Age

Meadows and Billington (2007) established that older markers tend to mark some questions more consistently compared to young markers. They however emphasized

further research is necessary to establish the cause of this scenario. Adding to that for GCSE English, and IGCSE Biology, the study revealed that male markers are inclined to marking definite items more consistently than female markers, and vice versa (Meadows and Billington, 2007; Suto et al., 2011).

2.4.3 Testing Environment

Greatorex and Bell's (2002a and b) study shows expansion of gender biasness. They explored not just the differences in the response of male and female assessors to learner's script of the opposite gender, but rather the affiliation amid self-perception of maleness or felinity of raters and scores for both male and female test takers. The study used students' scores (GCSE subjects) from English, history, design and technology as well as assessors. All assessors completed Bem Sex Role inventory. Griswold (1990) states that if the testing environment is distracting or noisy, the test taker will not be consistent through the testing process. The testing environment in the context of agriculture project would encompass school financial resources, the security of the project, availability of land and inputs, prevailing climatic conditions and cost of the project may also influence the reliability of project scores. Nyang'au et al. (2011) argue that when financial resources are scarce, the sustainability of projects would be costly to the schools. In such a scenario, the accessibility of inputs, tools and equipment, as well as security of the project, will be affected. Availability of land or otherwise will also have an impact in the event the school lacks land, it might be forced to acquire land from outside the school threatening security of the project work.

2.4.4 Test Characteristics of Secondary School Agriculture Project

In exercise of the project the KNEC provides the options for schools from which a school will select one based on how sustainable it is. It is worth to note that not all candidates will be interested in the same project as identified by the school (Ndirangu,2000). But as Kibet puts it, the learners have to be interested in the activity they are carrying out. If candidates are engaged in a project that they are not interested in, they may not direct their effort in it. Likewise, the project options provided by the council may be suitable in some ecological conditions but maybe quite unsuitable in others. The project guidelines provided by the council are supposed to be applied uniformly but the extent to which this uniformity exists can

only be ascertained when there is an independent external assessor of the project. The reason as to why the researcher seeks to establish whether the introduction of the concept of inter-raters in the assessment of agriculture would produce scores with high correlation to students' scores in the theory examination therefore improving reliability of agriculture examinations.

2.4.4.1 Test Length

It is generally viewed that lengthy tests yield high reliability. An analogy to the old axiom that "measure twice, cut once." This can be perceived as being quite reasonable. Most instructors would not base midterm grades on students' response to one multiple-choice item. Preferably, they would test on ground of 50 varying items. It can be attributed to the fact that for any question, measurement error shows broad mean of student score as a percentage.

Although reliability increases with test length, the results can only be significant with short for example a test of five items is increased by an additional 5 items reliability will increase considerably as compared to if the original test was fifty items, in this case, the effect would be negligent

2.5 Theoretical Framework

The study employed the theory of classical test Novick (1966) and described in classic texts such as Lord and Novick (1968) and Allen & Yen (1979/2002).

2.5.1 Classical Test Theory

The classical model was detailed by Novick (1968) and Gulliksen (1950). According to Schumaker, (2010) CTT is an emancipation of the early 20th Century approaches to measuring individual difference. Charles Spearman pointed out how to correct the correlation coefficient for attenuation due to error measurement and how to achieve the reliability index needed to make the correction. His finding is deemed the beginning of Classical Test Theory.

For decades, CTT was used to determine the reliability and other characteristics of the measuring instruments. According to Hambleton, and Jones, (1993) observed score (X) from psychometric instruments are thought to be composed of a true score

(T) that represents the subjects score that would be obtained if there were no measurement errors, and an error component (E) that is due to ME. Measurement error (ME) prevents one from attaining their true score.

$$X=T+E$$

The main assumptions emphasized by the CTT are: true scores and error scores are uncorrelated; the examinees' average error score is zero and parallel test error scores are uncorrelated. CTT's assumption is that, according to Magno (2009), each individual examinee has a true score (unobservable) that would be obtained if there were no measurement errors. Because the instruments used are imperfect, the score observed for each individual may differ from the true ability of an individual. This difference between the observed score and the true score is due to error of measurement. Error is often assumed to be a normal distributive random variable.

To test takers this implies that tests are fallible imprecise tools. The score an individual obtains is called the true score of the individual. Meaning that the true score for an individual won't change, even with repeated application of the same test. The observed score for this CTT is always the true score influenced by some degree of error, the influence of this error on the score being observed can be positive or negative.

The standard error deviation is used as the fundamental error measure in CTT. In practice, the test reliability and standard deviation of the score being observed are used to estimate the standard measurement error. The smaller the standard measurement error, the more certain is the precision with which the attribute is measured which also tells us that the individual score is close to the true score. Contrarily, the greater the SE, the less accurate is the attribute being observed.

In the case of the agriculture project scores awarded by the subject teacher, it is assumed that their correlation to students' performance in the theory agriculture examination is low due to the element of measurement error. On the other hand, it is speculated that the aspect of inter-raters in the scoring agriculture project will minimize measurement error thereby having scores with a high correlation to the students' score in agriculture theory examination.

2.6 Conceptual Framework

Independent variables in this research are; school-based agriculture project score by the teacher and agriculture project scores by inter-raters. Dependent variable is an observable aspect and is measured to establish the end product of independent variable. For instance, students' score in agriculture theory examination from the mock examination.

The problem of assessing performance of practical agriculture project in Kenyan Secondary Schools has been the lack of correlation between scores awarded by the agriculture teacher and that obtained by the student in agriculture theory examination. This motivated the researcher to introduce the aspect of use of inter-raters who were subject teachers with wide experience in the assessment of agriculture project with a view of improving reliability of the scores. It is hoped that with the use of inter-raters there will be consistency in scoring hence produce scores with high correlation to the students' score in theory examination as compared to scores by the agriculture teacher hence improve reliability of agriculture project scores.

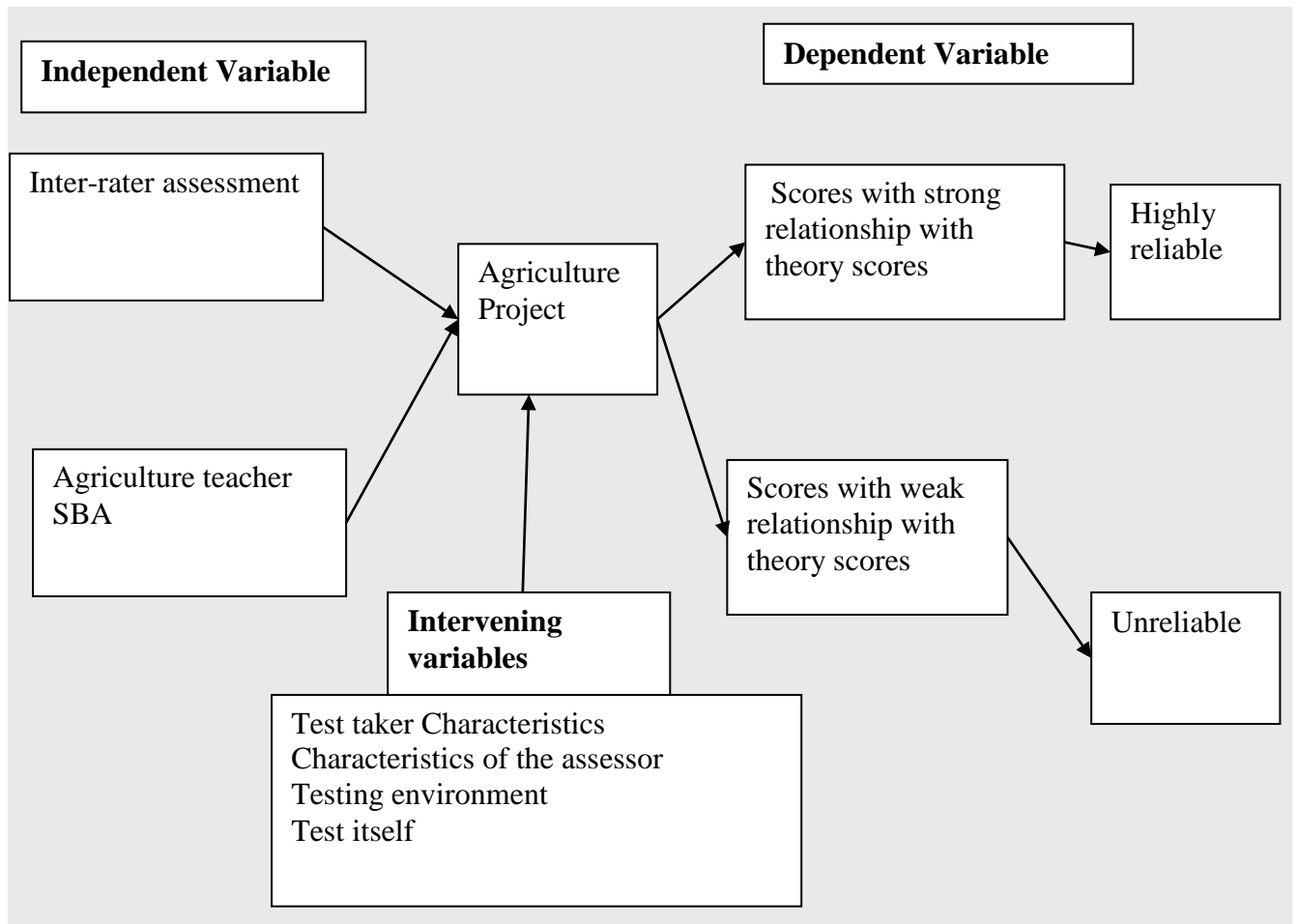


Figure 2: Conceptual frame work

Source: The researcher, 2020

CHAPTER THREE

RESEARCH METHODOLOGY

3.0 Introduction

The chapter outlines the developing strategy of the research and also explains the rationality. The chapter first introduces the research design. It goes ahead to point out data collection instruments and method of administration. Also specified is the target population and sampling procedures. Finally, the chapter will outline data analysis technique for the variables in accordance with the study.

3.1 Research Design

The study applied correlation methodology to establish the association between the variables and determine the magnitude to which subject teacher score, inter-rater score and students' score in theory examination relate.

3.2 Study Site

The study took place in public secondary schools in Matungu sub-county. Matungu sub-county is among 10 sub-counties in the larger Kakamega County. The sub-county occupies an area of 275.9sq km (approx) and has a populace of 146,563 with a population density of 598 per sq. km. (2019 Kenya Population and Census Volume 1). The area is characterized by fertile and arable land which until the collapse of the giant Mumias sugar factory, the land was mainly under sugar cane farming with most of the residents being farmers. Sugarcane farming was done on large scale this being the sole cash crop in the area and the giant income earner for the residents. The land has since been sub-divided and sold to immigrants into the sub-county, with small portions being left for subsistence farming of maize, beans, grounds, millet, sorghum and other subsistence crops. The strong reliance on agriculture could be seen as the reason why from the 42 schools in Matungu Sub-County, all the 40 public secondary offer agriculture as an examinable subject at KCSE. It is only the private schools in the area that do not offer agriculture.

Image 1: Kakamega County Sub-County

Kakamega County Map

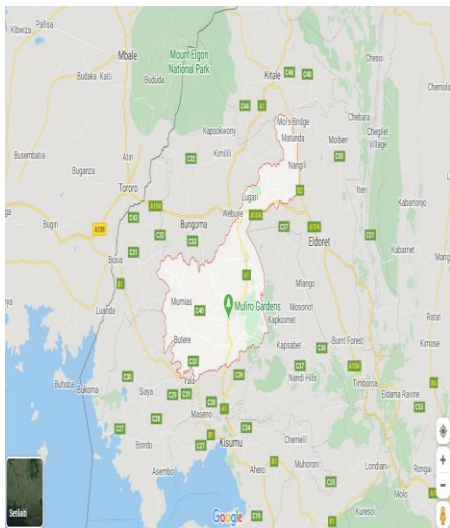
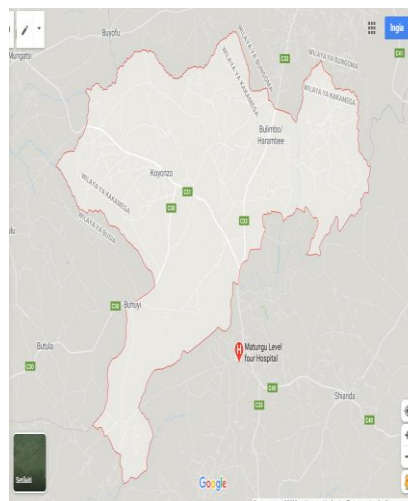


Image 2 Matungu

Matungu Sub-County Map



Site Map

3.3 Target Population

The study targets secondary schools that registered students for the 2020 KCSE examinations disrupted by COVID-19 in Matungu Sub-County. The Sub-County has an aggregate of 42 high schools. Out of this, 40 are public while only two are private and both (the two private schools) do not offer agriculture at form four. The total number of students registered for agriculture paper three 443/3 in Matungu Sub-County is 1254. (Source Matungu Sub-County Education Office).

3.4 Sample Size

The study used a sample of 12 agriculture teachers of the 12 sampled schools and 2 inter-raters who were from schools that were not sampled. Once the school was sampled, projects for all students registered for agriculture were scored by the subject teacher and the inter-raters. The sample of students' projects assessed by the teachers was calculated using a 95% confidence level and 5% error margin in accordance with table of sample size (Appendix I). The table was downloaded from

research consultants who developed it using Krejcie & Morgan's formula (1970) which is:

$$n = \frac{x^2 * N * P * (1-P)}{(ME^2 * (N-1)) + (X^2 * (1-P))}$$

n= ample size

X2= Chi-square for the specified confidence level at 1 degree of freedom

N=Population size

P=Population

ME= Measurement Error

Table 1: Sample size per cluster

The table below indicates sample size in each cluster:

S/No.	Zone	No. of School	Sample	Target Population	Sample
1	A	15	4	356	127
2	B	12	4	354	126
3	C	14	4	444	127
Total		40	12	1254	380

Source: Matungu Sub-county education office

3.5 Sampling Procedure

Being a mixed research model, cluster sampling was used. Sampling framework was the schools in accordance to the different zones in the Sub-county and school status. The clustering by zone and school status either extra-county or county schools was to take into consideration the varying students' demographic characteristics. This enabled the researcher to capture both urban and rural schools. Proportionate sampling was then conducted to have equal distribution of students across the 3 zones. From the 40 schools offering agriculture as examinable subject at KCSE in the area, 4 schools were sampled from each cluster giving a total of 12 schools with consideration of the only two extra-county schools in the area from two of the

clusters. For the sampled schools, the projects for all the students were assigned marks as already mentioned. The respondents were purposively sampled as they were to be the subject teachers charged with assessment of the form four students' project work. While the two inter-raters were also purposively sampled with the help of the Sub-county Director of Education who directed the researcher to two teachers who had a wide experience in assessment of agriculture projects having over 10 years experience. It is worth mentioning that the two inter-raters were both male.

3.6 Research Instruments

The main instrument that was used to collect students' scores as assigned by the subject teacher, inter-rater and students' theory score was a "score sheet for collection of student' score". (Appendix II). The score sheet was designed by the researcher. The score sheet captured the serial number then for ease of tracking and ensuring that the set of marks were for the same student across the three data sets, the students' index number was captured. This ensured that the students' marks were correctly matched for the subject teacher, inter-rater and the theory examination score. The "score sheet for collection of students' scores' was issued to the subject teachers and inter-raters basing on the total of students registered for agriculture examinations. This score sheet comprised of seven columns and the respondents were made aware that the assessment scores should be entered under their respective columns as displayed.

The study used students' agriculture project score as captured on the students' assessment sheet by the agriculture teacher as a total score. The inter-rater rating of the students' project was also captured as a total score. These scores were assigned following the criteria provided by KNEC on assessment of project 2020 (appendix V). The scores were subjected to test-retest reliability to establish the strength of association amongst teacher tally, theory tally as well as inter-rater tally and theory tally. Fleiss Kappa was also performed to establish concordance in inter-rater scores.

The agriculture theory examination score was retrieved from the sub-county office. This is the paper 1 and 2 theory examination sat jointly in the sub-county (appendix VI). The principal in charge of joint examinations in Matungu Sub-county availed

the score which was entered into the instrument for collection of students' scores for the respective candidates.

A questionnaire identical for both subject teacher and inter-rater (Appendix III and IV) respectively was also used to collect demographic information about the agriculture teacher and inter-raters. The questionnaire designed by the researcher consisted of 11 closed ended question divided into part I and II. Part I encompassed demographic information while part II was on general information pertaining the conduct of agriculture project in schools.

3.7 Data Collection Procedure

The researcher received an introductory letter from the University of Nairobi-Psychology department. Subsequently, authorization to conduct the research was sought from the National Commission for Science, Technology and Innovation (NACOSTI) through online application process. On receiving the NACOSI permit, the researcher presented the documents to the Sub-county Director of Education, Matungu sub-county who also drafted a letter of permission to conduct the research in the schools. Upon meeting the respondents, the aim of the research was elaborated and any concerns raised by the participants were clarified. The subject teacher was issued with the questionnaire which was self-explanatory together with the score sheet for capture of the students' score. The two inter-raters were also issued with the self-explanatory questionnaire for inter-rater and score sheets for capturing students' score.

3.7.1 Quantitative Data

A quantitative correlation study was carried out using descriptive analysis to enable the researcher measure the variables to assess the statistical relationship, According to Charles, (1988), Correlation will be used as it enables researcher to explore associations and make forecasts.

The descriptive analysis comprised of Agriculture teacher and inter-rater ratings of the 2020 KCSE students in project work. The study explored the relationship of the project marks with the agriculture theory examination results in comparison to its relationship with agriculture teachers and inter-rater rankings. Research on

correlation helps researchers to assess not only whether there is a relationship between variables but also the degree of the relationship between variables. (Gall, Borg et Gall, 1999). The relationship between agriculture teacher, inter-rater scores and theory examination scores was explored to address the research questions informing reliability of agriculture project scores.

The marks for agriculture theory examinations retrieved were original marks; real marks given/acknowledged prior to adjustment. Grounded on the method and protocols followed in the questioning and assessment of mock examinations the research supposes the marks obtained are both accurate and reliable.

Both agriculture teachers and inter-raters completed a paper and pencil questionnaire with 11 items to collect demographic data of the participants. The questionnaires were identical with part I on demographic factors including gender, age, teaching experience, teaching experience in agriculture, experience in assessment of agriculture project, and training in assessment of agriculture project.

Part II of the questionnaire gave information on the teaching of agriculture in 2020 which included the total number of students taught in the agriculture class, time dedicated to handling agriculture project work, maintenance of pictorial evidence of the students' agriculture projects as required and clarity of the KNEC marking scheme provided for assessment of agriculture project.

3.8 Validity and Reliability

Test development procedures for mock examinations incorporate processes like test specification, analysis, moderation and paneling matters that enhance validity and reliability of examinations (a sample of the sub-county mock agriculture mock examination paper is provided).

Reliability of the mock examinations the theory examinations in agriculture paper 1 and 2 is an assumption of regulated marking followed during assessment of students' response. In other words, this is a standard examination.

The marks for agriculture theory examinations retrieved were original marks; real marks given prior to adjustment. Grounded on the method and protocols followed in

the development of the question items and assessment of mock examinations the research supposes the marks obtained are both accurate and reliable.

3.9 Data Analysis

Raw data was scrutinized to ensure accuracy and completeness. Any errors and omissions noted were edited. The questionnaires were thoroughly checked and coded to enable information to be synthesized. The student scores and the responses on the questionnaires were keyed into excel and imported to the STATA version 16 software.

A mix of descriptive and inferential statistics was employed. The descriptive statistics presented frequency distributions tables, graphs, measures of central tendency by presenting means and measures of dispersion by establishing standard deviations for the variables, the minimum and maximum scores as awarded by the subject teacher, inter-raters and the theory examinations were also presented. To analyze each item of the closed ended questionnaire, frequency distributions tables indicating the means, standard deviations, percentages, minimum and maximum values and graphs to illustrate the information were generated.

Inferential statics was then employed for each objective as follows:

To study relationship of agriculture project marks by the teacher and inter-raters and the theory examination marks, the study established averages; statistics on t-test, regression analysis and test-retest correlation coefficient was calculated.

For objective 2; to study inter-rater concordance, the study employed Fleiss Kappa statistics. The statistic assumes that raters are independent observers. Coefficient k might vary from -1.0 to 1.0. If raters concur on the same or lower as projected by coincident only, k will be same as zero; and if the covenant surpasses the projected coincidental level, k will be bigger than zero; and if perfect covenant is found k approaches 1.0.

For objective four, Pearson's moment Correlation analysis was accomplished to examine the course and magnitude of relationship between subject teacher scores and scores generated by inter-raters.

For objective five; to test the hypothesis to establish the strength of association between, teacher, inter-rater and theory examination scores, chi square and Cramer's V at 5% significant level was performed

Since the questionnaires of both teachers and inter-raters consisted of open - ended, closed items, a scoring procedure was used to ensure consistency and accuracy during the data capturing process. All the closed items on the questionnaire were coded numerically and the responses captured onto the computer using Microsoft Access. The numerical code was captured on the system. The data was then exported to the STATA version 16 software.

3.10 Ethical Considerations

The researcher received an introductory letter from the Department of Psychology introducing the researcher to NACOSTI (appendix VII). A permit to conduct the research was granted by NACOSTI (appendix VIII). Upon presentation of the letters to the Sub-county Director of Education Matungu, a letter permitting the researcher to conduct the research within the sub-county was also granted (appendix IX). The two letters were presented to the principals and the agriculture teachers in the sampled schools. The teachers involved in the study gave verbal consent. The sampled teachers and inter-raters were explained to the purpose of the study and what the researcher expected of them before taking part. The respondents were assured that data generated in the research is purposefully for academic use.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

4.0 Introduction

This chapter analyses data collected with the aim of establishing the relationship between assessment procedure of agriculture project and reliability of students' scores in theory examinations in Matungu Sub-County. The findings on the demographic factors which include gender, age bracket, teacher's experience, teaching experience in agriculture, experience examining agriculture projects, training on assessment of agriculture projects, number of students taught, time spent on agriculture projects and finally the clarity of KNEC marking scheme are explained. Measures of central tendency and correlations are used to report and analyze the findings on the relationship between the dependent and independent variables. Chi squares, Cramer's V, T-tests and Pearson's correlation coefficients tested the type and strength of the relationships on variables on each objective.

4.1 Summary Statistics

Tables 2 to 11 represents summary of statistics for continuous confounding variables. Since the data is continuous, the summary gives the total observations, percentages, measures of central tendency (means) measures of dispersion (standard deviations). Finally, the minimum and maximum values are presented per category.

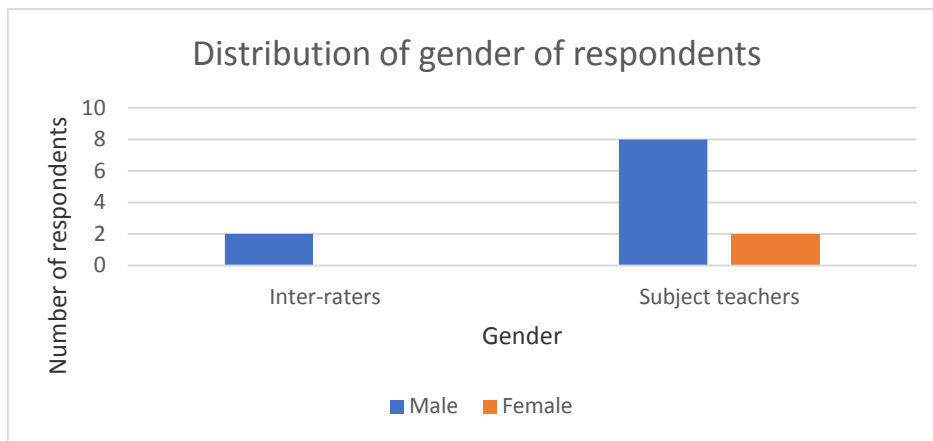
Table 2: Distribution of the gender of the respondents

<i>Gender of respondents</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Male	2 (100.00%)	8(80.00%)
Female	0 (0.00%)	2(20.00%)
Total	2	10

Table 2 presents the distribution of the gender of the inter-raters and subject teachers. All the 12 respondents revealed their gender with all inter-raters 2(100%)

being males while 8(80%) subject teachers were males and 2(20%) were female. The findings are also reflected in Figure 2.

Figure 2: Distribution of gender of respondents



It is important to note that agriculture as a science subject is usually combined with biology. Sciences are thought to be a male dominated area. This could be a possible explanation for the significantly high percentage of respondents being male.

Table 3: Distribution of the age of the respondents

<i>Age of respondents</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Less than 30	0(0%)	8(66.67%)
31-40 years	0(0%)	2(16.67%)
41-50 years	2(100%)	2(16.67%)
Total	2	12(100%)

Table 3 represents the results on the distribution of the age of the respondents. The findings indicate that all inter-raters 2(100%) were aged between 41-50 years while the majority of the subject teachers 8(66.67%) were aged below 30 years followed by 2(16.67%) who were aged between 31-40 years and 41-50 years respectively.

The inter-raters used in the study were purposively sampled on account of vast experience in assessment of agriculture and training in assessment of agriculture. Ruth and Murphy (1988) confirm this by noting that differences in marking of essay between trainee teachers and experienced teachers, trainees teachers marked essays more severely as compared to experienced examiners. Research has indicated that experienced teachers mark more consistently as compared to their novice counterparts. This is clearly reflected in the displayed data as the two inter-raters

were aged between 41-50 years implying a long duration in the area of agriculture assessment. It is also important to note that agriculture as a subject has a combination with biology. This being a science subject, a high percentage of teachers had just been recently absorbed. This explains the few years in teaching experience as compared to the inter-raters.

The findings are reflected in Figure 3.

Figure 3: Distribution of the age of the respondents

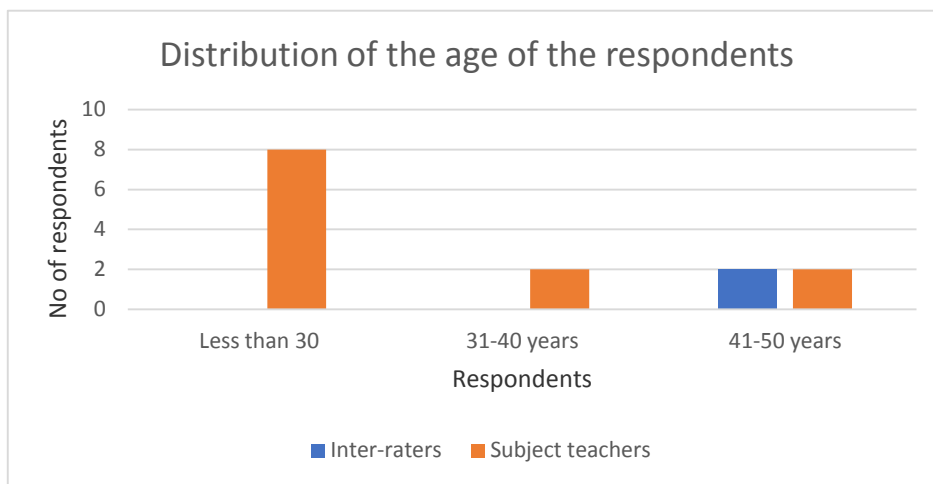


Figure 3 shows that majority of the respondents were aged below 30 years (57.14%), followed by those aged between 41-50 years (28.57%) and those aged between 31-40 years (14.29%) respectively.

Table 4: Summary statistics for inter-raters

<i>Variable</i>	<i>Observations</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Min</i>	<i>Max</i>
Teaching experience in agriculture	2	15.5	2.12132	14	17
Experience examining agriculture project	2	13.5	2.12132	12	15

Table 4 presents the summary statistics for inter-raters. The findings indicate that the average inter-rater experience teaching agriculture projects was 15.5 years ($M=15.5$, $SD=2.12$) with a minimum of 14 and a maximum of 17 years respectively. Similarly, the average inter-rater experience examining agriculture projects was 13.5 years ($M=13.5$, $SD=2.12$) with a minimum and maximum of 12 and 15 years respectively. This was in line with the intention of the researcher that

the inter-raters were to have a long teaching experience which according to research indicates that they are able to assess the projects consistently.

Table 5: Summary statistics for subject teachers

<i>Variable</i>	<i>Observations</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Min</i>	<i>Max</i>
Teaching experience in agriculture	12	5.333333	4.355422	2	14
Experience examining agriculture project	12	4.166667	3.927371	1	12

Table 5 presents the summary statistics for subject teachers. The findings indicate that the average subject teacher experience teaching agriculture projects was 5.3 years ($M=5.3, SD=4.4$) with a minimum of 2 and a maximum of 14 years respectively. Similarly, the average subject teacher experience examining agriculture projects was 3.9 years ($M=3.9, SD=3.9$) with a minimum and maximum of 1 and 12 years respectively.

Table 6: Proportions on whether respondent attended training on assessment of agriculture projects

<i>Attended training on assessment of agriculture projects</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Yes	2(100%)	10(83.33%)
No	0(0%)	2(16.67%)
Total	2	12

Table 6 shows that all inter-raters 2(100%) had attended training on assessment of agriculture projects while 10(83.33%) of the subject teachers had attended training on assessment of agriculture projects compared to 2(16.67%) subject teachers who had not attended any training on assessment of agriculture projects. The 83.33% of teachers who reported to have been trained could be attributed to the extensive nationwide training and workshops that were conducted in 2019.

Figure 4: Whether respondent attended training on assessment of agriculture projects

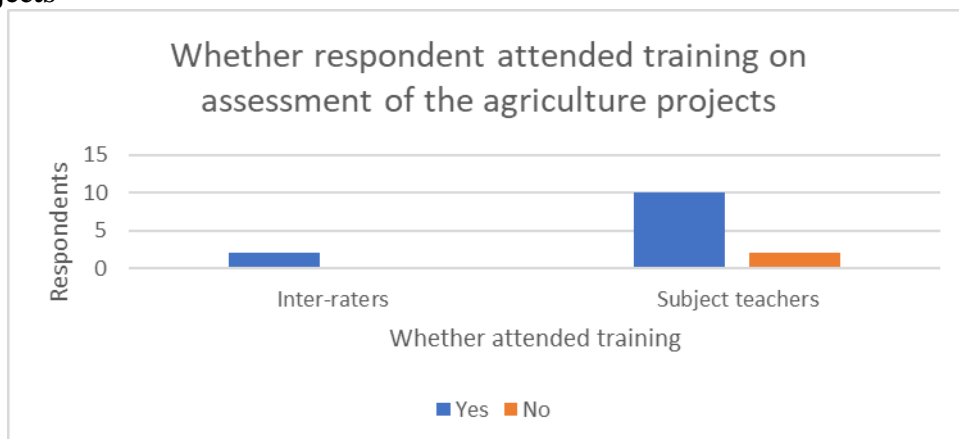


Table 7: Number of students taught

<i>Number of students taught</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Less than 50	1(50%)	11(91.67%)
50	1(50%)	0(0%)
100	0(0%)	1(8.33%)
Total	2	12

Table 7 presents results on the number of students taught. Specifically, 1(50%) of the inter-raters taught less than 50 students while 1(50%) of the inter-raters taught 50 students. Similarly, majority of the subject teachers 11(91.67%) taught less than 50 students compared to 1(8.33%) who taught 100 students. The same findings are indicated in Figure 5.

Agriculture is an elective subject; therefore on transiting to form three students make a choice of the subjects that they would like to pursue further in line with the provisions by the KNEC. This explains why the majority of respondents reported to have classes of less than 50 students.

Figure 5: Number of students taught

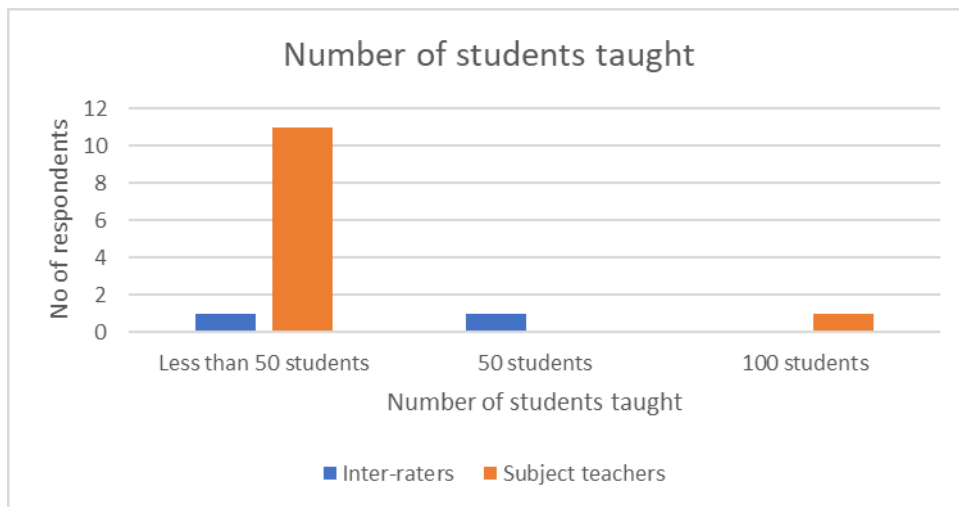


Table 8: Time spent on agriculture projects

<i>Time spent on agriculture projects</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Less than 5 periods	0(0%)	2(16.67%)
6-15 periods	0(0%)	9(75.00%)
17-20 periods	2(100%)	1(8.33%)
Total	2	12

Table 8 presents the results on the time spent on agriculture projects. Particularly, it was found that 2(100%) of the inter-raters spent 17-20 periods on agriculture projects. Similarly, majority of the subject teachers 9(75%) spent 6-15 periods on agriculture projects, while 2(16.67%) of the subject teachers spent less than 5 periods on agriculture projects and 1(8.33%) spent 17-20 periods on agriculture projects. The same findings are indicated in Figure 6. It is important to note that this time span refers to the duration between January, when the project assessment began to mid March. This is as a result of COVID-19 pandemic that forced closure of schools in March.

Figure 6: Time spent on agriculture projects

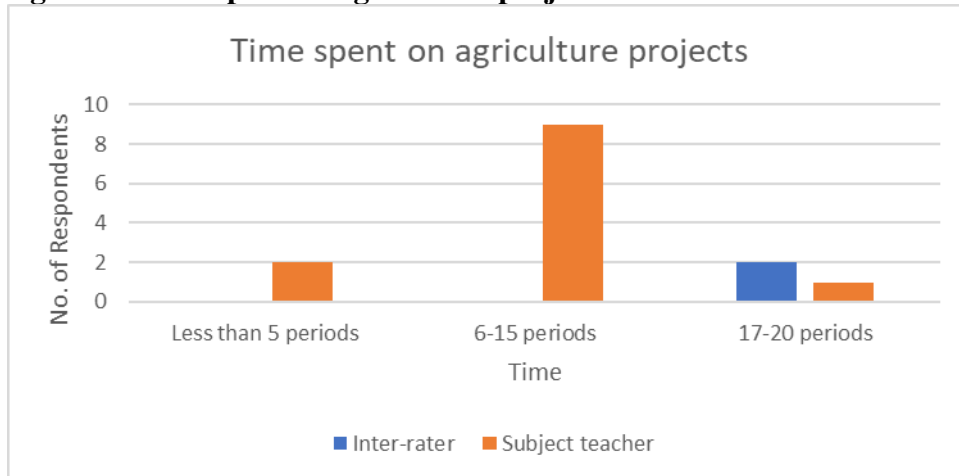


Table 9: Records and pictorial evidences

<i>Records and pictorial evidences</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Yes	2(100%)	12(100%)
No	0(0%)	0(0%)
Total	2	12

Table 9 presents the results on records and pictorial evidences. They indicate that 2(100%) of the inter-raters were keeping records and pictorial evidences compared to 12(100%) of the subject teachers who reported they were keeping records and pictorial evidences. The findings are further indicated in Figure 7.

Figure7: Whether respondents keep records and pictorials

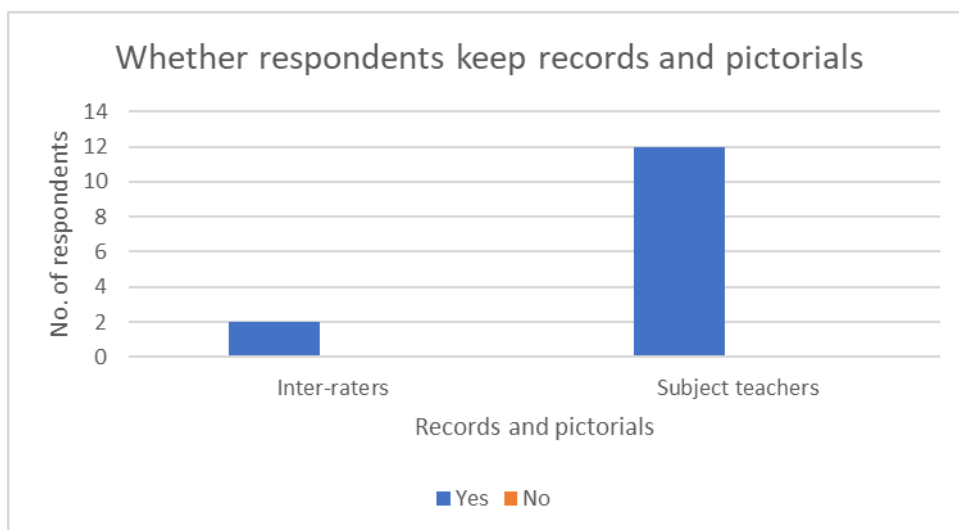


Table 10: Marking scheme provided by KNEC

<i>Marking scheme provided by KNEC</i>	<i>Inter-raters</i>	<i>Subject teachers</i>
Yes	1(50%)	11(91.67%)
No	1(50%)	1(8.33%)
Total	2	12

Table 10 presents results on use of marking scheme provided by KNEC. They indicate that 1(50%) of the inter-raters were using marking schemes provided by KNEC compared to 1(50%) who were not using marking schemes provided by KNEC. Similarly, 11(91.67%) of the subject teachers were using marking schemes provided by KNEC compared to 1(8.33%) who were not using marking schemes provided by KNEC. The same findings are reflected in Figure 8.

Figure 8: Whether respondents use KNEC marking scheme

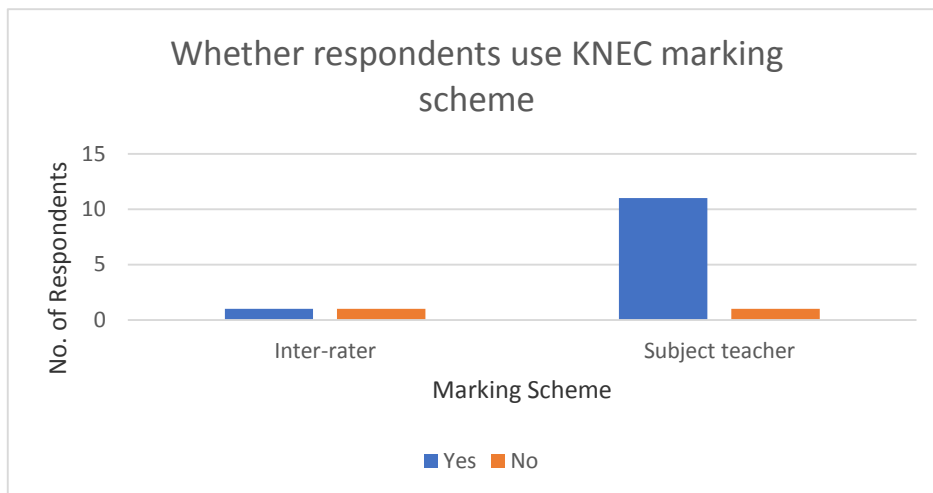


Table 11: Summary statistics on the continuous variables

<i>Variable</i>	<i>Observations</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Min</i>	<i>Max</i>
Subject teacher score	380	42.28158	4.264776	30	49.5
Inter-rater 1	380	37.59474	4.503763	24	47
Inter-rater 2	380	37.75263	4.66581	24	47
Average inter-rater score	380	37.67368	4.541668	24	46
Theory score	380	21.64079	7.080244	6	36

In Table 11, the summary statistics for the continuous variables are presented. Particularly, the average score by the subject teacher was 41.3 ($M=42.3$, $SD=4.3$) with a minimum and maximum of 30 and 49.5 respectively. The average score from

inter-rater 1 was 37.6 ($M=37.6$, $SD=4.5$) with a minimum and maximum of 24 and 47 respectively while the average score from inter-rater 2 was 37.8 ($M=37.8$, $SD=4.7$) with a minimum and a maximum of 24 and 47 respectively. Similarly, the average inter-rater score from the two inter-raters was 37.7 ($M=37.7$, $SD=4.5$) with a minimum and maximum of 24 and 46 respectively. Finally, the average theory score was 21.6 ($M=21.6$, $SD=7.1$) with a minimum and maximum of 6 and 36 respectively.

4.2 Establishing the average of scores of the teacher and inter-rater score and determining their reliability in relation to theory examination

The first objective was to establish the average of the scores of the subject teacher and inter-raters and determine their reliability in relation to theory examination scores in Matungu Sub-county by establishing the test-retest reliability coefficient.

The following section shows analysis on frequencies, means, standard deviations, paired t-test, regression analysis on reliability of average inter-rater scores and theory and average subject teacher and theory examination scores.

As shown in Table 11, the average score by the subject teacher was 42.3 ($M=42.3$, $SD=4.3$) with the minimum and maximum scores being 30 and 49.5 respectively. The average inter-rater score was 37.7 ($M=37.7$, $SD=4.5$) with the minimum and maximum scores being 24 and 46 respectively. The average theory score was 21.6 ($M=21.6$, $SD=7.1$) with minimum and maximum scores being 6 and 36 respectively. The paired t-test was used to determine the reliability of the teacher and inter-rater scores in relation to theory scores.

Table 12: Distribution of the proportions of the independent variable in the dependent variable

<i>Variable</i>	<i>Theory Examination Scores</i>				
	<i>Observations</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Min</i>	<i>Max</i>
Subject teacher scores	380	42.28158	4.264776	30	49.5
Average inter-rater scores	380	37.67368	4.541668	24	46
Inter-rater 1 scores	380	37.59474	4.503763	24	47
Inter-rater 2 scores	380	37.75263	4.66581	24	47

From Table 12, the average theory examination score that were influenced by subject teacher scores was 42.3 ($M=42.3$, $SD=4.3$) with a minimum of 30 and

maximum of 49.5 respectively. The average theory examination score that was influenced by the average inter-rater scores was 37.7 ($M=37.7, SD=4.5$) with a minimum of 14 and maximum of 46 scores respectively. Similarly, the average theory examination score that was influenced by scores from inter-rater 1 was 37.6 ($M=37.6, SD=4.5$) with a minimum and maximum of 24 and 47 respectively. The average theory examination score that was influenced by scores from inter-rater 2 was 37.8 ($M=42.3, SD=4.3$) with a minimum and maximum of 24 and 47 respectively. These findings are also presented graphically in Figure 9.

Figure 9: Distribution of the proportions of the independent variables in the theory examination scores

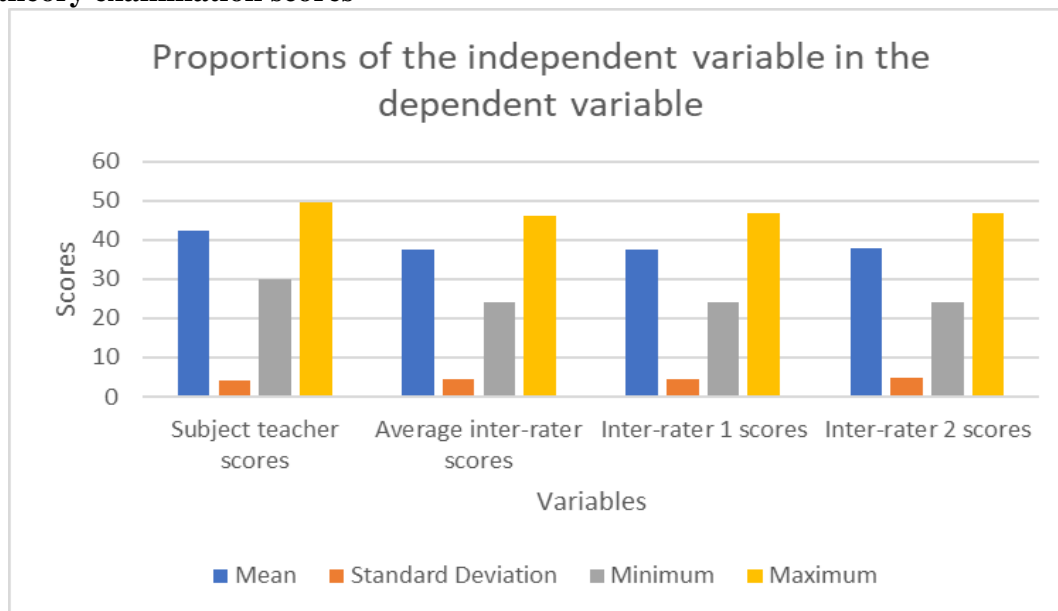


Table 13: Paired t-test on reliability of subject teacher scores and theory scores

<i>Degrees of Freedom (DF)</i>	<i>t-statistic</i>	<i>p-value</i>
379	67.5934	0.0000

Table 13 shows that the difference in the scores assigned by subject teachers and those on theory was statistically significant, implying that the observed difference was not by chance and that teacher and theory scores were thus reliable.

Table 14: Paired t-test on reliability of inter-rater scores and theory scores

<i>Degrees of Freedom (DF)</i>	<i>t-statistic</i>	<i>p-value</i>
379	52.8270	0.0000

Table 14 shows that the difference in the scores assigned by the inter-raters and those on theory was statistically significant. This means that the observed difference was not by chance and that inter-rater and theory scores were reliable.

Table 15 presents the regression results on the effect of subject scores and inter-rater scores on theory examination scores. The findings indicate that scores by subject teacher had a positive and statistically significant effect on theory examination scores. Specifically, it was found that subject teacher scores improved theory performance by 0.4280 points on average ($\beta=.4280$, $t=3.18$, $p=0.002$) while inter-rater scores improved theory performance by .5237 points on average ($\beta=.5237$, $t=4.14$, $p=0.000$).

Table 15: Regression results on effect of subject teacher and inter-rater scores on theory

<i>Theory score</i>	<i>Coefficient</i>	<i>SE β</i>	<i>p-value</i>
Constant	-16.1855 (-5.43)	2.9820	0.000
Subject teacher score	.4280 (3.18)	.1346	0.002
Inter-rater score	.5237 (4.14)	.1264	0.000
<i>Number of observations</i>	380		
<i>R²</i>	0.3271		

4.3 Establishing the inter-rater concordance of the scores generated in agriculture project

The second objective aimed at establishing whether there was agreement in scores generated by the inter-raters in agriculture project in Matungu Sub-county.

Below is an analysis on Fleiss Kappa to establish whether there was concordance between the two inter-raters in the scores generated in agriculture project.

Table 16: Fleiss' Kappa inter-rater concordance

<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>
0.4632	0.0256	18.08	0.000

Since the Fleiss' Kappa concordance coefficient is 0.4004, it indicates that there was moderate agreement between the two inter-rater scores. According to McHugh

(2012), Fleiss' kappa values ≤ 0 indicate no agreement, 0.01-0.20 indicate none to slight agreement, 0.21-0.40 indicate fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 indicate almost perfect agreement.

4.4 Correlating subject teacher scores and scores generated by the inter-raters

The third objective sought to establish the correlation between subject teacher scores and scores generated by the two inter-raters.

The following table shows the results of the analysis of the correlation of subject teacher versus inter-rater scores.

Table 17: Correlating subject teacher scores and scores generated by the inter-raters

<i>Correlation coefficient</i>	<i>Subject teacher scores</i>
Inter-rater 1	0.8386
Inter-rater 2	0.8521

Table 17 presents the correlates the subject teacher scores with those of the inter-raters. The findings indicate that the correlation coefficient between subject teacher scores and inter-rater 1 was 0.8386 while that between subject teacher scores and inter-rater 2 was 0.8521. The coefficients were positive and close to 1 indicating a positive and strong relationship between the subject teacher scores and scores by the two inter-raters.

4.5 Establishing reliability coefficient of the teacher versus the inter-rater

The fourth objective sought to establish the reliability coefficient between subject teacher score and inter-rater score in Matungu Sub-county using the Pearson moment coefficient.

Table 18: Pearson moment coefficient between subject teacher scores and inter-rater scores

<i>Correlation coefficient</i>	<i>Subject teacher scores</i>
Average inter-rater scores	0.8535

Table 18 presents the correlation coefficient between subject teacher scores and those by the inter-raters. The coefficient was close to 1. The findings indicate that the Pearson correlation coefficient was 0.8535 indicating a positive and strong relationship between the two variables.

4.6 Establishing the strength of association between teacher score, inter-rater score, and the theory examination score

The fifth objective aimed at examining the strength of association between subject teacher, inter-rater and theory examination score in agriculture in Matungu Sub-county. The following is a summary of the chi square analysis to establish whether there is association between the variables. The results of Cramer's V to establish the strength of the association are also presented.

Table 19: The strength of association between teacher score and theory examination score

Subject teachers	Theory examination			
	Degrees of Freedom (DF)	Chi-Square Statistic (χ^2)	P-value	Cramer's V (Φ)
	1144	1300*	0.007	0.3577

Note: * means statistically significant at the 5% level of significance

Table 20: The strength of association between average inter-rater score and theory examination score

Average inter-rater score	Theory examination			
	Degrees of Freedom (DF)	Chi-Square Statistic (χ^2)	P-value	Cramer's V (Φ)
	1540	1600	0.285	0.3437

Table 19 and 20 presents the Chi-Square and Cramer's V results on the strength of association between subject teachers and theory examination, and average inter-rater score and theory examination scores. The Chi-Square results indicate that the association between average inter-rater and theory examination was statistically significant as indicated by ($\chi^2=1300$, $p=0.007$, $\Phi=0.3577$) while that between subject teacher scores and theory examination was statistically insignificant ($\chi^2=1540$, $p=0.285$, $\Phi=0.3437$). The strength of the association was determined using Cramer's V which returned strong association as indicated by $\Phi=0.3577$ and $\Phi=0.3437$ respectively.

CHAPTER FIVE
DISCUSSION, CONCLUSION AND RECOMMENDATIONS

5.0 Introduction

The study sought to establish the relationship between the type assessment procedure of agriculture project and reliability of students' score in theory examinations in Matungu Sub-county. This chapter discusses the internal and external validity, demographic factors and main findings in relation to 1. Average of scores of the teacher and inter-rater score and their reliability in relation to theory examination 2. Inter-rater concordance of the scores generated in agriculture project 3. Correlation of subject teacher scores and scores generated by the inter-raters 4. Reliability coefficient of the teacher versus inter-rater scores 5. Strength of association between subject teacher, inter-rater and theory examination scores. Also presented are results on the confounding variable measured, conclusions drawn and finally recommendations.

5.1 Internal and External Validity

The study aimed at examining the relationship between the type of assessment procedure (either by subject teacher or inter-rater) and reliability of students' scores in agriculture theory examination. The students' project work was assigned two sets of scores, one by the subject teacher and another average score of the two inter-raters. The third set of scores was retrieved from the joint sub-county agriculture theory mock examinations that were sat by the students. Bearing in mind that the study required teachers to give the scores, teachers were apprehensive at first citing confidentiality of students' score. The researcher took time to explain to the teachers that the study was legal and authorized by presenting the research permit and also the permission letter from the Sub-county office. This won the trust of teachers and they were reassured that the scores given were to be used strictly for academic purposes only and that the scores will be strictly treated with the confidentiality that they deserve. Considering the correlational intend to which the scores were to be subjected, the researcher envisaged a situation where with this knowledge the participants would want to conform. This is referred to as the Hawthorne effect understood as the difference in feeling/conduct by contestants which may arise simply as result of taking part in the study (Drew, Hardman and Hosp, 2008). The researcher controlled for Hawthorne effect by concealing the correlational intend of

the study. This therefore checked for intentional conformity of teachers and inter-rater rating in the project scores.

All the sampled schools chose option A for this year's agriculture project which involved establishment of a tree nursery and also sat for the same joint sub-county mock examination. This limits generalization of the finding nationally since schools in other areas might have chosen option B of the project which involved goat rearing and also different sub-counties sit different theory examinations depending on their area, hence the findings can only be generalized to Kakamega county bearing in mind that choice of project is based on ecological conditions the county having homogenous ecological conditions makes it possible for all schools to have gone by the same choice of project.

5.2 Summary of Findings

The following is a brief on the main findings derived from the study to establish whether there is a relationship between type of assessment procedure of agriculture project (assessment by subject teacher or inter-raters) and reliability of students' score in agriculture theory examinations in Matungu Sub-county.

- 1) The results indicate that the reliability coefficient between subject teacher scores and theory scores was .4280 while that of inter-rater score and theory examination score was .5237. The reliability coefficients of both teacher and inter-rater are between 0-1 meaning a positive correlation. The paired t-test results on reliability of subject teacher scores was $t=3$ while the t-test on reliability of inter-rater scores and theory scores was $t=4$.
- 2) Inter-rater concordance of the scores generated by the inter-raters was .4632 indicating a moderate agreement between the inter-raters. This implies that the inter-raters agreed moderately by awarding a similar score to the same students' project work.
- 3) The correlation coefficient between subject teacher scores and scores generated by inter-rater 1 and 2 was 0.8386 and 0.8521 respectively indicating a strong positive correlation.
- 4) The Pearson moment coefficient between the subject teacher score and inter-rater scores was 0.8. This is a strong positive correlation of the variables.

- 5) From the results, the study revealed a statistically significant chi-square association between subject teacher, inter-rater and theory examination scores as indicated by ($\chi^2=1600$, $p=0.000$, $\Phi=0.4003$) and ($\chi^2=1144$, $p=0.007$, $\Phi=0.3577$). The strength of association between inter-rater score and theory score was stronger as compared to that between subject teacher and theory score according to Cramer's V that statistically showed the strength of association.

5.3 Discussion of the Findings

The following is a discussion on the results of this study regarding the relationship between the type of assessment procedure of agriculture project and reliability of students score in agriculture theory examinations in Matungu Sub-county. According to Meadows and Bellington (2005), a test is useful if there is a correlation between the score on the test (predictor) and scores on whatever one is trying to predict (the criterion). In this case the scores of a student in agriculture project (predictor) should be predictive of the same students' performance in the theory examinations.

5.3.1 Establishing the average of scores of the teacher and inter-rater score and determining their reliability in relation to theory examination

The results showed a statistically significant relationship between subject teacher and theory scores as well as inter-rater versus theory examination scores. This implies that the observed difference is not by chance confirming reliability of the scores. The average score by the subject teacher was 42.3 ($M=42.3$, $SD=4.3$) with the minimum and maximum scores being 30 and 49.5 respectively. The average inter-rater score was 37.7 ($M=37.7$, $SD=4.5$) with the minimum and maximum scores being 24 and 46 respectively. The average theory score was 21.6 ($M=21.6$, $SD=7.1$) with minimum and maximum scores being 6 and 36 respectively. The paired t-test shows that the difference in the scores assigned by subject teachers and those on theory was statistically significant, implying that the observed difference was not by chance and that teacher and theory scores were thus reliable. According to Magno (2009), each individual examinee has a true score (unobservable) that would be obtained if there were no measurement errors. In the event that measurement error is eliminated, scoring of students' project work is based on the

provided criteria and therefore other factors unrelated to the test do not affect the rating yielding reliable scores. Findings on regression analysis on effect of subject teacher scores and inter-rater scores had a positive and statistically significant effect on theory examinations.

It is crucial to note that though according to literature, SBA scores by subject teacher have always been viewed to be inconsistent and prone to teacher biases, the reliability of scores in this study have greatly improved which can be attributed to the extensive nationwide training and workshops that were conducted to equip teachers with the knowledge, skills and attitudes relevant in order to improve reliability of such scores. This minimized the error factor where initially teacher scores were assigned haphazardly basing on other factors other than the assessment criteria greatly contributing to inconsistencies and a bid to concentrate scores around a certain value that failed to mimic normal distribution. As reported by the respondents, a high percentage under-went training in assessment of project work.

5.3.2 Establishing the inter-rater concordance of the scores generated in agriculture project

The results indicated Fleiss Kappa concordance that was moderate between the two inter-raters. According to McHugh (2012), Fleiss' kappa values ≤ 0 indicate no agreement, 0.01-0.20 indicate none to slight agreement, 0.21-0.40 indicate fair agreement, 0.41-0.60 moderate agreement, 0.61-0.80 substantial agreement, and 0.81-1.00 indicate almost perfect agreement. Ayodele (1989) notes that inter-rater reliability was used to determine the scale to which different examiners provided rational measure of the same phenomena. According to Rudner and Boston (1994), use of more than one scorer increases reliability in a similar way that multiple test increase reliability of standardized tests. Considering that people are well known for deviations, we must ponder on how we can regulate reliability in observations of two examiners. It is crucial to determine rater reliability. As earlier noted, use of more than one rater increases reliability since each marker's mistakes incline towards compensating for others mistakes (Thorndike & Thorndike-Christ, 2010). According to Thorndike and Thondike-Christ (2008), biasness in scoring of performance-based assessment can be notably decreased by employing more than one rater (Airasian, 2005; Airasian & Russell, 2008; Thorndike & Thorndike-Christ, 2010). He further claims that when a single rater is used, it is noted that one rater

may award consistently higher scores while other awards lower scores for the same PA. The established inter-rater concordance supports that view that use of multiple raters has the potential of reducing inconsistencies in scoring of PA and can be used to resolve the element of error introduced by subjectivity of a single rater.

5.3.3 Correlating subject teacher scores and scores generated by the inter-raters

There was correlation between the scores assigned by inter-rater 1 and subject teacher scores as well as the scores assigned by inter-rater 2 and subject teacher scores. The analysis was to establish whether there were any differences in the way two raters scored the project in relation to the subject teacher scores. The correlation coefficient for inter-rater 1 and 2 was .8386 and .8521 respectively.

5.3.4 Establishing reliability coefficient of the teacher versus the inter-rater

The findings indicate that the Pearson correlation coefficient was 0.8535 indicating a positive and strong relationship between the two variables. This was generated from the average score of the two inter-raters. It could be inferred that there is a strong positive correlation between the subject teacher scores and inter-rater scores. Daniels and Schouten (1970) and Owoyemi, (2000) note that a prediction on the subsequent exam could be shaped with fair success on grounds of the results of previous examinations. These findings relate to those of Andala et al. (2014) study which showed strong positive linear correlation between the mock and KCSE examination. Both the Pearson correlation and the spearman's rho correlation gave a high positive correlation of .949 and .942 respectively significant at 0.01%. As a result of the high positive correlation, the study concluded that mock examinations were reliable.

This means that agriculture subject teachers are well equipped and are conducting assessment based on the laid down procedure. Findings in the various studies indicate that, the assessment procedure approves or taints the evaluation process, the end results endowed to students and its consistency. How do the scores assigned by the subject teacher in the project work compare with those of standardized test score? The most direct comparison between performance in agriculture project and achievement as shown here was to compare student score in the project to

achievement in theory examination which are standardized. Ideally there would be a strong correlation between students score in the theory examination and in the project work. As highlighted earlier, recent studies have shown that scores in agriculture project lack consistency with the same student scores on standardized tests in this case the agriculture theory examinations paper one and two.

5.3.5 Establishing the Strength of association between subject teacher, inter-rater and theory examination score.

The Chi-Square results indicate that the association between average inter-rater and theory examination was statistically significant while that between subject teacher scores and theory examination was statistically insignificant. The strength of the association was determined using Cramer's V which returned strong association leading to rejection of the null hypothesis that there was no association between the variable.

It is evident that when the random error is reduced to a minimal level, score preciseness and reproducibly can be generalized to supplementary evaluation tests and related tests. Reliability, therefore, is viewed as how dependable a measure is. Like in this case the presence of associations between the variables implies that the scores are reliable. It can be deduced that it is the estimate of scores accuracy and dependability. In other words, the degree to which a score measures the behaviour being assessed rather than other factors that cause score variation. In this case, the scores are thought to be reliable if we would get duplicate results if the test were done on different occurrences. This implies that it should be possible then for student scores in agriculture project to be correlated positively to the theory agriculture examinations. The existence of an association between the three variables imply that the scores are reliable and free from errors caused by other factors other than the test itself.

5.4 Conclusion

Following the study findings, a number of conclusions were drawn in relation to reliability of agriculture project scores in Matungu Sub-county.

First, it was established that there was a significant correlation between the scores generated by inter-raters in the project work and theory examination scores as

compared to the correlation coefficient between subject teacher scores and theory examination scores.

Secondly, inter-rater concordance was seen to be moderate implying the inter-raters moderately agreed. Meaning, the two inter-raters tended to agree on a number of occasions by awarding same score to the work of same student. As already discussed, the inter-raters are independent and the students are unknown to them. Therefore, their scoring of students work is objectively done in accordance to the assessment criteria minimizing the error element as no factors other than the assessment criteria affected their scoring.

Thirdly, there was a strong positive correlation between the subject teacher and inter-rater scores. As earlier mentioned, extensive training and workshops on group 4 subjects with a project component had been conducted country wide. This is evident from the research findings which sought to establish whether the respondents had undergone training in assessment of agriculture project where 85.71% indicated that they had undergone training on assessment of agriculture project. It is assumed that this armed the teachers with the required skills, knowledge and attitude in such assessment generating reliable scores. Broadfoot (1994) argued that teacher training to obtain suitable skills is crucial in assessment. The improved assessment practices hence reliable scores can also be attributed to clear marking scheme provided by KNEC. 85.71% of the respondents indicated that the marking scheme was clear.

Lastly, the results also indicate statistically significant association between subject teacher, inter-rater and theory examination scores.

5.5 Recommendations.

1. The main aim of the study was to establish the relationship between assessment procedure of agriculture project and reliability of student scores in agriculture theory examinations. Performance assessments equip the learner with practical skills. Such assessments should be given priority. It should be a clear reflection of students' ability. When the assessment is objectively carried out according to the correct assessment procedure, there should be a strong correlation between the students' ability as seen in the

project work and their ability in theory examinations. Agriculture projects should be given valued consideration by teachers and the entire school administration in order to uphold the quality of assessment methods and to ensure assessments are carried out transparently because as noted, project assessment predicts future academic performance like in this case KCSE this will ensure reliable scores.

2. There is need to continuously empower teachers charged with assessment of agriculture project with knowledge, skills, attitudes and competencies that are necessary to ensure objectivity in assessment so as to realize project work scores that can be correlated to the same students score in theory examinations which are standard.
3. The body charged with assessment and the quality control department can explore the possibility of incorporating the concept of multiple scoring of agriculture projects to eliminate biases introduced by the subject teachers. Literature suggests this as a remedy to subject teacher biases and way of improving reliability of project scores.
4. In PA it is important to emphasize proper application of marking schemes which clearly reflect the performance criteria and levels of performance. This offers guidance to scorers to ensure scores are not just randomly assigned but are derived in accordance with the specific performance criteria.
5. The worldwide COVID-19 pandemic had far reaching effects on the area of project work assessment. In this case, although agriculture project is to be assessed in two milestones, assessment was only done once. The country has never experienced a pandemic that caused a social crisis and subsequent lockdown there is limited knowledge about how to deal with such a situation. By the time the students get to form three, they already have chosen agriculture as one of the examinable subjects. It would be important for the practical aspect of agriculture to be a continuous process where students are assessed continuously on various agricultural practices the moment they step in form three unlike the current system where project is reserved till the final year.

5.6 Recommendations for Further Research

1. Considering that there are other subjects with a project component, research should be carried out across other disciplines to establish the relationship between type of assessment procedure and reliability of students score in theory examinations.
2. A study can also be carried out to incorporate the students to seek the understanding of students' views on reliability of their scores in the project assessment.
3. Educational reforms are at an advanced stage in the county, CBC emphasizes SBA that is assessment of learning. It is important for studies to be carried out to establish reliability of other performance assessments in line with the CBC.

REFERENCES

- Abiri, J.O.O. (2006). Elements of evaluation measurement and statistical Techniques in education. Ilorin: Library Publications Committee, University of Ilorin.
- Adeyemi T.O. (2008). Predicting students' performance in Secondary Certificate *Administrators*. USA: Sage Publications.
- Albano (2016) *Inter-rater reliability*. Chapter 6.
Cehsol.unl.edu/aalbano/Intromeasurement/mainch.T.html
- Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Company.
- Allen, M.J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press
- Andala H.O, Digolo, O. & Kamande, M. (2014). *Reliability of Mock Examinations for Prediction of the Kenya Certificate of Secondary Examination (KCSE) Results*.
- Armes, J (2016) Quantifying the Qualitative: Increasing the Reliability of Subjective Language Assessments Master's Projects and Capstones. 338. <https://repository.usfca.edu/>
- Ayodele, S.O. (1985). Assessment practices and the english teacher. In J.D. Medugu An investigation of the assessment practices of industry technology teachers in Adamawa State. An unpublished M.Ed. Thesis. Abubakar Tafawa Balewa University, Bauchi
- Beck AT, Steer RA, Carbin MG. Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clin Psychol Rev. 1988;8(1):77–100. [[Google Scholar](#)]
- Bello, A& Tijani (2003) A. *Training Needs of Teachers in School-Based Assessment in Anglophone West African Countries*. West African Examination Council
- Black, P. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*. 5(1): 7-74.
- Broadfoot, P., (1994). Approaches to quality assurance and quality control in six Countries.
- Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, 52, 1, 29–46.

- Brown, Richards. Coughlin, Ed (2007). The predictive validity of selected. Benchmark Assessment. *Regional Educational Laboratory Mid-Atlantic. Cambridge: University of Cambridge School of Education.*
- Bull, K.S and Kimball, S.L. (2000). Basic measurement theory, objectives and needs Assessment. Oklahoma: Oklahoma State University
- Calvo-Mora, A., Leal, A., & Roldan, J. L. (2006). Using enablers of the EFQM model to manage institutions of higher education. *Quality Assurance in Education*. 14 (2) 99-122.
- changed assessment practices, and what promise does the revised National
- Chong, S., & Ho, P. (2009). Quality teaching and learning: A quality assurance framework for 559 initial teacher preparation programs. *International Journal of Management in Education*, 3, 560-314.
- CLASSICAL_TEST_ANALYSIS.pdf. Retrieved on 13 August, 2014.
- Cohen, L., Manion, L. & Morris, K. (2000). *Research Methods in Education. 5th edition*. London: Routledge Falmer. Congressional Research Service.
- Cronbach, L. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass. Curriculum Statement hold? *Perspectives in Education*, 21 (1): 119 -133.
- DeVillis, R.E., "Scale Development: Theory and Application," *Applied Social Science Research Method Series*, Vol. 26 (Newbury Park: Sage Publishers Inc., 2006).
- Education: Principles, Policy and Practice, Carefax, Oxfordshire.
- Elbert, C. D., & Baggett, C.D. (2003). Teacher competence for working with disabled students as perceived by secondary level agricultural instructors in Pennsylvania. *Journal of Agricultural Education*, 44(1), 105-114.
- Examination from performance in Junior Secondary Certificate Examinations in Ondo State, Nigeria. *Humanity and Social Sciences Journal* 3 (1); 26-36.
- Fisher, R and Lewis, M (2002) Examining teaching in the literacy hour. In Fisher, R, Brooks, G and Lewis, M (eds) *Raising standards in literacy*. London, Roulledge.
- Ganesh, T., "Reliability and Validity Issues in Research," *Integration and Dissemination Research Bulletin* 4, (2009): 35-40.
- Goddard-Spear, M. (1984) The biasing influence of pupil sex in a science marking exercise. *Research in Science & Technological Education*, v2 n1 p55-60
- Graziano, A.M. & Raulin, M.L., *Research Methods: A Process of Inquiry*, 6th Ed. (Boston, MA: Allyn & Bacon, 2006).

- Greatorex J (2005). Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. *J of Vocational Education & Training* 57(2), 149-164;
- Greatorex, J. & Bell, J.F. (2002a) Does the gender of examiners influence their marking? Paper presented at the Learning communities and assessment cultures: Connecting research with practice, University of Northumbria.
- Greatorex, J., Baird, J., & Bell, J.F. (2002) 'Tools for the trade': What makes GCSE marking reliable? Paper presented at the EARLI Special Interest Group on Assessment & Evaluation, University of Northumbria, UK, August 2002.
- Greenberg, K (1992). *Validity and reliability issues in the direct assessment of writing*. WPA writing programme Administration. Volume, 16 No. 1-2.
- Griffith, S. A. (2005). *Assuring fairness in school based assessment: Mapping the boundaries of teachers' involvement*. A paper presented at the 31st Annual Conference of the International Association for Educational Assessment. Abuja,
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.
- Haradhan, M. (2017) Two criteria for good measurements in research: Validity and Reliability. Munich Personal RepEc Archives. Premier University Chittagong, Bangladesh.
- Harlen, W. (1998). Classroom assessment: A dimension of purposes and procedures. In K. Carr (Ed.), *SAME papers* (pp. 75-87). Hamilton, New Zealand: Centre for Science, Mathematics, and Technology Educational Research
- Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for Centre, social science research unit, Institute of Education, University of London. Retrieved 6th May, 2009, from: <http://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=6WO5QivP0Q4%3d&tabid=119&language=en-US>
- Harlen, W. (2005). 'Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes', *Research Papers in Education*, 20, 3, 245-270
- Hoxby, Robertson, Symons, S. Chee, P (2003)., Factors that influence Bruneian students not to enroll in secondary school Agriculture subject. Darusalam: Brunei
- <https://newsblaze.co.ke/knec-posts-new-assessment-sheets-to-be-used-for-evaluating-project-based-examinations-this-year/>
- Hulela, K. (2017) The practice of scaling down practical assessment components of agriculture in Junior Secondary Schools curriculum. A synthesis of teacher

perception. *Journal of Education and Training Studies*. Vol 5, No. 36: March 2017

- John, A. C. (2015) Reliability and validity: Asine Qua Non for fair assessment of undergraduate technical and vocational education projects in Nigerian Universities. *Journal of Education and Practice*. Vol. 6, No. 34, 2015
- Johnson, S. (2011). A Focus on Teacher Assessment Reliability in GCSE and GCE. In: Ofqual Reliability Compendium (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].
- Jones, B. E. (2001) *Checking the checkers – a report on a script re-checking exercise undertaken in the Manchester office of the AQA, Summer 2001*. AQA Research Report, RC/154.
- Kenya Literature Bureau (1992). *Secondary School Syllabus Vol.2, Nairobi*, Author
- Kenya National Examinations Council. (2018). *Kenya Certificate of Secondary Education Examination Report (2013) Vol.2, Nairobi*, Author
- Keyton, J., King, T., Mabachi, N.M., Manning, J., Leonard, L.L. & Schill, D., Content Analysis Procedure Book (Lawrence, KS: University of Kansas, 2004)
- Kibett, J.K. (2002). *Effect of project based learning on student performance in secondary school agriculture* (unpublished Ph.D Thesis). Njoro, Kenya; Egerton University
- Kim, S.C. and Wilson, M. (2009). ‘A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model’, *Journal of Applied Measurement*, **10**, 4, 408–423.
- Kinyua & Okunya (2014). *Validity and reliability of teacher-made tests: Case study of 74 year 11 physics in Nyahururu District of Kenya*. African Educational Research Journal 2(2),
- Klenowski, V., & Wyatt-Smith, C. (2008). “Standards-driven reform Years 1–10: Kline, T. J (2005) *Psychological testing a practical approach to design and evaluation*. Sage publication, UK.
- Liao, SC, Hunt, EA & Chen, W (2010). *Comparison between inter-rater reliability and inter-rater agreement in performance assessment*. August 2010, Volume 39 No. 8.
- Livingston, S. A (2018) Test reliability. Basic concepts. Education Testing Services. Research Memorandum. ETS- RM-18-01

- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley Publishing Company
- Magnusson, D. (1967) *Test theory*. Reading. MA: Addison Wesley.
- Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson and S. L. Hershberger (Eds.), *the new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum, pp. 129-152
- Masole, M. T. (2011) *Enhancing the quality of performance assessment in agriculture in Botswana schools*. Phd Dissertation. University of Pretoria.
- Masters, G. N., & McBryde, B. (1994). *An Investigation of the Comparability of Teachers' Assessments of Student Folios*. Tertiary Entrance Procedures Authority. Brisbane.
- Maxwell, G. S. (2004). *Progressive assessment for learning and certification: Some lessons from School-based assessment in Queensland*. A paper presented at the third Conference of the Association of Commonwealth Examination and Association Boards. Nadi, Fiji.
- McHugh, M. (2012). Inter-rater reliability: the kappa statistic, *Biochemia Medica*, 22(3), 276-282.
- McMillan, J.H. (2001). *Essential Assessment Concepts for Teachers and*
- Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].
- Meadows, M. and Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection*. London: QCA [online]. Available: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA_104980_marker_selection.pdf [12 November, 2012].
- Moderation an optional extra?", A paper presented at the Australian Association for Research in Education Conference, Brisbane.
- Murphy, R.J. (1978). 'Reliability of marking in eight GCE examinations', *British Journal of Educational Psychology*, 48, 2, 196–200. Cited in:
- Mwanyumba, D and Mutwiri, J.G (2009), *Challenges associated with implementation of control mechanism in Public examinations and how the Kenya National Examinations Council (KNEC) has handled some of these challenges*. A paper presented at the 27th Annual Conference of the Association for Educational Assessment in Africa (AEAA) Yaounde, Nairobi .The Kenya National Examinations Council.

- Newstead, S. E. & Dennis, I. (1994) Examiners examined: The reliability of exam marking in psychology. *The Psychologist: Bulletin of the British Psychological Society*, v7 p216-219
- Nitko J. A. (1994). A framework for concepts and procedures of using continuous
- Njabili, A. F., Abedi, S., Magesse, N. W., & Kalole, A. M. (2005). *Equity and school-based assessment: The case of Tanzania*. Paper presented at the 31st Annual Conference of International Association for Educational Assessment (IAEA). Abuja, Nigeria
- Novick, M.R. (1966) *The axioms and principal results of classical test theory* *Journal of Mathematical Psychology* Volume 3, Issue 1, February 1966, Pages 1-18
- Nyang'au, M. K, Kibett, J. K & Ngesa F.U, (2011) *Perception of school principals and agriculture teachers towards factors influencing initiation of secondary school agriculture projects*. *Middle-East Journal of Scientific research* 9(4)
- Oberholzer, A. (1998). *Research Project: School-Based Assessment 1998. Draft Document for Discussion. Unpublished article*. KwaZulu-Natal.
- Othuon, L. and N. Kishor, 1994. Hierarchical linear modelling of predictive validity: The case of Kenya certificate of primary education examination. *Studies in*
- Owoyemi, N., 2000. Moderation and standardisation of continuous and terminal
- Phillips, F. K, (2007). *School based assessment: The need, the reality, the future A perspective from the Independent Examinations Board of South Africa* *Independent Examination Board, South Africa*
- Popham, W.J. (2001). Why Standardized Test Scores Don't Measure Educational Quality. *Educational Leadership*, 56(6) 8-15.
- Porter, J. M. & Jelinek, D. (2011). Evaluating Inter-rater Reliability of a National Assessment Model for Teacher Performance, *International Journal of Educational Policies*, 5(2), 74-87.
presented at the University of South Africa. Pretoria, October, 1996.
problems of assessment and prediction of academic performance, Council for cultural cooperation of the council of Europe; George Harrap Co. Ltd, Publishing Ltd.
- Queensland Studies Authority. (2009). Student assessment regimes: getting the balance right for Australia. Queensland Government *resources information*.
- Royal-Dawson, L. (2004) Is teaching experience a necessary condition for markers of Key Stage 3 English? AQA Research Report, RC261.

- Sadler R (1989) Formative assessment and the design of instructional systems. *Instructional Science* 18: 119-144.
- Satterley, D. (2000). The quality of external assessment. In W. Harlen (Ed.) *Enhancing quality in assessment*. <file:///D:/quality.of.external.assessments.htm>
- Schumacker, R. E. (2010). *Classical Test Analysis*.
<http://appliedmeasurementassociates.com/ama/assets/File/>
- Scottish Examination Board (1992) *Investigation into the effects of the characteristics of candidates and presenting centres on possible marker bias*. Edinburgh: Scottish Examination Board
- Scottish Qualifications Authority Spiller,D (2009) *Principles of Assessment* . University of Waikato,
- Shohamy, E., Gordon, C., & Kramer, R. (1992) The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, v76 n1 p27-33.
- Sieborger, R., & Macintosh, H. (1998). *Transforming Assessment: A guide for South African Teachers*. Cape Town: Juta.
- Spear, M. (1984) The influence of halo effects upon teachers' assessments of written work. *Research in Education*, v56 p85-87.
- Stanley, G., & Tognolini, J. (2008). "Performance with respect to standards in public examinations". A paper presented at the Annual Conference of the International Association for Educational Assessment. Cambridge
- Stemler, S.E. (2004). 'A comparison of consensus, consistency and measurement approaches to estimating interrater reliability', *Practical Assessment, Research & Evaluation*, 9, 4 [online]. Available: <http://PAREonline.net/getvn.asp?v=9&n=4> [12 November, 2012].
- Stiggins, R. J. (1997). *Student-Centred Classroom Assessment* (2nd ed.). New Jersey: Prentice-Hall, Inc.
- Suto, W.M.I. and Nadas, R. (2008). 'What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers', *Research Papers in Education*, **23**, 4, 477-497. Cited in: Suto, I., Nadas, R. and Bell, J. (2011). '.
- Tisi J., Whitehouse, G., Maughan S. and Burdett, N. (2013). *A Review of Literature on Marking Reliability Research* (Report for Ofqual). Slough: NFER
- Traub, R.E., & Fisher, C.W. (1997). *On The Equivalence of Constructed Responses and MultipleChoice*

- Tuckman, B.W. (1985). *Measuring Educational Outcomes*. Honduras: Brace Jovanovich International.
- WAEC (1993). *A research report on "Ghana basic education certificate examination (BECE)": Relationship between Internal and external assessments*. WAEC publications. Accra department of WAEC research division.
- Weigle, S. (1999) Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative & Qualitative Approaches. *Assessing Writing*, v6 n2 p145-178.
- Wells, C. S. and Wollack, J. A (2003) *An instructor's guide to understanding test reliability*. Testing & Evaluation Services. University of Wisconsin. W. Johnson St. 373 Madison.
- Wikimedia Foundation (2006). Qualification and assessment. Retrieved on August, 25, 2006 from http://en.wikibooks.org/wiki/SA_NCS:qualification_assessment
- Wild, C. L., & Ramaswamy, R. (2008). *Improving Testing: Applying Process Tools and Techniques to Quality*. New York: Lawrence Erlbaum
- William, D. (2000) Reliability, validity, and all that jazz. *Education*, v29 n3 p9-13.
- Williamson, C (2017) *Teachers' Role in School-Based Assessment as Part of Public Examinations*. US-China Education Review B, June 2017, Vol. 7, No. 6, 301-307 doi: 10.17265/2161-6248/2017.06.005 University of West Indies (UWI), Mona, Jamaica
- Wood, R. (1991) *Assessment and Testing: A survey of research*. Cambridge: Cambridge University Press
- Yarnold, P.R., "How to Assess the Inter-Method (Parallel-Forms) Reliability of Ratings Made on Ordinal Scales: Emergency Severity Index (Version 3) and Canadian Triage Acuity Scale," *Optimal Data Analysis*, 3(4), (2014): 50-54.
- Zhao, Y. (2009). *Catching up or leading the way: American education in the age of globalization*. Alexandria, VA: ASCD.

APPENDICES

Appendix I: Sample Size Form

Required Sample Size[†]
from: **The Research Advisors**

Confidence = **95.0%** Confidence = **99.0%**

Population Size	Confidence = 95.0%				Confidence = 99.0%			
	Degree of Accuracy/Margin of Error				Degree of Accuracy/Margin of Error			
	0.05	0.035	0.025	0.01	0.05	0.035	0.025	0.01
10	10	10	10	10	10	10	10	10
20	19	20	20	20	19	20	20	20
30	28	29	29	30	29	29	30	30
50	44	47	48	50	47	48	49	50
75	63	69	72	74	67	71	73	75
100	80	89	94	99	87	93	96	99
150	108	126	137	148	122	135	142	149
200	132	160	177	196	154	174	186	198
250	152	190	215	244	182	211	229	246
300	169	217	251	291	207	246	270	295
400	196	265	318	384	250	309	348	391
500	217	306	377	475	285	365	421	485
600	234	340	432	565	315	416	490	579
700	248	370	481	653	341	462	554	672
800	260	396	526	739	363	503	615	763
973	276	434	596	884	395	566	712	919
1,000	278	440	606	906	399	575	727	943
1,200	291	474	674	1067	427	636	827	1119
1,829	318	549	835	1537	487	778	1083	1647
2,000	322	563	869	1655	498	808	1141	1785
2,500	333	597	952	1984	524	879	1288	2173
3,500	346	641	1068	2565	558	977	1510	2890
5,000	357	678	1176	3288	586	1066	1734	3842
7,500	365	710	1275	4211	610	1147	1960	5165
10,000	370	727	1332	4899	622	1193	2098	6239
25,000	378	760	1448	6939	646	1285	2399	9972
50,000	381	772	1491	8056	655	1318	2520	12455
75,000	382	776	1506	8514	658	1330	2563	13583
100,000	383	778	1513	8762	659	1336	2585	14227
250,000	384	782	1527	9248	662	1347	2626	15555
500,000	384	783	1532	9423	663	1350	2640	16055
1,000,000	384	783	1534	9512	663	1352	2647	16317
2,500,000	384	784	1536	9567	663	1353	2651	16478
10,000,000	384	784	1536	9594	663	1354	2653	16560
100,000,000	384	784	1537	9603	663	1354	2654	16584
264,000,000	384	784	1537	9603	663	1354	2654	16586

† Copyright, The Research Advisors (2006). All rights reserved.

The recommended sample size for a given population size, level of confidence, and margin of error appears in the body of the table.

For example, the recommended sample size for a population of 1,000, a confidence level of 99%, and a margin of error (degree of accuracy) of 3.5% would be 575.

Change these values to select different levels of confidence.

Change these values to select different maximum margins of error.

Change these values to select different (e.g., more precise) population sizes.

Appendix II :Score Sheet for Collection Of Student's Scores

Please use this form to enter the students' score as indicated. (You are allowed to use more forms if the candidature exceeds the provision).

S/N	INDEX NO.	MARK IN AGRIC PROJECT (443/3) MILESTONE I	INTER-RATER SCORE 1	INTER-RATER SCORE 2	AVERAGE INTER-RATER SCORE	THEORY EXAM SCORE.
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.						

Appendix III :Agriculture Teacher Questionnaire

Dear teacher, thank you for agreeing to participate in this study. Please fill in the blank spaces and tick (✓) the appropriate boxes.

SECTION A: Part 1: Biographical Information

1. Name: _____ 2. 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. Teacher experience _____ years.
4. Teaching experience in agriculture _____ years.
5. Experience in examining agriculture project _____ years
6. Have you ever attended training on the assessment of agriculture project?
Yes No

Part 2: Information regarding teaching of agriculture in the year 2020

Please answer the following questions with respect to the year 2020 teaching of agriculture.

7. How many students do you teach (in agriculture?)
- Less than 50 50 100 150 200
8. How much time was spent on agriculture project work?
- Never (Less than 5 periods)
- Sometimes (6-15 periods)
- Often (17-20 periods)
- Regularly (more than 20 periods)
10. Have records and pictorial evidences been maintained for students' performance on the agriculture project?
Yes No
11. Do you think the marking scheme provided by KNEC for assessing agriculture project is clear? Yes No

Appendix IV :Inter-Rater Questionnaire

Dear teacher, thank you for agreeing to participate in this study. Please fill in the blank spaces and tick (✓) the appropriate boxes.

SECTION A: Part 1: Biographical Information


1. Name: _____ 2. 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. Teacher experience _____ years.
4. Teaching experience in agriculture _____ years.
5. Experience in examining agriculture project _____ years
6. Have you ever attended training on the assessment of agriculture project?
Yes No

Part 2

Please answer the following questions with respect to the year 2020 teaching of agriculture.

7. How many students do you teach? (in agriculture)
Less than 50 50 100 150 200
8. How much time was spent on agriculture project work?
- Never (Less than 5 periods)
- Sometimes (6-15 periods)
- Often (17-20 periods)
- Regularly (more than 20 periods)
10. Have records and pictorial evidences been maintained for students' performance on the agriculture project?
Yes No
11. Do you think the marking scheme provided by KNEC for assessing agriculture project is clear? Yes No.

Appendix V: Agriculture Project 2020 Assessment Sheet



THE KENYA NATIONAL EXAMINATIONS COUNCIL
Kenya Certificate of Secondary Education

443/3 AGRICULTURE PROJECT 2020 ASSESSMENT SHEET

CANDIDATE'S NAME _____ INDEX NO _____

SCHOOL CODE _____ SCHOOL NAME _____

PROJECT A: ESTABLISHMENT OF TREE NURSERY

MILESTONE 1: Nursery establishment (To be uploaded by 31st March 2020)

ACTIVITY	ASSESSMENT PERIOD			SCORE AWARDED	MAXIMUM SCORE	PICTORIAL EVIDENCE AVAILABLE (YES/NO)
	JAN.	FEB.	MARCH			
1. PORTFOLIO OF EVIDENCE					08	
2. SOIL PREPARATION					16	
3. TREE NURSERY DESIGN					12	
4. POTTING					20	
5. PRICKING OUT					24	
6. SHADING					08	
7. WATERING					12	
Total out of 100					100	
Total out of 50 (divide the above total by two (2))					50	

Declaration by the Subject Teacher and School Principal
This is to declare that this candidate undertook the project and adhered to the conditions specified in the KNEC instructions. The evidence maintained in the candidate's project portfolio is a true reflection of the candidate's performance on the project.

	NAME	TSC NO	SIGNATURE	DATE
Agriculture teacher				
School principal				

Official school stamp

©2020 The Kenya National Examinations

**Appendix VI: Sample Question Paper for Joint Sub-County Agriculture Exams
Paper 1**

Name: Index No.:
Candidate's Sign.....
Date:.....

443/1
AGRICULTURE
PAPER 1
SEPTEMBER 2019

MATUNGU SUB-COUNTY JOINT EVALUATION EXAM

**443
Agriculture
Paper 1
2 hours**

INSTRUCTIONS TO CANDIDATES:

- *This paper consists of three sections, A B and C.*
- *Answer all questions in section A and B and any two questions from section C.*
- *All answers must be written in the spaces provided.*

For Examiner's Use Only.

SECTION	QUESTION	MAXIMUM SCORE	CANDIDATE'S SCORE
A	1-15	30	
B	16-20	20	
C	21-23	20	
		20	
	TOTAL SCORE	90	

SECTION A (20 MARKS)

1. State **FOUR** ways through which Agriculture contributes to Industrial development. (2mks)

2. state **FOUR** form of horticultural farming. (2mks)

3. State **FOUR** characteristics of shifting cultivation. (2mks)

4. Name **FOUR** beneficial biotic factors that influence crop production. (2mks)

5. State **FOUR** factors to consider when choosing the type of irrigation to use in the farm. (2mks)

6. State **FOUR** farming practices that ensure minimum tillage. (2mks)

7. Give **FOUR** reasons why land clearing should be carried out. (2mks)

Part B (10 Marks)

8.a) Name **THREE** vegetative planting materials used for propagating pineapples. (1 ½ mks)

b) Give **THREE** reasons why farmers are encouraged to use certified seeds for planting. (1 ½ mks)

State any **FOUR** effects of rainfall on Agriculture. (2mks)

10. State five importance of drainage as a land reclamation method. (2 ½ mks)

11. Give **TWO** reasons for adding well rotten organic manure to a compost heap. (2mks)

12. List **FOUR** soil factors that may lead to low crop yield. (2mks)

13. Give **FOUR** conditions that have led to fragmentation and subdivision of agricultural land in Kenya. (2mks)

14. State **FOUR** cultural methods of soil conservation. (2mks)

15. State **TWO** effects of Aphids on plants. (1mk)

SECTION B. (20 MARKS).

16. The diagram below illustrates a field management practice carried out to a fruit crop.



a) Identify the practice illustrated above. (1mk)

b) Give **TWO** reasons for carrying out the practice illustrated above. (2mks)

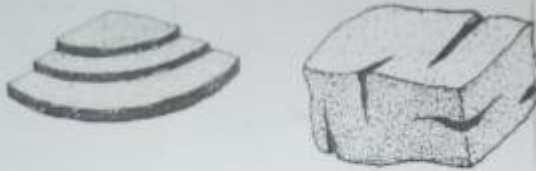
c) Name **ONE** crop onto which the practice is applicable. (1mk)

17. Study the weeds illustrated below and answer the questions that follows.



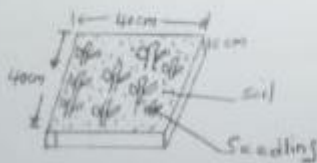
- Identify the weeds shown above. (2mks)
- Give economic importance of each of the weeds A and B. (1mk)
- State the advantage of weed C on the farm. (1mk)
- Give **TWO** reasons why weed D is difficult to control. (1mk)

18. The illustrations below represents types of soil structures. Study them carefully and answer the questions that follow.



- Identify the soil structure X and Y. (1mk)
- State **TWO** reasons why X is not suitable for growing maize. (2mks)
- Give **one** way of improving the soil structure X. (1mk)

19. Observe the following diagram and answer the questions that follow.



SECTION C (40 MARKS)

21. a) Describe **TEN** practices that a farmer should carry out to ensure uniform germination of seeds. (10mks)
b) Outline **SEVEN** advantages of land consolidation in land tenure reforms. (7mks)
c) List **THREE** factors considered in designing a crop rotational programme. (3mks)
- 22.a) Describe the establishment of kales under the following sub headings.
- i) Nursery preparation. (6mks)
 - ii) Transplanting (3mks)
 - iii) Field practices. (3mks)
- b) Explain **EIGHT** factors considered when spacing crops. (8mks)
- 23.a) State and explain **FIVE** factors influencing the efficiency of pesticides in crop production. (10mks)
- b) State **five** advantages of rotational grazing. (5mks)
 - c) Describe the procedure of harvesting tea. (7mks)

Appendix VII: Letter Of Introduction – UON



UNIVERSITY OF NAIROBI
COLLEGE OF HUMANITIES AND SOCIAL SCIENCES
FACULTY OF ARTS
DEPARTMENT OF PSYCHOLOGY

Telegrams: Varsity Nairobi
Telephone: 318262
Fax: 3245566
Telex 22095 varsity Ke Nairobi, Kenya

P.O. BOX 30197, 00100
NAIROBI
KENYA

8th October, 2020

The C.E.O,
National Commission for Science Technology and Innovation,
P. O. Box 30623, 00100
Nairobi

Dear Sir

REF: LETTER OF INTRODUCTION- MILDRED WERE-REG NO.
E58/84934/2016

The above named is a student in the Department of Psychology pursuing a Master of Education (Measurement and Evaluation). She has requested for a letter of introduction to enable her to collect data. She has successfully defended her research proposal at the Department. Her topic of research is: THE RELATIONSHIP BETWEEN TYPE OF ASSESSMENT PROCEDURE OF AGRICULTURE PROJECT AND THE RELIABILITY OF STUDENT SCORES IN AGRICULTURE IN MATUNGU SUB-COUNTY. Your kind support is highly appreciated

Sincerely,

A handwritten signature in blue ink, appearing to read 'C. Kimamo', written over a light-colored rectangular background.

Dr. C. Kimamo
Chair,
Department of Psychology

Appendix VIII: NACOSTI Research License



RIJTI RII OI ALNYA

Ref No: **891195**



**NATIONAL COMMISSION FOR
SCIENCE, TECHNOLOGY & INNOVATION**

Date of Issue: **14/October/2021**

RESEARCH LICENSE



This is to Certify that Ms. Mildred Ware of University of Nairobi, has been licensed to conduct research in Kakamega on the topic: RELATIONSHIP BETWEEN TYPE OF ASSESSMENT PROCEDURE OF AGRICULTURE PROJECT AND RELIABILITY OF STUDENTS SCORE IN AGRICULTURE THEORY IN MATUNGI SUB-COUNTY for the period ending: **14/October/2021.**

License No: **NACOSTI/20/7144**

Applicant Identification Number: **891195**


Director General

**NATIONAL COMMISSION FOR
SCIENCE, TECHNOLOGY &
INNOVATION**

Verification QR Code



NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

Appendix IX: Matungu Sub-County Permission Letter

MINISTRY OF EDUCATION, SCIENCE & TECHNOLOGY

Telegram:
Tel: 0202661941
Email: matunguedu@yahoo.com

When replying please quote our

Ref. and date



REPUBLIC OF KENYA

SUB COUNTY EDUCATION OFFICE,
MATUNGU SUB-COUNTY,
P.O. BOX 960 - 50102
MUMIAS.

Date: 15TH OCTOBER, 2020.

REF: MTG/ADM/26/1/61

To All
Principals of secondary schools
MATUNGU SUB- COUNTY

RE: MILDRED WERE: REG.M.E 58/84934/2016 RESEARCH

Permission has been granted to the above named student of the University of Nairobi pursuing Master of Education in measurement and evaluation to carry out research in secondary schools in Matungu Sub-County.

Accord her the necessary support.

Thank you.



Immaculate Obari
IMMACULATE OBARI
SUB- COUNTY DIRECTOR OF EDUCATION
MATUNGU SUB-COUNTY.