University of Massachusetts Amherst
ScholarWorks@UMass Amherst

Doctoral Dissertations 1896 - February 2014

1-1-2004

Content validity of independently constructed curriculum-based examinations.

Elias Watson Jani Chakwera
University of Massachusetts Amherst

Follow this and additional works at: https://scholarworks.umass.edu/dissertations_1

Recommended Citation
Chakwera, Elias Watson Jani, "Content validity of independently constructed curriculum-based examinations." (2004). *Doctoral Dissertations 1896 - February 2014*. 2371.
https://scholarworks.umass.edu/dissertations_1/2371

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

CONTENT VALIDITY OF INDEPENDENTLY CONSTRUCTED CURRICULUM-BASED EXAMINATIONS

A Dissertation Presented

by

ELIAS WATSON JANI CHAKWERA

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

May 2004

School of Education

© Copyright 2004 by Chakwera
All Rights Reserved

CONTENT VALIDITY OF INDEPENDENTLY
CONSTRUCTED CURRICULUM-BASED EXAMINATIONS

A Dissertation Presented

by

ELIAS WATSON JANI CHAKWERA

Approved as to style and content by:

Stephen G. Sireci, Chair

Ronald K. Hambleton, Member

Arnold D. Well, Member

Andrew Effrat, Dean
School of Education

DEDICATION

To Rose, for her love, patience, and encouragement

ACKNOWLEDGEMENTS

Many people contributed to my success in accomplishing the work of this dissertation. In a special way, I wish to thank the following: Dr. Stephen G. Sireci, Associate Professor in the Research and Evaluation Methods Program and Co-Chair of the Center for Assessment, University of Massachusetts Amherst, for chairing my dissertation committee and providing guidance throughout the study including reading all draft work; Dr. Ronald K. Hambleton, Distinguished Professor in the Research and Evaluation Methods Program and Co-Chair of the Center for Assessment, University of Massachusetts Amherst, who served as member of my committee and keenly followed on my progress and advised at different stages of the study; Dr. Arnold Well, Professor in the Department of Psychology, University of Massachusetts Amherst, for agreeing to serve as external member on my committee and providing guidance and support; Messrs. D. K. Khembo and N. A. C. Mhango, for conducting training of participants and leading the test review teams; Mr. C. M. Selemani-Mbewe, for assisting in the data collection; and all participants in the study.

Finally, I thank the directors of the UPIC project at the Center for International Education, University of Massachusetts Amherst, and the Secretary for Education, Science and Technology in Malawi, for providing financial support.

ABSTRACT

CONTENT VALIDITY OF INDEPENDENTLY
CONSTRUCTED CURRICULUM-BASED EXAMINATIONS

MAY 2004

ELIAS WATSON JANI CHAKWERA, B.Ed., UNIVERSITY OF MALAWI

M.Ed, UNIVERSITY OF MALAWI

Ed.D. UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Stephen G. Sireci

This study investigated the content validity of two independently constructed tests based on the Malawi School Certificate History syllabus. The key question was: To what extent do independently constructed examinations equivalently sample items from the same content and cognitive domains? This question was meant to examine the assumption that tests based on the same syllabus produce results that can be interpreted in similar manner in certification or promotion decisions on examinees without regard to the examination they took. In Malawi, such a study was important to provide evidence for the justification for using national examination results in placement and selection decisions.

Based on Cronbach's (1971) proposal, two teams of three teachers were drawn from six schools that were purposefully selected to participate in this study. Each team constructed a test using the Malawi School Certificate of Education (MSCE) History

syllabus. The two tests were put together in a common mock examination, which was first piloted before the final form. Two hundred examinees from the participating schools took the common mock examination. Paired scores from the two tests and the same examinees' scores on MSCE History 1A were used in the analysis of testing the mean difference of dependent samples and variance comparison. Subject matter experts' ratings were used to evaluate content and cognitive relevance of the items in the test.

The findings indicate that MSCE syllabus was a well-defined operational universe of admissible observations because independently constructed tests equivalently tapped the same content. Their mean difference was not statistically different from zero and the mean of the squared difference scores was less than the sum of the split-half error variances. It was therefore, concluded that the two independently constructed were statistically equivalent. The two tests were also found to be statistically equivalent to the 2003 MSCE History 1A. However, the presence of stray items indicated syllabus looseness that needed redress to improve content coverage. Inadequacy in the rating of cognitive levels was noted as a problem for further research. The need to improve examinations was advocated in view of the their great influence in instruction and assessment decisions or practices.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

ABSTRACT vi

LIST OF TABLES xiii

CHAPTER

1. INTRODUCTION

1.1 Contextual Setting of the problem 1
1.2 Purpose of the Study and Research Questions 2

1.2.1 Definition of the Problem 2
1.2.2 Practical Relevance of the Study 5
1.2.3 The use of Test Samples in Content Validation 6

1.3 The Malawi Education System and Examination Validity 7

1.3.1 The Malawi Geographical Setting and School Enrolments 8
1.3.2 The Malawi Examination System 9

1.4 Concerns of Content Validity 11

1.4.1 The Role of Content Validity 12

1.5 Significance of and Limitations of the Study 13

1.5.1 Significance of the Study 13
1.5.2 Limitations of the Study 18

2. LITERATURE REVIEW 22

2.1 Importance of Theoretical Background for Content Validity 22
2.2 Development of Validity Theory 22

2.2.1 Categories of Validity 24
2.2.2 A Shift Toward a Unified Concept of Construct Validity 25

2.3 Prominent Role of Content Validity as Test Quality 27

2.3.1 Defining Content Validity 28
2.3.2 Concepts Related to Content Validity 30
2.3.3 How Content Validity Relates to Construct Validity 34

2.4 Content Validation Procedures 37

2.4.1 Subjective Methods 37
2.4.2 Empirical Methods 42
2.4.3 Pseudo-Empirical Methods 43

2.5 Content Validity Studies in Malawi 44
2.6 How Duplicate Test Construction Relates to Content Validity 45

3 METHODOLODGY 48

3.1 Selection of the Sample of Participants 48

3.1.1 Description of Participating Schools 48
3.1.2 Description of Participants 50

3.2 Data Collection Instruments (Curriculum-Based Tests) 52

3.2.1 MSCE History 1A Examination 52
3.2.2 Common Mock Examination 53
3.2.3 Class Tests 54
3.2.4 Summary of the MSCE Tests Used in the Study 54

3.3 Data Collection Instruments (Surveys) 55

3.3.1 Teacher-Participants' Surveys 56
3.3.2 Student-Participants' Survey 57
3.3.3 Subject Matter Experts Rating Scales and Surveys 58

3.4 Procedures for Data Collection with Duplicate Tests 58

3.4.1 Common Mock History Examination 59
3.4.2 Review of Duplicate Tests' Items 59
3.4.3 Administration of the Common Mock Examination 61
3.4.4 Impact of Duplicate Test Construction Activities on Teacher Participants 62

3.5 Procedures for Data Collection with Teacher and Student Surveys 63

3.5.1 Teachers' Survey on Adequacy of Student Preparation 63
3.5.2 Students' Survey on Adequacy of Instruction they Received 64
3.5.3 Collection of Item-Topic and Item-Cognitive Relevance Ratings 65
3.5.4 Collection of Data on Test Quality 66

3.6 Data Analysis Procedures 67

3.6.1 Mean of the Squared Difference Scores and Sum of the Error Variances of Mock A and Mock B 68
3.6.2 Computation of Descriptive Statistics and Proportion of Agreements 69

3.7 Summary of Data Collection and Analysis Procedures 71

3.7.1 Conclusion 72

4 PRESENTATION OF RESULTS AND ANALYSIS 73

4.1 Content Validity Evidence from the Duplicate Constructed Tests 73

4.1.1 Computation of the Mean of Difference and Standard Errors of Measurement 74
4.1.2 Paired Dependent Sample Mean Difference t-Test and Repeated Measures ANOVA 76

4.2 Judgmental Evidence for Content and Cognitive Skills Relevance of the Items 77

4.2.1 Test Designers' item Allocation to Content and Cognitive Domain Area 78
4.2.2 Content Relevance Ratings 81
4.2.3 Inclusion of Items Not Related to the MSCE History Syllabus 83
4.2.4 Cognitive Relevance Ratings of Items Across Tests 85

4.3 Perceived Overall Quality Impressions on the Mock Tests and MSCE 1A 87

4.3.1 Appropriateness of Language and Item Clarity 87
4.3.2 Overall Test Difficulty 90

4.4 Equivalence of Independently Constructed Mock Examinations 92

4.4.1 Participants' Impressions about the Test Construction Experiment 92

4.5 Evidence of Examination Influence on Instruction and Assessment Practices 94

4.5.1 Teachers' Instructional Emphasis Prior to Mock Examination 95
4.5.2 Evidence for Assessment of Examinees During Instruction 97
4.5.3 Students' Perception of Instruction and Assessment Prior to National Examination 99

4.6 Conclusion 103

5 DISCUSSION AND CONCLUSION 104

5.1 Statistical Equivalence of Independently Constructed Duplicate Tests 105

5.1.1 Content Validity Evidence from Statistical Equivalence of Duplicate Tests 105
5.1.2 Test Specifications and SMEs' Content Ratings Validity Evidence 107
5.1.3 Lessons from the Cognitive Skills Rating Relevance Ratings 110

5.2 Construction of Equivalent Duplicate Tests 112

5.2.1 SMEs and Students' Perception about Test Quality 113
5.2.2 Impact of Examinations on Classroom Practices 115

5.3 Implications and Recommendations of the Study 118
5.4 Conclusion and Summary 123

APPENDICES

A SET 1 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MOCK A) 125

B SET 2 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MOCK B) 136

C SET 3 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (CLASS TEST) 146

D SET 4 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MSCE 1A) 155

E QUESTIONNAIRE OF TEACHERS' PERCEPTION ON EXAMINEE PREPARATION 165

F QUESTIONNAIRE OF TEACHERS' IMPRESSIONS ABOUT THE IMPACT OF THE DUPLICATE TEST CONSTRUCTION EXPERIMENT 174

G STUDENTS PERCEPTION ABOUT THEIR PREPARATION FOR MSCE HISTORY EXAMINATION 176

H QUESTIONNAIRE ON STUDENTS' OPINION ABOUT TEST QUALITY 178

I CONTENT COVERAGE RATING SCALE FOR MSCE HISTORY ITEMS 179

J COGNITIVE SKILL RATING SCALE FOR THE MSCE HISTORY ITEMS 183

K SAMPLE CONSENT LETTER 185

L SUBJECT MATTER EXPERT'S QUESTIONNAIRE ON OVERALL TEST QUALITY 186

BIBLIOGRAPHY 187

LIST OF TABLES

Table | Page

1.1 MSCE History Percentage Pass Rates 1992 to 2003 11

3.1 Summary Description of Test Instruments Developed from the MSCE History Syllabus 55

3.2 Summary of Data Collection and Analysis Procedures 71

4.1 Means, Standard Deviations and Standard Errors of Measurement of Mock Tests 75

4.2 Summary of One-Way Repeated Measures ANOVA 77

4.3 Test Specification Tables for Mock A, Mock B, and MSCE 1A 79

4.4 SMEs Average Rating Determination of the Number of Items per Content Domain Area 81

4.5 Mean Aiken Values for Items Per Topic Across 83

4.6 Items Identified as Drawn From Outside the Content Domain Across Tests and their Aiken's Validity Indices (V) 84

4.7 Summary of Average Item Cognitive Ratings Per Domain Area Per Topic 86

4.8 Percentage of Students' Ratings for Overall Test Clarity of Language Vocabulary 88

4.9 Summary of One-Way Repeated Measures ANOVA for Students' Ratings on Clarity and Language Vocabulary Appropriateness Across Tests 88

4.10 Percentage of SMEs' Ratings on Clarity of Test Items 89

4.11 Summary of One-Way Repeated Measures ANOVA of SMEs' Ratings on Item Clarity 90

4.12 Percentage of Students' Overall Rating of Test Difficulty 90

4.13 Summary of One-Way Repeated Measures ANOVA of Studnts' Ratings on

Item Difficulty 91

4.14 Frequency Ratings of Teachers' Impressions about Duplicate Test Construction Experiment 93

4.15 Mean and Standard Deviation of Teachers' Ratings on Importance and Coverage of Syllabus Topics 96

4.16 Percentage of Teacher Agreement on Topics on which Students Were Assessed 98

4.17 Percentage of Students' Choice of Topic Coverage Categories Across Schools 100

4.18 Mean and Standard Deviation of a Distribution of Student Percentages of Choice of the Most and Least Adequately Prepared Topics 101

CHAPTER 1

INTRODUCTION

1.1 Contextual Setting of the Problem

To construct statistically parallel or equivalent tests is a desired goal for repeated testing over the same objectives and content domain. In high-stakes testing, an achievement of this goal would enhance the validity of score interpretations based on students' performance on the tests. Qualifications that are certified by the same examination over the years are considered equivalent on the assumption that the tests that are drawn from the same domain to assess the expected competencies are equivalent and would classify persons in the same way from year to year. In verifying this, it is of particular interest to determine how tests would equally sample items from the same domain and retain similarity in test difficulty.

This study investigated content and cognitive relevance of independently constructed tests based on the Malawi School Certificate of Education (MSCE) History syllabus (Curriculum frameworks). In doing this, a duplicate test construction experiment was conducted to deeply examine the problem of content validity. As an introduction, this chapter describes the context of the problem that was investigated in section 1.1. Section 1.2 outlines the problem, its practical relevance and research questions; Section 1.3 describes the importance of the problem in the context of Malawi Education System; why the problem fits into the content validity question is discussed in section 1.4; and section 1.5, describes the significance and limitations of the study.

1.2 Purpose of the investigation and Research Questions

The extent to which tests equivalently sample the same domain on repeated testing is of great importance in high-stakes testing with increased accountability. This expectation is based on the assumption that each sample of the tests equally draws from the universe of admissible observations on which a person with a given ability in the subject matter would retain the same score on repeated testing within the standard error of measurement, regardless of the sample of the test administered. Further to this assumption is the belief that the content domain is so clearly defined and that a universe of items that could be tapped exists (Cronbach, 1971). A clearly defined domain would mean that any competent test designer would come up with items that would adequately represent the domain. On the other hand, the existence of the item universe would ensure that items of equal difficulty would be drawn by equivalent test samples. The purpose of this study was to provide a practical investigation of these assumptions and draw insights that could inform on test construction and curriculum development. The study used the MSCE History syllabus as the universe definition on which the investigation was focused.

1.2.1 Definition of the Problem

The MSCE syllabus defines the content of the subject matter to be covered for those studying History in the last two years of secondary education. It guides both teachers and test designers about what to include in instruction and assessment of those who study the subject. While the syllabus remains the same, tests and examinees change

overtime and yet decisions that are made out of such examinations are treated as similar as if examinees took the same examination. The question of interest was whether two teams using the same sampling guidelines would draw up equivalent tests if they worked independently in selecting items for MSCE History examination. The investigation involved a duplicate test construction experiment that was proposed by Cronbach (1971). The focus of the investigation was to find out how well each test sampled the defined operational universe and how the two tests compared in their sampling of the universe. In each case, the investigation was concerned with content validation of the duplicate-constructed tests. The comparison was extended to the 2003 MSCE 1A examination to determine the extent to which the duplicate tests individually sampled the content domain in relation to the national examination on which the two tests were modeled.

The experiment used both judgmental ratings from subject matter experts (SMEs) and empirical data from examinee performance in the tests to derive content validity evidence. Because MSCE is an achievement examination, examinee preparation was considered critical for the assumptions in the investigation to hold. Issues of student preparation to establish a link between assessment and classroom practices were examined to get more information for the equivalence of the tests.

The experiment provided an opportunity to find out whether teachers and examiners had a common understanding of the MSCE History syllabus. In high-stakes examinations, both examinees and instructors are concerned about how well examinees should prepare for the upcoming examinations. But, increasing anxiety to meet the demands of examinations have often led to negative consequences, such as, curricular

reduction, examination drills rather than teaching, and teacher modeled dishonesty (Popham, 2003; Chakwera, Khembo, & Sireci, in press). These undesirable effects can be avoided if items on a test match the instructional objectives that are articulated in a syllabus. To ensure such a match is one way of demonstrating that both the test constructor and the instructor have a common understanding of the curriculum that is taught and assessed.

The following were the questions investigated in this study:

- To what extent do independent examinations developed from the same syllabus equally the domain of tasks described by the educational objectives contained in the syllabus?
- How well do items on each mock subtest represent the items on the national examination in terms of content and cognitive levels of achievement?
- How well do the two subtests in the mock examination, end of term class test, and the national examination compare in quality in terms of item clarity and difficulty?
- How well do items on end of term class tests represent the MSCE History examination content?
- To what extent do classroom practices reflect the impact of MSCE examinations in History instruction?

In exploring these questions, the validity of MSCE History examinations and the influence of high-stakes testing on classroom assessment practice were established. By involving teachers in developing parallel common tests, both their expertise and experience were broadened with respect to their understanding of the History

curriculum and assessment process. In that way, the investigation provided an opportunity for teachers to upgrade their assessment skills, which must have enhanced their professional growth.

1.2.2 Practical Relevance of the Study

The major purpose of this study was to investigate how well independently drawn up tests would adequately assess the same content and cognitive domains. Because teachers of History drawn from six schools formed the two teams that independently constructed the duplicate tests, the experiment also demonstrated whether MSCE History teachers and national examination developers had common understanding of the History syllabus. In both cases, the comparison of students' performance on the tests was the basis for understanding the relationship between the duplicate tests and the national examination in the same subject. Examining the correlations between paired scores and testing the hypothesis of the dependent sample mean difference between the two independently constructed tests and with the MSCE 1A accomplished this. The study used end-of-term class tests to investigate the emphasis that teachers put on the topics that were taught in anticipation of the national examination. The study also established the degree of relevance between items in each test and topics of the syllabus to determine content validity of each test. With this information, it became possible to compare the adequacy of content coverage of each test with other tests. In addition, the study compared the tests in terms of their weighting with respect to the levels of cognitive abilities that the items of each test represented.

Teachers are the most important stakeholders, apart from their students, of the national examinations. As instructional designers they always aim at enabling their students to achieve maximally on the high-stakes examinations by focusing their instruction on the content and skills they believe to be examinable. This experiment had practical relevance because it directly compared teachers' work in assessment to that of national examination designers. While playing the role of test designers, the teachers did not only understand the MSCE syllabus better but also the link between instruction and assessment. This was another practical relevance of the duplicate test construction experiment because it promoted professional growth of the teachers. This was considered important in view of the fact that despite their involvement in student assessment, "many teachers receive little or no training in how to construct good achievement tests" (Gronlund, 1977, p. 1). The study attempted to analyze performance characteristics and cognitive skills that the items measured, as teachers understood them. This was important because the teachers' understanding of students' performance characteristics and the state of cognitive skills that are expected is the basis for formulating desired instructional goals and objectives (Nitko, 1989). For these reasons, this experiment was conducted from the context of acceptable goals that would promote mutual benefits for the researcher as well as the participants, including the examinees.

1.2.3 The Use of Test Samples in Content Validation

While teaching artifacts can provide evidence of how examinations influence instruction, teacher-made tests can provide information about how examinations influence classroom assessment practices. It should, therefore, be reasonable to treat the

teacher-made tests in each examinable subject as alternate samples of the national examinations when appropriate parallel procedures in their development are satisfied. For this reason, the study examined the content validity of teacher-made revision test (classroom tests administered by individual teachers prior to the experiment), mock examination (two independently constructed tests on which the experiment was based), and national examination (MSCE 1A). The two teams were assumed to have similar skills and used same test construction guidelines and the MSCE History syllabus but independently constructed test specifications and items, and they also independently selected their items to build the final tests. The extent of agreement for the two test forms in content representation in relation to the curriculum topics and objectives contained in the MSCE syllabus was an indicator of sample equivalence and content coverage adequacy. A high correlation between scores from the two forms of the mock examination and between each mock subtest and the national examination were used as the measure of the extent of agreement in terms of how well the two tests similarly assess the same levels of cognitive domain.

1.3 The Malawi Education System and Examination Validity

Formal education in Malawi was introduced by a number of Christian missionaries in the last quarter of the 19th century. In those early years, each mission designed its own curriculum in the schools under their authority. The common aim however, was the teaching of reading, writing and arithmetic and most missionaries did not see the need for higher education for Africans (Banda, 1997). Due to the diversity of the curriculum and objectives of different missions, it was not possible to have common

examinations until the 1940s when secondary education begun. The limited growth of secondary schools increased the importance of examinations, which became a tool for selection as secondary schools emerged in the 1940s. Since then three national examinations have been administered at school level. The Primary School Leaving Certificate Examination (PSLCE) is administered after 8 years of primary (Grade 8); Junior Certificate Examination (JCE) is administered at the end of second year of secondary (Grade 10); and the MSCE examination is administered for the exit of secondary education (Grade 12). Before independence from the British colonial authority, secondary education was the highest formal learning that Africans would get. When, in 1965, the University of Malawi was opened, the secondary exit examination assumed the gate-keeping role as the other exams in the school system.

1.3.1 Malawi Geographical Setting and School Enrolments

Malawi is a land-locked country in southeast Africa lying between 9 and 17 degrees south of the Equator. It shares boundaries with Tanzania to the north and northeast; Zambia to the west and northwest; and Mozambique to south, southeast, and southwest. The country covers a surface area of 118,500 square kilometers of which 94,276 are land and 24,224 are water (mainly covered by the Lake Malawi). The country has three administrative regions, namely the north, which is mountainous and about 2,500 meters above sea level; the Center is largely a plain at an attitude of 1,500 meters and a strip of lower attitude along the Lake Malawi; and the South with scattered mountains but largely low flat plains at 500 meters above sea level (Tindall, 1992). The population of Malawi is currently about 11 million people with a growth rate of 1.9%.

There are about 5,000 primary schools, which have enrolled about 3.1 million kids of between 6 to 16 years old. Approximately 152,000 students are enrolled in public secondary education (Ministry of Education, 2002).

The Malawi education system comprises 8 years of primary education, 4 years of secondary, and an average of 4 years of university or tertiary education. The Primary Education Sector, which constitutes the backbone of Basic Education, is a priority sector just as basic education is considered to be a human right. In 1994, Government introduced the Free Primary Education Policy (FPE) with an intention of achieving Universal Primary Education (UPE) (PIF, 2000).

The PSLCE enrolls the highest number of examinees as majority of young people exit formal education at the primary level. In the last few years this enrolment has gone up to slightly over 150,000. The pass rate for the PSLCE has often been around 75%. Of those who pass the examination, only 22% get selected to secondary school. The MSCE is another high-stakes examination because it is used to select about 3% of those who go to university (Chakwera, et al., in press). The two high-stakes examinations tend to drive the education system as both instructional and assessment decisions are evaluated in terms of how well they effectively support examinees to compete favorably in the selection. Therefore, validity of such examinations is of paramount importance for the "cutthroat" decisions to be accepted.

1.3.2 Malawi Examination System

The Malawi National Examinations Board (MANEB) is responsible for the development and administration of all school and teacher education (primary)

examinations. MANEB operates as a non profit-making organization that is fully financed by the Malawi government. It often gets pressurized due to high number of examinations and examinees involved at each level. For example, at the MSCE level MANEB administers 50 examinations in 24 subjects and it registers over 150,000 examinees for the PSLCE. Despite such a large-scale assessment process, the MANEB has limited financial and human resources yet administration costs and examinee numbers continue to rise each year. Because of increasing administrative problems psychometric issues such as, scaling and equating tend to be sidelined in view of competing priorities over limited funding. Large-scale studies on validity and reliability are not carried out due to resource limitations.

Content validity is ensured during test construction by moderation (item review) exercise prior to final compilation of the test papers. Moderation is the final review that gives an opportunity for SMEs to check on the clarity, specificity, item difficulty and the correctness of the selected keys in case of objectively scored items. By reviewing items on a complete test, SMEs also check on content coverage and adequacy of the time allocated to the test. MANEB also ensures standardized practice during administration and scoring of each examination. For PSLCE and JCE, MANEB uses letter grades for the award of certificates while a 9-point scale is used for the MSCE examination. This scale has four categories, and these are: Distinction for points 1 and 2; Credit for points 3 to 6; Pass for points 7 and 8; and Fail for point 9.

However, the need for psychometric research on the examinations is apparent. Table 1.1 presents deteriorating and fluctuating pass rates for MSCE History over the past ten years:

Table 1.1. MSCE History Percentage Pass Rates 1994 to 2003

Year	# Examinees	# Passed	% Pass Rate
1994	7762	4267	54.97
1995	9848	3802	38.61
1996	8753	4365	49.87
1997	10711	3960	36.97
1998	13531	4164	30.77
1999	13236	4825	36.45
2000	14408	7047	48.91
2001	17723	7046	44.81
2002	15463	9802	63.39
2003	9063	6092	67.22

Observed fluctuations in the pass rates from year to year may not only be due to student differences. It was felt that the quality of examinations that are administered every year might contribute to the inconsistency in the pass rates. This analysis provided a justification for this study, which aimed at evaluating content validity of independently constructed tests and correlate them with the criterion based on the same syllabus.

1.4 Concerns of Content Validity

The pressure of high-stakes testing raises the concern of adequacy of students' preparation for both schools and examinees. In achievement tests, such as the Malawi national examinations, the critical validity question is: how adequately do both instruction and examination cover the content domain of interest? The content domain for such examinations is operationally defined by topics and objectives that make up the teaching and assessment syllabus. Therefore, concerns about adequacy of preparation and test coverage of domain of interest are closely related issues of content validity.

1.4.1 The Role of Content Validity in Assessment

Content validity is an aspect of test quality that shows the extent to which the items in a test represent the domain of tasks from which they are drawn. It also demonstrates a match between the items and objectives measured by the test as well as determining the relevance of such objectives. According to Cronbach (1971) ensuring content validity begins "by translating each educational objective into task specifications in relation to behavior relevant to each topic" (p. 457). In assessment of achievement that is based on instruction, both instructors and examiners need to have a common understanding of the content domain, and both must similarly translate the content domain into educational objectives. Otherwise lack of a common understanding would result in a mismatch between student preparation and test coverage. In such a context, an evaluation of content validity would seek to establish whether instructors and test developers have similar understanding of the breadth, depth or extent of the domain of interest as expressed by the syllabus. The similarity of understanding can be established if content validity of several forms of tests drawn from the same domain, provide evidence of adequate and relatively equal representation of the domain of interest. With respect to national examinations, one needs to evaluate both classroom tests and national examinations to establish the extent to which both teachers and examiners understand the curriculum.

Studies have often shown that high-stakes examinations tend to influence instruction and assessment practices. Steeves, Hodgson, and Peterson (2002) observed the tendency for teachers to cover material that they believe would most likely be

included in assessment to ensure students' success. Such a practice is often referred to as teaching to the test, which can have positive as well as negative consequences. One commonly cited negative effect of teaching to the test is the narrowing of the curriculum resulting from intentional elimination of those parts of the curriculum that are not covered in the high- stakes tests (Steeves, et al., 2002; Shepard, et al., 1992; Chakwera, et al., in press; Mucheru, 2003). Because one objective of national examinations in Malawi is to monitor educational progress, it is necessary to frequently evaluate how examinations relate to what is taught. This is important because the national examinations in Malawi are really high-stakes at every level of the education system. Teachers often sideline innovations that emphasize coverage of curriculum objectives for their own sake because of the "cutthroat" competition that the "bottle necked" education system entails.

1.5 Significance and Limitations of the Study

This study met both theoretical and practical requirements for an education research. The new knowledge from its findings is likely to influence test construction practices and illuminate the existing knowledge of content validity. Its significance and caution for interpretation of results in view of the limitations of the study are described in this section.

1.5.1 Significance of the Study

It should be expected that in high-stakes assessments, both the test constructors and instructors must articulate similar interpretations of the curriculum while

constructing tests or prepare examinees for the tests, respectively. But as Ebel (1962) observed, such common understanding is not always possible because in some cases test specifications exist in the mind of the test constructor and written test construction guidelines tend to focus more on the format rather than the content of the test. Until recently, MANEB did not use test specifications when constructing tests and so the specifications used to be in the mind of the test constructors, as in the observation cited above. The current practice however, is that the team of test constructors is asked to draw up a table of test specifications to be used in the construction of the test. This means that each team of test constructors comes up with its own test specification for the test it develops each year. But generally, the teams use the same guidelines in terms the type of items and the layout of the papers.

This was a good reason for the study, as it aimed at establishing the extent to which both teachers' and national examination constructors agree on the content to be assessed in MSCE History examination. The specifications for each test were examined in terms of the weighting of the topics that were selected by each team to establish the importance of the topics in the syllabus. Such an approach was considered appropriate to check on the MANEB assumption that different teams that independently draw up specifications would construct equivalent tests when provided with the same guidelines. The use of one test specification table for subsequent tests is considered favorable for high-stakes testing because results from such tests receive similar interpretation on the basis of equivalent content coverage. In principal, high-stakes testing restricts attention to content validity to the selection of test materials and the universe description without considering the persons to be tested (Cronbach, 1971). Yet, as Ebel (1962) observed,

> Seldom are the test specifications sufficiently explicit and so comprehensive that competent test constructors, working independently, could be expected to produce forms on which raw scores are essentially equivalent. (pp. 21-22)

Apart from examining the item-objective/topic relevance and weighting of the test specifications, this study also evaluated the decisions that influenced instructors in making tests that suited the preparation they offered to their students. By evaluating the content validity of the end of term tests, information about the emphasis that the topics that were actually taught received was obtained. In addition, both teacher participants and examinees provided information about the preparation they made in view of the anticipated national examination. It was expected that results from such analyses would help MANEB to determine preparation of instructions to item writers, which could be useful in narrowing the gap of misunderstanding between examiners and instructors. In addition, the results of the study have provided MANEB with feedback on the current practice, which should help in shaping the direction for test development.

The findings of this research were expected to lead to improvements in test development as well as enhance score interpretation in national examinations. As Ebel (1962) observed,

> If more systematic and standardized processes of test production could be developed and used, our educational measurements should become not only more consistently reproducible, but what is perhaps even more important, they should become more meaningful. (p. 22)

It was therefore, expected that in response to the findings of the study MANEB would be more explicit in stating the operations by which tests are produced so as to ensure that different test constructors have same understanding of the guidelines and the content domain on which items are developed. The extent to which the systematic and

standardized processes of test construction are achieved in Malawi would be a major step toward removing teacher anxiety that has often led to collaborated cheating.

The teacher participation in duplicate test construction experiment helped to lay foundations for teacher collaboration and collegiality, in neighboring schools. With increasing accountability of high-stakes testing, teachers in different schools tend to see each other as competitors and this limits opportunities for sharing knowledge and resources that can benefit students. The duplicate test construction experiment benefited both teacher participants and examinees because the common mock examination had broader focus due to team input than would have been the case if a single teacher developed it. It is expected that after this experiment the teacher participants would still find it appropriate to work together to achieve common objectives. In a country like Malawi, with varying teacher qualifications and critical shortage of qualified teachers in secondary schools, collaboration in assessment could be one way of improving students' preparation and increase teacher and students' confidence in national examinations. To the extent that this objective is achieved, this study would go a long way in enhancing professional growth through academic collaboration.

The duplicate test construction experiment was conducted to confirm or disconfirm the assumption that a test is a representative sample of content domain. Cronbach (1971) noted a danger of vagueness in any reference to content universe arising from the understanding that content is an ill-shaped and undifferentiated mass. He also observed that even when the content domain is defined, a universe of items does not exist except for the items in the test. Therefore, if the two test construction teams would produce equivalent test forms, some evidence for content universe in

MSCE History would be established. The results were also expected to shed light on clarity of the operational definition of the universe of admissible observations and test construction guidelines that MANEB uses.

The ultimate objective of this study was to contribute to improvement in test construction and student preparation for high-stakes testing. It is therefore, justified in as much as the findings would help teachers and test constructors to establish a common understanding of the content domain on which they focus teaching and assessment. It was also educationally justified because of opening up opportunities for teacher collaboration, which has been lacking due to competition among teachers perpetuated by increasing accountability of high-stakes testing. Through this research, both teachers and examinees are likely to achieve their assessment goals in national examinations. The common mock examination accorded the examinees an opportunity for an external test prior to national examinations, which most likely enhanced their preparation by providing feedback on their mastery of the subject matter.

Following this study, MANEB would most likely prepare test construction guidelines based on empirical evidence. Because MANEB trainers were used to conduct test construction training, the results from this study provided some feedback on the effect of the training that was offered. The extent of agreement in content validity of the two independently constructed tests gave insight about what test constructors need to know. It should be expected that results of this study would spark the desire for similar studies in other subject areas, which eventually could lead to improvement of the entire examination system. Some insight on curricular ambiguities became apparent from the

differences in the syllabus or objective interpretations that the two test construction teams made. To that end, the study set out to inform on the focus of curriculum review.

1.5.2 Limitations of the Study

Procedurally, it is not possible to completely duplicate the construction of a test because persons who are in independent teams have different belief systems regarding what is important in the content domain. The teams also work in different conditions, which can influence them differently on the material they select for assessment. Such differences become apparent when teams of test constructors develop test specifications independently. In the present study, there were remarkable differences between the constructors of the national examination and those of the common mock tests. MANEB uses trained test developers but classroom teachers were used for the common mock test construction. Although MANEB-like training was given to the teachers, their item construction or selection of items would probably have been guided by their knowledge of the examinees that they teach while national test developers would be more material focused and less examinee minded. This should be noted as a departure from the ideal that Cronbach (1971) proposed. The design of the current study was chosen to establish ownership of the common mock tests with the teachers so that their classroom obligations could be exploited in favor of the experiment. Otherwise, it was not possible to use MANEB for such an experiment, as it would seem a waste of examinees' time if it were linked to live examination. It would also have been difficult to justify its expenditures when the national examinations are plagued with insufficient funding.

Of practical importance is the fact that the duplicate test construction experiment was also meant to equip teachers with skills of test construction. To ensure this objective, the same constructors were used to review each other's items to give the teachers the practical experience of test review. The two teams swapped the tests to ensure both neutrality and objectivity in the review. However, by promoting independent review through the swapping of the tests, the test construction independence might have been compromised. An attempt to retain test construction independence was made by not allowing the teams to remove items or reword item stems. The reviewers were instructed to work only on grammar and factual errors to remove ambiguity and improve factual accuracy. The extent to which this practice might have affected the results of this experiment limits interpretation of its findings.

Another limitation connected with the use of classroom teachers to construct, administer, and score examinations of this kind in their own schools is that one could not guarantee standardized practices as is done in MSCE national testing. Because teachers have always competed over getting students to pass the national examinations, there could be a possibility that at some point, cheating elements might have occurred and contaminated the results of the common test. If that happened, then the common mock tests would not ably represent the assessment that the national examinations represent. Issues of fairness were however stressed during the experiment, but the failure of the study to enforce standardized practices during the administration and scoring of the tests in the manner that MANEB does, is a serious limitation.

Absence of a common table of test specifications is another limitation of this study because content comparison between the test forms was not based on the same

proposed item allocations. Instead, it was based on the extent to which the items reflected the weighting that the teams attached to each topic of the syllabus. For this reason, the test specifications were part of what was to be compared between different teams and not part of the guidelines provided to teams of test constructors, as would have been the case in an ideal duplicative test construction process. However, this departure was considered inevitable to fit the study in the context of the MANEB practices, but this limits interpretation of the findings to the Malawian and similar contexts. The interpretation of the results was therefore, made in terms of overall representation of the content as described by the MSCE syllabus within the limits of the teams' test specifications. And the weighting of the topics by each team as provided by the specifications reflected the extent to which teams of test developers differ in selecting portions of the content domain to be assessed in each test.

This study used multiple-choice items in the two duplicate constructed tests. In a typical MSCE History examination the multiple-choice items constitute only one part of the full test. The other component comprises short-answer and essay type items. For this reason, the findings of the study cannot be generalized to the all History examinations because the other component could have different weighting of content and cognitive domain coverage.

A final notable limitation became apparent after the data collection. This was about the frequent teacher change over in instructional assignments. It was assumed in the study design that the same teachers would have taught the same students from form 3 up to form 4, which the study targeted. However, 67% of the teachers that were involved in the duplicate test construction experiment had just taken over the classes in

form 4 and so they did not feel responsible for what happened while the students were in form 3. This was a limitation because the teachers' information of how the topics that others' taught were covered would be from the second source. For these reasons, interpretation of the findings on this study could not be generalized beyond the context in which the study was conducted.

CHAPTER 2

REVIEW OF LITERATURE

2.1 Importance of Theoretical Basis of Content Validity

The way in which test items represent the objectives or topics of the syllabus from which the items are drawn, is of interest in achievement testing. Content validity is the quality of a test, which is satisfied when the sample of test items is representative of the domain of knowledge being measured. To appreciate the importance of content validity it is necessary to review the literature on the validity theoretical framework. This chapter is therefore, an attempt to examine the theoretical basis for content validity as test quality and procedures for establishing its evidence. The review of literature, which follows, is reported in four sections. Section 2.1 discuses theoretical basis of content validity; Section 2.2 is about the development of validity theory; section 2.3 describes important role of content validity in determining test quality; section 2.4 discuses content validation procedures; section 2.5 presents relevant content validity studies done in Malawi; and section 2.6 describes how the duplicate test construction experiment relates to content validity.

2.2 Development of Validity Theory

Over the decades, validity has ensued academic debate because of conceptualization changes that the term has gone through. Cronbach (1971), Angoff (1988), and Sireci (1998b), outlined some of the critical changes that have dominated the arguments and are responsible for the theoretical basis that determine validation

studies. Angoff (1988) observed that from the 1930s to the 1950s, the principle of prediction dominated the concept of validity and correlation was a procedure for checking validity evidence. It is no wonder that definitions of validity from the scholars of that period as quoted by Angoff (1988) include

> A correlation of scores on a test with some other objective measure (Bingham, 1937); a test is valid for anything with which it correlates (Guilford, 1946); it is a correlation between the fallible and infallible measure of a trait (Lindquist, 1942). p. 20

This conceptualization of validity emphasized the need to have external link with some other measure to establish validity of a test. According to Jenkins (1946) psychologists were more concerned with construction of predictors without realizing the criterion problem. They very much chose the criterion on the basis of convenience rather than its adequacy, which became apparent in the period after the Second World War. Therefore, for some time the problem of the criterion dominated the debate on validity. For example, Ebel (1961) noted that the problem of the criterion, very much reflected on the logical and operational limitation of the concept of validity itself and he further stated:

> Even in those rare cases where criterion measures have been painstakingly devised, the validity of the test is not determined unless the validity of the criterion has been established. This requires a criterion for another criterion, and so on ad infinitum. We can pursue such an infinite regress until we are weary without finding a self-sufficient foundation for a claim that a test is valid. It is an unhappy fact that the general conceptual definition of validity provides no firm basis for operational definitions of validity. (p. 642)

In view of these problems, other aspects of validity began to receive attention. Cureton's (1950) chapter on "Validity" in the first edition of Educational Measurement apart from reflecting on aspects of relevance and reliability also recognized the

existence of content validity. A new way of conceptualizing validity had emerged with this chapter.

2.2.1 Categories of Validity

Another important development of the 1950s was the recognition of four categories of validity: predictive-, status- (concurrent), content- and congruent validity (Sireci, 1998b). And Sireci further noted that this change was reflected in the Technical Recommendations for Psychological Tests and Diagnostic Tests of 1952 and 1954 Standards for Educational and Psychological Testing (APA, 1952; APA, AREA, & NCME, 1954). This was a significant step for the development of content validity because it was distinguished as the type of evidence that demonstrates how well a test taps a particular domain area without necessarily showing the usefulness of a test for any purpose. The 1952 Technical Recommendation for Psychological Tests and Diagnostic Tests as quoted by Sireci, (1998b, p. 88) recognized this special role and stated:

> Content validity refers to the case in which the specific type of behavior called for is the goal of training or some similar activity. Ordinarily, the test will sample from a universe of possible behaviors. An academic achievement test is most often examined for content validity. (p. 468)

This recognition of content validity marked a departure from the domination of statistical validation studies that used correlation of the predictor and the criterion as validity evidence. According to Angoff (1988) predictive and concurrent validity were also introduced as types of validity in the same 1952 Technical Manual and these retained the need for statistical evidence. Yet, evaluation of content validity would be

accomplished by judging how well the content of the test sampled the class of situations in a domain of reference. With this development, test validation began to focus on a test quality rather than its relationship with some other measure as previously conceptualized. Content validity was therefore, deemed particularly important for achievement and proficiency measures and its responsibility begun to shift to the test constructor. It became a quality that should be ensured as a test is being constructed. Emphasis on the need of a table of test specifications prior to test construction is one way of ensuring content validity.

2.2.2 A Shift Toward a Unified Concept of Construct Validity

According to Sireci (1998b), another shift in favor of construct validity emerged in 1955 with the presentation of a paper by Cronbach and Meehl (1955). From then on, the conceptualization of validity began to move towards the unitary concept and the definition of validity began to reflect this shift in thinking. This conceptualization of validity did not only exalt the place of construct validity but also posed new insight on the definition of content validity. For example, Lennon (1956) defined content validity in terms of how well the sample of responses of an examinee to the items of a test represented the real or hypothetical universe of situations of interest to the person interpreting the test scores. In this definition, responses rather than only items, were featured in the definition of content validity and therefore the interaction between the test content and examinee responses was to become the focus of content validation.

According to Sireci (1998b), Lennon is reputed to have provided a justification for the use of content validity, which was later supported by Ebel (1956) who argued

that content validity was the foundation of construct validity. But Sireci (1998b) recognized that it was Loevinger (1957) who made a major shift toward a unitary concept of construct validity when she developed the concept of substantive validity to incorporate the concerns of content validity in the construct validity framework. The shift into a unified concept was then adopted in the Standards (APA, AERA, & NCME, 1966; 1974), and further expounded by Cronbach (1971). But Messick (1988; 1989) forcefully suggested elimination of aspects of validity such as content- and criterion-related in favor of construct validity. Messick (1988) argued,

> Typically, content-related inferences are inseparable from construct-related inferences. What is judged to be relevant and representative of the domain is not the surface content of test items or tasks, but the knowledge, skill or other pertinent attributes measured by the items or tasks. (p. 38)

However, the elimination of content validity has not succeeded because the evidence of item representation to the domain of interest is important in determining the validity of any inference resulting from test score interpretation.

Today, the most acceptable meaning of validity refers to an evaluation of the adequacy and appropriateness of the interpretations and uses of assessment results (Linn, 2000; Messick, 1989). According to the Standards (AERA, APA, NCME, 1999) the concept of validity refers 'to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of a test' (p. 9). In general, current definitions of validity emphasize score interpretation and use as the basis of validity evaluation. This shifts test validation from focusing on a test and scores to meaning of scores in the context of their use. According to Cronbach (1971), the process of validation refers to collection of evidence to support the inferences that are

drawn from test scores. For the Standards (AERA et. al., 1999) the process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation. Although the idea of a unitary concept in favor of construct validity has been pushed far enough, scholars are in agreement that all different sources of validity evidence should be tapped to fully justify construct validity interpretation of test scores. In this respect, content validity must be ensured prior to justification of construct validity.

2.3 Prominent Role of Content Validity as a Test Quality

Evidence about the relevance and representation of test items to the domain of content is an indispensable contribution of a test validation process. Such a contribution is meaningful when one evaluates the usefulness of a test with respect to what the test measures and the purposes it serves. Therefore, whether our concern is about the construct, test purpose, or degree of domain representation, we shall find content validity evidence useful. As it was pointed out earlier, content validity with its emphasis on the congruence of items to content domain cannot be ignored in achievement testing where scores denote mastery of the content being measured.

Although Messick, (1988, 1989) dismissed content validity as a form of validity, he acknowledged the importance of evidence relating to adequacy of domain coverage and representativeness of items or tasks. This researcher maintains the view that there can be no construct validity if content validity of a test is not ensured because full interpretation of test scores depends on how well the test represents what it measures. This view is shared by Sireci (1998b) who stated "Obviously, we cannot, and should

not, evaluate test scores without first verifying the quality and appropriateness of the tasks and stimuli from which the scores were derived" (p. 103). Messick (1989) recognized construct underrepresentation as one threat to construct interpretation. This makes content validation important because it is the sure way of checking on adequacy of construct representation. Therefore, whether content validity is derived to meet the requirements of construct validity or in its own right, both the test developer and user should strive to satisfy it, especially in achievement tests.

2.3.1 Defining Content Validity

Content validity refers to the extent to which test items represent the tasks, skills or knowledge of a domain of interest as described by course objectives. In case of achievement tests, such objectives describe the limits of the breadth and depth of the content domain and become the guide for instruction and assessment. Tests are samples of possible sets of items that can be drawn from the domains they purport to measure. Task or sample representativeness to the domain of interest is, especially, important for high-stakes assessments because such tests tend to emphasize a philosophy of realism with the belief that exactness and precision are inherent in measurements (Ediger, 2002). In view of this, most high-stakes assessments base their critical decisions such as promotion, selection, or certification on scores from a single test without regard to accumulated students' achievement reports obtained in the course of learning. It is therefore, imperative that the tests, which provide such scores, should regularly be evaluated for content validity. Otherwise, if content representation is lacking, such tests can undermine students' achievement and lead to serious misinterpretations of their

achievements. It should be noted that score interpretation always has its basis on the content domain of interest for which the score may denote mastery or non-mastery.

Lennon (1956) made similar observation and stated:

> We propose in this paper to use the term content validity in the sense in which we believe it is intended in the APA Test Standards, namely to denote the extent to which a subjects' responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothetical universe of situations which together constitute the area of concern to the person interpreting the test. (p. 91)

For Lennon, it was not enough to appraise content validity by taking into account only the questions in a test but to include the responses because content validity inheres in the responses too. This was seen to be consistent with the test user's intention of making inferences about behaviors, which are best displayed by the responses of examinees to items.

For most high-stakes achievement tests, score interpretation leads to classification of students into passers or failures and sometimes those who pass are further classified into subcategories such as proficient, advanced, credit, or distinction. In some cases, schools are held accountable to improve instruction, student learning, grade promotion, and certification on the basis of such testing (Ryan, 2002). Therefore, conducting regular content validity studies can help both test developers and users to make more defensible decisions regarding the articulated effects of high-stakes testing than when content validity evidence is lacking. When scores assume prominence in classifying people into masters and non-masters, content validity evidence can support the accuracy of the decisions that are based on scores so that people are not erroneously classified (Hambleton, 1984). Lennon (1956) concluded,

> When a test is intended to serve as the basis for determining how well an individual would perform at the present time on some universe of situations, and when circumstances do not permit measurement of responses of the entire universe, then it is clear that the test user must rely upon correspondence between test sample and universe in deciding what confidence he may have on the inferences made on the basis of the test. (p. 297)

2.3.2 Concepts Related to Content Validity

Other important concepts encountered in validity theory and can sometimes be confused with content validity are: face validity, curricular validity and instructional validity. According to Crocker and Algina (1986), face validity refers to the extent to which items on a test appear to measure the constructs that are meaningful to typical examinees. In achievement testing that follow instruction, face validity is important to gain students' confidence in the test's fairness or to gain examinee voluntary cooperation when required (Thorndike, 1997). But Angoff (1988) described face validity as appearance of validity and he noted that although the term emerged as early as the 1940s, it has not received much technical interest.

With respect to curricular validity, the term has been used to describe how well the items on a test represent the curricular objectives of an institution. According to Yallow and Popham (1983), curricular validity is the correspondence between a test and the objectives of the curriculum. In the case of achievement examinations, there seem to be some relationship between the curricular and content validity. Quite often, the objectives of the curriculum describe the limits of the content domain and test specifications or blueprints are drawn from curricular objectives that describe the content. Despite a common basis for both content validity and curricular validity, the two terms cannot be used interchangeably because they have different emphasis.

Curricular validity deals with relevance of the test to objectives of an institution's curriculum while content validity is about congruence of items with the specified domain objectives that are often presented in test specifications. But the evidence for both is derived from objectives of the same curriculum on which achievement test is focused. Therefore, content validity overlaps with curricular validity with regard to domain representativeness. Messick (1989) contended that curricular validity should be called curricular relevance, or the judged fit of the test to the immediate objectives of the curriculum. But he also noted that some scholars regard it as insufficient for claiming test relevance when instructional experience is not ensured.

This argument leads us to instructional validity, another term that has sometimes been confused with content validity. Yallow and Popham (1983) described instructional validity as " the correspondence between the test and instruction" (p.12). It demonstrates how well the instruction has prepared examinees for the test that measures the domain on which the examinees were instructed. Because of its focus on how examinees are prepared for a test, Yallow and Popham (1983) further argued that it should be called adequacy-of-preparation rather than instructional validity. As such, instructional validity refers to how the use of the test justifies the preparation of the examinees in the content being measured. Again, Messick (1989) dismissed instructional relevance as a form of validity but admits that it refers to how well the test items represent the immediate objectives on which examinees were actually instructed in. The concern on instruction should be a reminder to high-stakes test developers to ensure that such assessments should focus on what students have had an opportunity to

learn. Otherwise, the decisions that are made on the basis of the scores of such tests cannot be justified.

Instructional relevance is therefore, different from content validity, which makes inferences about the domain of interest and has nothing to do with the way the examinees were prepared for the test. It is true that in achievement tests, and especially in high-stakes examinations, policy makers worry about examinee preparation as much as they worry about item-objective congruence because of the way scores are interpreted. According to Ediger (2002) meaningful testing ensures assessing what pupils have had an opportunity to learn, as listed in their objectives of instruction that teachers used. But Yallow and Popham (1983) argued that adequacy of preparation is not necessary for one to make sensible inferences about what scores signify. Inadequacy of examinee preparation cannot make a test that has adequately sampled the domain to be measured invalid. However, it can also be argued that ensuring content validity provides a basis for checking adequacy of preparation, which when ignored, can amplify negative consequences of a test.

Regarding consequences of testing, Cronbach (1988) in response to Yallow and Popham stated: "You may wish to exclude reflections on consequences from the meanings of the word, 'validation', but you cannot deny the obligation" (p. 6). The concern of test consequences and its obligation on validators has led to the emergence of a new term of "consequential validity", which has joined the old validity terminology debate. Messick (1989) espoused the importance of test consequences when he advocated their inclusion in the validation process with respect to test use and score interpretation. He felt that consequential validity referred to the latent, sometimes

undesirable, and unplanned effects of an examination. For example, in Malawi the use of a single examination to select students for higher educational opportunities has led to rampant cheating in examinations (Chakwera, et al., in press; Malunga, 2000).

Quite often, in high-stakes testing, the negative consequences of examinations are cited by the critics of testing with emphasis on perceived damages on individuals, schools, and the education system as a whole. Steeves et al. (2002) observed that high-stakes testing has led to teaching to the test, which has consequently made teachers depend on tests for making instructional decisions. It is therefore, argued that apart from narrowing the curriculum, teaching to the test has in the long run rendered the test to be an object of instruction rather than the outcome of previous instruction because teachers spend more time on drill and practice than on hands-on curricular activities. A similar effect is observed with respect to assessment of Agriculture in Malawi where practical aspects of the course are no longer emphasized in instruction because the MSCE examination in Agriculture no longer assesses projects, as was previously the case (Chakwera, et al., in press). In recognition of the effect of assessment in the implementation of a curriculum, Wiggins, as quoted by Miller and Linn (2000) felt that reforms in education should as well be made through assessments. Miller and Linn further agued that high-stakes assessments are often implemented with the assumption that teaching to the test will be one of the positive effects of the assessment program.

However, in recognition of some negative consequences of testing, the Standards for Educational and Psychological Testing (AREA et al., 1999) state that evidence about consequences is relevant to validity when it can be traced that a source of invalidity is construct underrepresentation or construct irrelevant components. Lane

and Stone (2002) provided an example for gathering evidence of test consequences and they stated:

> When a state assessment and accountability program is used to improve educational practice by influencing curriculum and instruction and to hold Schools' accountable, evidence for the following propositions might be deemed necessary:
> 1. School administrators and teachers are motivated to adapt the instruction and curriculum to the standards.
> 2. Professional support is being provided.
> 3. Instruction and curriculum will be adapted.
> 4. Students are motivated to learn as well as to put forth their best effort on assessment
> 5. Improved performance is related to changes in instruction. (p. 26)

Obviously, such an evaluation process may need validators of a test to seek evidence beyond the traditional methods of focusing on item adequacy and meaning of scores. For this reason, Crocker (2002) and Ryan (2002) recommend the inclusion of stakeholders such as administrators, students, teachers, businessmen or prospective employers to provide information regarding consequential validity of a high-stakes examination. And Frye et al. (2002) reported encouraging results from a curricular validation study, which involved a number of stakeholders.

2.3.3 How Content Validity Relates to the Construct Validity

The relationship between content validity and constructs being measured has been the basis for the arguments toward a unitary concept of validity. Construct validity refers to the extent to which test items measure the psychological traits that represent the domain of interest. Advocates of a unitary concept of validity (Anastasi, 1982; Cronbach, 1971 & 1984; Messick, 1988, 1989) contend that the focus on the interpretation or meaning of test scores puts emphasis on construct validity and such an

understanding reduces the other validity types into sub-components of the same thing. However, Tenopyr (1977) as quoted by Sireci (1998b), argued that "construct and content validity cannot be equated .....Content validity deals with inferences about test construction: construct validity involves inferences about test scores" (p. 50). This argument helps to illuminate the role of content validity in relation to construct validity. In view of this understanding, Sireci and Geisinger (1995) advocated that test developers should ensure content validity by adequately defining and specifying content domain, drawing up test blueprints that adequately represent the content specifications, and constructing test items that adequately represent the test blueprint. Careful test development can ensure test quality and enable scores to derive meaningful inferences about examinees with respect to the construct being measured. Interpretations of examinees' test performance are meaningless when we cannot describe and establish congruence of items to specifications (Popham, 1984). It is therefore, reasonable to argue that content validity should be the first quality that test developers should satisfy in a test.

While acknowledging the existence of an embracing relationship between content and construct validity, whereby the latter is inclusive of the former, the two validity aspects deserve separate validation too. In terms of gathering validity evidence, it should be noted that construct validation tends to be more involving and requires more rigor as provided in generalizability framework (Brennan, 2000; Frye, Szauter, Litwins, & Ofoegbu, 2002) than content validation. But it can also be argued that construct validity is incomplete without evidence of content validity. Ensuring representation of the construct being measured is the ultimate role of content validity.

Therefore, it should be a matter of common sense that inferences about test scores cannot be adequately expressed unless the instrument from which the scores are derived is valid. According to Sireci (1998b), the word 'validity' in content validity refers to credibility, or the soundness of assessment instrument itself for measuring the construct of interest. Therefore, ensuring content validity of an instrument is a prerequisite to score based evaluations such as construct validity. But while construct validity requires supporting evidence from content validity, it is misleading to contend that content validity's inability to provide exhaustive evidence for a test reduces it to something less than validity. Ensuring item and domain congruence and appropriateness in task or skill sampling should be a desirable professional obligation for test developers and users. Therefore, content validity needs to be regularly carried out to ensure such a noble objective even when resources are not available for a fully fledged construct validation study.

By examining different terms relating to test validity, this section has tried to highlight the importance of content validity amid the controversy that has questioned its existence over the years. In general, scholars agree on both the existence and importance of content- related evidence in testing. But those who advocate the elimination of content validity from the validity concept are doing so on grounds that it does not adequately cover the issues of score interpretation and use, which they contend should be the focus of test validation. It is clear from these arguments that item representativeness of the domain of interest, or item and domain congruence, cannot be ignored in achievement testing. In high-stakes testing there would be no justification for the inferences derived from test scores unless the content validity of the test that

produces such scores is established. It is therefore, concluded that content validity remains an important quality to be satisfied in testing, especially when test scores are believed to represent student achievement. Content underrepresentation is a serious threat to any type of validity evidence that could be of interest in an evaluation of an instrument.

2.4. Content Validation Procedures

There are several procedures that have been used to assess content validity of a test. Sireci and Geisinger (1992) classified these methods into subjective, empirical, and pseudo-empirical.

2.4.1 Subjective Methods

This category represents the most traditional approach to content validity that essentially uses subject matter experts (SMEs) to give judgments about test item relevance to the content domain being assessed. The most typical index of this method is the SMEs' average proportion of item ratings, which indicates the proportion of items that assess the objectives of a local curriculum or frameworks. This index can be computed when SMEs make dichotomous decisions about how each item match the listed objectives. The formula that is used for this index is:

$$p = \frac{\sum_{j=1}^{J} n_j}{NJ} \tag{2.1}$$

where n_j is the number of items matched to the correct objective by judge j,

N is the number of items on the test, and

J is the number of judges. An alternate to this formula is:

$$p = \frac{\sum_{i=1}^{N} n_i}{NJ} \tag{2.2}$$

where n_i is the number of judges who rated item i as a match to its objective (Crocker, Miller & Franks, 1989, pp. 180-181).

When interpreting this index it is important to remember that:

a) It is sensitive to the total number of items and becomes more stable when number is not less than 30 items;

b) It does not consider the weight of the particular objectives being measured by the items;

c) It does not consider the degree of certainty with which the judges matched the items; and

d) It is sensitive to the judges' ratings depending on the phrasing of the validity question.

Item-Objective-Congruence Index and its variation (Hambleton, 1980; 1984) are also used to ascertain content validity. This index is appropriate for assessing content validity of criterion-referenced tests and it is symbolized as:

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} - \sum_{i=1}^{N}\sum_{j=1}^{n} X_{ijk} + \sum_{j=1}^{n} X_{ijk}}{2(N-1)n} \tag{2.3}$$

where:

I_{ik} is the index of item-objective congruence for item k on objective i,

N is the number of objectives (i = 1, 2, . . ., N),

n is the number of content specialists (j = 1, 2, . . .,n), and

X_{ijk} is the rating (-1, 0, +1) of item k as a measure of objective i by content specialist j (Hambleton, 1984, p. 209).

With this index, each item is assessed on how well it matches with the objectives that were specified in the test blueprint. In so doing, the congruence of items and objectives is determined and content validity of the test established.

Sometimes content validity index (Lawshe, 1975) is computed to determine how the test fits the content domain. This ratio is appropriate for determining content validity of a job performance domain. It is symbolized as:

$$CVR = \frac{n_e - (N/2)}{(N/2)} \tag{2.4}$$

where n_e is the number of panelist rating the item as essential and N is the total number of panelists. The CVR values range from –1.00 to 1.00, and whenever fewer than half the panelists rate an item as essential, its CVR value is negative (Crocker, Miller & Franks, 1989 p. 186). As a percentage index, the CVR is a linear transformation of the proportion of the judges who rate the item as essential. It is sensitive to number of items and degree of certainty with which the judges match an item to an objective.

Another statistic, which involves SMEs' rating of items, is the Aiken Validity Index (Aiken, 1980). This content validity index is symbolically expressed as:

$$V_i = \frac{\sum_{i=1}^{c-1} in_i}{N(c-1)} \tag{2.5}$$

where v_i is the content validity index for item i,

c is the number of categories of the scale used to rate the item,

i is a value or weight assigned to categories (lower category is assigned a 0, the next lowest category receives a weight of i, value of 1; and so forth)

n_i is the number of judges who rate an item into the ith category (0, 1, 2 . . . c – 1)

N is the total number of judges (Crocker, Miller & Franks, 1989 p. 186)

The values of this validity index range from 0 to 1 and the index is different from CVR and item-objective congruence indices whose values range from –1.00 to + 1.00. In addition, the Aiken validity index has an accompanying method for calculating the probability of obtaining that particular distribution of judges' rating by chance. This probability is symbolically expressed as:

$$p = \frac{N!/c^N}{n_0!n_1!...n_{c-1}!} \tag{2.6}$$

With increased number of categories the calculation of probability becomes complex. The normal approximation to p is therefore, used instead and it is symbolized as:

$$z = \frac{N(c-1)(2V-1)-1}{\sqrt{N(c-1)(c+1)/3}} \tag{2.7}$$

The computed z value is contrasted to the critical value of z from a standard normal z table at a specified level of alpha. The test of significance is important to check on chance influence on the final allocations of the items to domain areas when many raters are used.

According to Crocker, Miller and Franks (1989) the Klein and Kosecoff's Correlation is also used as an index of relationship between the importance of an objective and the proportion of the judges who match an item to that objective. The sum of these proportions across items that measure an objective is denoted $P_k = \sum pik$, and represents the total number of items that measure a particular objective. And the Pearson Product Moment correlation (r) is then computed between the mean importance weighting and the P_k values for the objective to obtain an overall index of the degree of fit. This index is sensitive to the number of objectives, number of judges, variations in number of items assessing each objective, importance rating variations and specification of instructions to judges.

When the degree of fit between the test and the curriculum is required, the index of relevance suggested by Morris and Fitz-Gibbon (1978) as quoted by Crocker, Miller and Franks (1989) is useful. To obtain this index, SMEs are expected to perform the following activities: a) identify the objectives that match each item, b) rate the importance of the objective, c) determine suitability of the item format, and d) establish the appropriateness or estimated level of difficulty of the item for local student population. Once these ratings have been determined, the test selection decision can be guided by three separate indices: Index of Coverage, Index of relevance and Grand average.

As it has been observed in this section, all these approaches to content validity greatly depend on the subjective judgments from SMEs who rate items in relation to provisions of test blueprints. These procedures have been criticized for exposing test blueprints or test specifications, a practice that can bias the evaluators toward test

developers' inclinations on item/domain distribution. Cronbach (1988) and Ryan (2002) warned test evaluators to guard against "confirmationist bias," which is a tendency to look for supporting evidence in the validation of assessments with regard to interpretation and uses. In their criticism of the subjective methods of validation, Sireci and Geisinger (1995) picked out the exposure of the test blueprints as the major weakness of subjective procedures in content validation because such a practice prevents the SMEs from revealing their unique perceptions of the underlying content structure.

2.4.2 Empirical Methods

Validation approaches in this category seek empirical techniques to analyze item response data to discover underlying content response structure. According to Sireci and Geisinger, (1992) advocates of these methods apply multidimensional scaling, factor analysis and cluster analysis (Napior, 1972; Oltman, Sticker, & Barrows, 1990) to get dimensions, factors and clusters presumed relevant to content domain. The use of test response data is consistent with the definition of validity with respect to inferences about test scores and leans more toward construct validation than content validation. But Sireci and Geisinger (1992) observed that the dependence on item data to derive test validity evidence is the major weakness of the empirical methods. Analyses in these validation approaches tend to ignore inherent item characteristics that are central to content validity. These methods concentrate on irrelevant factors to content validation procedures such as item difficulty, ability level, examinee variations, motivation, and guessing. For this reason, empirical methods do not seem to adequately assess content

validity as the subjective approaches do. However, the need for empirical analyses in content validation studies cannot be overemphasized in the field of psychometrics in which measurements count. Deville (1996) observed that for a long time psychometricians longed for "empirical techniques that would link data from ratings provided by subject-matter experts (SMEs) with data from actual responses to a test/instrument. Such a technique would link content and construct evidence" (p. 127). Depending on the validity question that is being investigated the use on both examinee scores and SMEs' rating may call for techniques that can meaningfully analyze such data to establish validity evidence.

2.4.3 Pseudo-Empirical Methods

In an attempt to merge the two approaches discussed above, the pseudo-empirical methodology uses SMEs' judgments, but ignores item response data. Once the SMEs' judgments are obtained, Sireci and Geisinger (1992) observed that analyses would proceed with factor analysis (Tucker, 1961), MDS, and Cluster analysis to obtain factors, dimensions and clusters that describe SMEs' perception of the content domain. In such an approach, the MDS analysis of item-similarity data provides unique information beyond that gathered by the traditional content validity studies (Sireci, 1998a). Strong evidence of content validity is expected when both traditional ratings of item relevance or congruence and item-similarity ratings are included in content validation study.

The approach of gathering and analyzing content validity data in this study pretty much followed the pseudo-empirical methodology. The duplicate test

construction experiment that this study examined used both examinee scores and SMEs' ratings to establish content validity evidence for the two tests. The test scores were used to determine how the mean difference squared compared with the sum of error variances of the half tests of each of the common mock examination forms. In another development, Crocker, Llabre and Miller (1988) considered content validity under the framework of the generalizability (G) theory because of multifaceted nature of sources of information leading to such evidence. It is noted that despite the presence of multiple ratings of items' fit in content validation process, none of the procedures discussed earlier in this paper allows for systematic isolation of variance sources that influence content experts' rating. Therefore, G-studies are seen to be the way out of this problem.

This section has discussed procedures for conducting content validity studies. The data collection methodology is essentially linked to subjective involvement of SMEs. However, the analysis process with several statistical techniques helps to improve on the validity of the approaches in producing findings that can be empirically defensible. The analyses conducted with the pseudo-empirical methods were considered appropriate for this study because neither test specifications, nor blueprints are made available prior to test construction assignment in Malawi. Therefore, item-objective or topic relevance ratings remain the most important data that can be used to derive evidence for the content validity evaluation.

2.5 Content Validity Studies In Malawi

Chulu and Sireci (2002) evaluated the content validity of the MSCE Physical Science examinations. The results indicated that most of the content areas were well

represented while a few areas were underrepresented. These results contradicted earlier findings that demonstrated lack of content validity in MSCE Physical Science and failure for the Papers 1B and II in the same subject to complement each other (Mwanza, 1998). However, the 1998 study reported that although the papers were lacking in content representation, the items on the papers were well spread in terms of cognitive abilities. Earlier studies include Mbunge (1986) and Ndalama (1986) who analyzed content validity for Junior Certificate Geography and History, and Agriculture examinations, respectively. Both studies found that the content coverage of the examinations was generally satisfactory although some topics of the content were consistently underrepresented over the years.

All the studies conducted in Malawi have used the traditional or subjective method of content validation because they aimed at obtaining an index of item-objective congruence. As noted earlier, such analyses are prone to "confirmatory bias" because their data collection exposes the test developers' item selection criteria to the SMEs. This may result in judgments that are the same as those expected by the test developers because the exposure of the test developer's criteria for item selection makes it easy for the SMEs to deduce what is expected of them. In so doing, the test may not be independently evaluated because the raters may indirectly be influenced by the test developers' decisions as reflected in the test blueprint.

2.6 How the Duplicate Test Construction Relates to Content Validation

Without undermining the problem of "confirmatory bias", the current study used SMEs' ratings to complement the results of dependent samples t-test derived from

scores from the two tests administered to the same examinee under the duplicative test construction experiment (Cronbach, 1971). In addition, this approach recognized the degree of correlation between two independently constructed tests that are based on the same frameworks and objectives as well as item-objective congruence indexes as measures of content validity. Through the duplicate test construction approach the study provided insights about how well teachers and examiners understand the test content domain as described in the frameworks. In the Malawi context, such an approach to content validity was appropriate because MANEB does not use readily prepared test specifications or blueprints when developing tests. Instead, each test construction exercise begins with the development of test specification grid, which is used to determine the items that are included in the test. By allowing each team of test constructors to develop its own table of test specifications, it is assumed that test developers have common understanding of the content domain and that equivalent test specifications and test forms can be developed from the syllabus given the same test construction guidelines. The duplicative test construction experiment was suitable to examine the basis for such assumptions.

For achievement tests such as MSCE, the duplicate test construction experiment would provide the most evidence for MSCE examination validity. It would answer questions, which Rulon (1946) considered critical when he wrote:

> We have, then, two general ways of asking about validity of achievement test, one growing from the other. The first is, Are the materials of this test, and the processes called for on these materials, the same things and processes we are trying to teach the children? The other is, If they are not, do we have evidence that scores on this test go hand in hand with those we would obtain with a test about which the answers to our first questions were in affirmative. Thus we see that the direct observation of the

> things and processes, which are the aims of instruction, is the final proof of validity, as compared to the correlation coefficient of validity, which is at best secondary. (p. 292)

The experiment provided an opportunity to check the relationship between instruction and assessment because teachers were involved in designing the test. It was also possible to examine the correlation of the scores from the two forms of the mock examination as well as with the MSCE 1A because same examinees took all the three tests.

This chapter has discussed the literature on validity with an emphasis on content-related validity. The literature review is not by any means exhaustive because the topic of validity is very wide and cannot comprehensively be treated in a single chapter because of the diversity of ideas around it. As Kane (2002) observed, the evidence required for validation depends on the proposed interpretation, and it is entirely possible for one or more of these interpretations to be valid, while other interpretations are invalid. Similarly, in treating validity literature one has to tow a line of thought that is consistent with the investigation to be carried out. Further, reference to literature is made where appropriate in other parts of this report. The next chapter discusses the methodology that was employed to collect and analyze the data of this study.

CHAPTER 3

METHODOLOGY

Consistent with guidelines for duplicate test construction experiment as provided by Cronbach (1971), two teams of teachers drawn from six schools in Zomba district were formed. The study used both test scores and survey information from different sources to answer the questions that were investigated. This chapter describes the methodology for the sample selection in section 3.1; descriptions of data collection instruments are in sections 3.2 and 3.3: data collection procedures are described in section 3.4; and finally, data analysis procedures are outlined in section 3.5. An attempt to link the data and analysis to a particular research question has also been made.

3.1. Selection of the Sample of Participants

This study involved selected schools, teachers, and students of History at the MSCE level. Details of the schools and each participating group are described in this section.

3.1.1 Description of the Participating Schools

Six secondary schools from the Zomba district were selected for the study. This was a convenient sample in terms of proximity of the schools to each other, which made their access easy and also ensured adequate supervision of the experimental process. Despite being convenient, the sample was considered representative in terms of school size and type, teacher qualification, and teacher-student ratio. All the schools

were mixed boarding except one, which was a mixed day school. Two of the schools were private while four were public. Teachers who were teaching History in Form 4 in the six schools participated in the experiment. While the teachers were involved in the test construction, review, and scoring the tests, examinees who took the test as mock examination became the unit of analysis for this study because correlation of their paired scores and the mean difference were the major focus of the evaluation. By involving the six schools a cluster sample of 200 students participated in the study. This number of participants was considered large enough to ensure stable statistics for the analyses whose discussions and conclusions have been reported in the study. The report has not mentioned any school by name or description because the study focused on group data rather than school comparisons.

While a validity study of this nature would be better conducted on a randomly selected sample of examinees to maximize on the accuracy of the statistics (Henryson, 1971), a cluster sample that included whole classes in each school was preferred to capitalize on teacher involvement. The duplicate tests were administered in the same schools of the teachers who constructed the tests so that the teachers could consider the tests as part of their normal class assessments. In so doing, it was expected that the students who would earn a grade out of the tests would take the exercise seriously. The teachers administered and scored the tests just as they always did with ordinary mock examinations.

3.1.2 Description of Participants

There were several groups of participants who took part at different stages and on specific activities of the study. The first group comprised teachers of History drawn from the six schools that were selected for the duplicate test construction experiment. These teachers had taught MSCE History for not less than three years. Because the experiment targeted form 4, which is an examination class, there were differences regarding the length of time each teacher had been with the same class. Some had taught the classes from the time the students were in form 3 while others took over the classes in form 4. For this reason, there were variations in the manner they commented on students' preparation. Despite these differences, they were all qualified teachers of History with either diploma or degree that had a History background. In Malawi, teachers qualify to teach in secondary school when they either have a diploma, with a minimum of three years or a degree, with a minimum of four years of college/ university education.

The second group of participants consisted of students who took the Mock and MSCE examinations in the six participating schools. These students were 221 and they were all in their final year of secondary education. In terms of age, they were between 17 and 21 years old. They also had varied backgrounds. Some of them were repeating their classes while others were not and yet others had changed schools within the two years of instruction on the MSCE History syllabus. As a result of this, their responses about their levels of preparation were sometimes different even when they were in the same school.

The final group of participants comprised 15 SMEs who were engaged as judges to rate the content and cognitive relevance of the items in the tests. They were purposefully selected, through Education Division Offices, from the population of experts in the History subject matter. Some were teachers of History while others were involved in education but had taught History for not less than three years before. The SMEs were drawn from five of the six divisions and there were 9 men and 6 women. Only those SMEs who were not engaged by MANEB in setting, reviewing or scoring national examinations, were selected. Such a selection criterion was deemed necessary to avoid prejudiced judgments resulting from familiarity with certain items or the test developers. In addition to this criterion, judges were also selected to ensure geographical and gender representation. This report used ratings of 14 SMEs only because one of them dropped out.

In addition, there were two trainers that were used during the duplicate test construction experiment. One was a test development officer from MANEB and the other was a test reviewer who also supervises History scoring exercise for MANEB examinations. Both of these experts had worked for MANEB in those roles for more than 3 years. The two trainers were engaged in test construction and item review training that the teacher participants went through at the beginning of the duplicate test construction experiment. During the test review exercise both trainers worked as desk leaders rather than participants. Their major role was to provide technical advice on test construction process as the exercise progressed. Subject matter issues were left to the participants so that the belief systems of the trainers could not influence the content

decisions that the teams made. The two trainers kept rotating between the teams so that their expertise could equally benefit the two teams.

3.2. Data Collection Instruments (Curriculum-Based Tests)

This section describes four tests: MSCE History Paper 1A, Mock A, Mock B and Class Tests that were used to collect data for the study. These examinations were all based on the MSCE History syllabus.

3.2.1 The MSCE History 1A Examination

This test is the multiple-choice component of the MSCE History examination and it comprises 60 items. Trained test developers that MANEB engages every year draw it up. This study used the 2003 examination to obtain examinee score and SMEs' content and cognitive skill relevance ratings. Because MANEB could not provide a Table of Test Specifications for the 2003 MSCE History 1A, the researcher constructed a blueprint from the test to facilitate interpretation of SMEs ratings and comparisons with the other tests.

In a typical MSCE History examination, the second component, MSCE History 1B comprises short-answer and essay questions. Each of the questions carries 20 score-points. The short answer items are often in sets of questions focusing on a topic or two while the essay item quite often has one question or two complementary questions focusing on the same topic. The two components (MSCE 1A and 1B) are considered both complimentary and independent in terms of syllabus coverage and cognitive skill representation. They are often administered on the same day to save on students'

preparation time. Both components draw half of the items from each of the two broad sections of the MSCE History syllabus, the Central Africa and World History Since 1900. The final grade in the subject is determined by translating the composite raw score from the two components into a number grade using a 9-point scale, which was described in chapter 1.

3.2.2 Common Mock History Examination

The common mock examination consisted of 120 multiple-choice items that were developed by the two teams that participated in the duplicate test construction experiment. Part A of the common mock examination comprised sixty items from Team A and Part B contained the other 60 items from team B. As is the case with the MSCE History 1A, each part had one half of the items from each of the two broad content domain sections of the MSCE syllabus. The two parts were deliberately made distinct to enable easy collection of students' ratings of their impressions on the quality of each subtest. In addition, combining the two tests fulfilled a practical need, as students did not have to write the same examination for the second time. These practical considerations outweighed problems of practice and item positioning effects, which otherwise could have affected the results due to this approach.

The items were developed after training that was aimed at standardizing the skills of the teacher participants in test planning, multiple-choice item construction, editing and reviewing. After the first day's training, the six teacher participants were divided into two teams that were independently instructed to draw up test specification tables and develop a 60-item multiple–choice test based on the MSCE History syllabus.

The team members used their test specification tables to guide their allocation of the item construction tasks to individuals. By using the topics of the syllabus, the subject matter to be tested was defined while the cognitive abilities proposed by Bloom et al. (1956) were used to specify abilities to be tested. They continued to work in separate teams until the final forms of the tests were built.

3.2.3 Class Tests

Sixty multiple-choice and short answer items were compiled from various schools to provide samples of items that teachers developed prior to their participation in the duplicate test construction. Each teacher was asked to give a copy of items from the latest terminal test and the researcher compiled the items into a set of sixty items similar to the mock subtests and MSCE 1A. To accommodate the different formats that were used and ensure consistency with the common mock items only multiple choice and short-answer items were selected. This set was meant to demonstrate the topics that were favored in classroom assessment and the type and quality of the items. Therefore, a selection of items from various topics was made to build the test.

3.2.4 Summary of the MSCE History Tests Used in the Study

Three tests, Mock A, Mock B, and MSCE 1A were used to collect scores on student performance on MSCE History. But for the purposes of item relevance ratings, their original format was concealed to avoid prejudice and bias that would result from SMEs' familiarity with the tests' backgrounds. The items were then given to SMEs as Sets 1 and 2 for the duplicate tests A and B and sets 3 and 4 for the class test and MSCE

1A, respectively. Copies of these sets are attached to this report as Appendices A to D. The four tests described in this section can be summarized as in Table 3.1 as follows:

Table 3. 1. Summary Description of Test Instruments Developed From MSCE History Syllabus

Test	# Items	How Developed	How Administered	When?
MSCE 1A 2003	60	MANEB	Separate component	October,
Mock A	60	Team A of Duplicate Test Construction Experiment	Subset of Common Mock Examination	July, 2003
Mock B	60	Team B of Duplicate Test construction Experiment	Subset Common Mock Examination	July, 2003
Class Tests	60	Selected from various Class Tests	Not administered	

3.3 Data Collection Instruments (Surveys)

A number of questionnaires were developed to collect information on various areas of assessment practices. Such areas included teachers' instructional decisions about the content they selected to teach in preparation for the 2003 MSCE examination; teachers' opinion about their involvement in the test development activities of the study; pupils' opinions about adequacy of their preparation and quality of the tests they took; and finally, SMEs' opinions about the quality of the tests they judged. Descriptions of the surveys are as follows:

3.3.1 Teacher-Participants' Surveys

The first teacher-participant's survey aimed at collecting information about how the teachers prepared their students for the 2003 MSCE History examination. This was accomplished by a questionnaire that focused on how the teachers selected topics that they taught and how they covered those topics in view of the importance that they attached to the topics. The questionnaire also asked the teachers to indicate the topics on which they assessed their students during instruction. One specific objective of this survey was to gather evidence about the impact of national examinations on classroom instruction and assessment practices. The questionnaire was completed prior to the test construction training exercise so that the teachers would provide information that would not be biased by their knowledge of the duplicate test construction objectives or their interaction with other team members.

Responses on this questionnaire provided information regarding the belief systems that individual teachers used in determining the topics that they taught and on which they assessed their students over the period of instruction. On the same questionnaire, teachers were also asked to indicate the reasons for their choice of topics and the emphasis they put on each topic in the course of instruction. It was felt that the information obtained through this questionnaire would provide evidence on how examinations influence classroom instruction and assessment practices. The information would also help to explain the allocation of items to topics in the tables of test specifications for the duplicate test construction. A copy of the questionnaire is presented as Appendix E.

The second survey aimed at evaluating how each activity of the duplicate test construction experiment contributed to professional growth of the teacher participants in assessment. This questionnaire provided an opportunity to assess the impact of the training that the teacher- participants received. In a way, this acted as a means of ascertaining validity of the activities that were carried out during the test development and administration exercises. This was necessary to accomplish the objective of enhancing assessment skills in the participants of the duplicate test construction experiment. A copy of this questionnaire is presented in Appendix F.

3.3.2 Student-Participant Surveys

The student-participants completed two surveys, one at the beginning and the other at the end of the test construction experiment. The first questionnaire sought to collect information about how well the examinees felt prepared for the 2003 History examination. It provided an opportunity for students to include specific details about the topics that were taught, and number, type and content focus of the tests they had taken during instruction period. This information was used to verify and compliment the self-reporting data that were obtained from teachers about how they prepared the students for the national examination. A copy of the questionnaire is presented in Appendix G.

The second student survey collected information about students' opinions regarding the quality of the three tests, the 2003 MSCE History Paper 1A, Common Mock Form A, and Common Mock Form B that they took. The students were asked to indicate the test that best suited their preparation and to judge the quality of the tests in terms of how well the items matched the content they covered in instruction. In

addition, they were also asked to comment on item clarity, difficulty, and level of the language vocabulary used. A copy of the questionnaire is presented as Appendix H.

3.3.3 Subject Matter Experts Rating Scales and Survey

Two rating scales were constructed to collect ratings for the relevance of items to topics of the History syllabus and to the degree to which each item measured cognitive skills of the Bloom's Taxonomy. The content relevance ratings were based on a six-point rating scale in which a 1 represented "not at all relevant" and a 6 represented "highly relevant." Similarly, for a cognitive skill relevance rating scale, a 1 represented "does not measure this skill at all" and a 6 represented "measures this skill very well." An even-numbered rating scale was chosen for either judgment task to avoid neutral responses from the judges when in doubt. In each rating scale, SMEs used numbers 2 through to 5 to denote the intermediate degree of relevance. With such scales, each item was rated against each content topic and cognitive skill. Copies of the two rating scales are presented as Appendices I and J for content and cognitive relevance ratings, respectively.

3.4 Procedures for Data Collection with Duplicate Tests

The instruments mentioned in the preceding two sections, were used to collect various types of data that have been used in the analyses of the study. All participants on whom data were collected using these instruments indicated their willingness to participate in the study by completing letters of consent. Headteachers completed consent forms on behalf of their students. A sample copy of a consent letter is presented

in Appendix K. The consent letters guaranteed the participants' freedom to participate or pull out from the study without being prejudiced.

Details about the data collection procedures with the duplicate tests were as follows:

3.4.1 Common Mock History Examination

The compiled tests were first piloted in two schools outside the Zomba district before item review. The pilot tryout was conducted on a cluster sample of two schools, which allowed for the participation of typical students. Although such a sample was at best a compromise in view of great disparities among schools its practical convenience outweighed its limitations. A simple tryout was considered sufficient because the purpose of the duplicate test construction was to develop tests that would match the MSCE, which was the current assessment tool of the syllabus. The tryout was meant to check on the directions for the two forms that would be administered together, the extent to which the marked keys were really correct, and the extent to which the multiple-choice alternatives were plausible. Therefore, the tryout was basically a way of getting supplementary information to guide reviewers during final item editing and test review.

3.4.2 Review of Duplicate Tests Items

The two teams swapped papers for the review as a practical requirement of the test development training they received. This was done to provide the participants with the hands-on experience of item review exercise which otherwise would have been

skipped had other people conducted the review. Such a practice was a departure from what testing agencies do and could have probably compromised the independence of the duplicate tests. By swapping the tests, the review teams reviewed the tests objectively without being defensive about a particular item associated with them. The items were reviewed in terms of factual errors and item statistics from the pilot testing were used to locate item inadequacies. The reviewers were not allowed to reword or replace item stems but could improve on the alternatives to ensure factual and grammatical correctness. This process improved the items that were included in the final versions of the tests. Both the MANEB test development trainer and reviewer participated in the editing and review exercise as desk leaders. By involving these experts the teacher-participants were accorded an opportunity to learn how MANEB conducts similar exercises. It also ensured that the two test forms would be comparable in quality to the national examination papers.

For the purpose of administration, the two tests were combined into one test in a common mock examination with a total of 120 items. However, the two tests were kept distinct in parts A and B for the purposes of further data collection. This was necessary to ensure that examinees would be able to determine the overall quality of each test when responding to the second questionnaire. This practical requirement outweighed the need to control for item position effects. Therefore, items in Mock B followed those in Mock A in the common mock examination. Because the combined tests doubled the length of a typical MSCE History multiple-choice component, it became necessary to double the length of administration time for the common mock examination to 2 hours 20 minutes instead of 1 hour 10 minutes. This time extension was not unusual for the

MSCE examinees because most MSCE papers run up to 2 hours 30 minutes and a few go up to 3 hours. The extension of time was therefore, not expected to adversely affect performance of examinees in the common mock examination.

3.4.3 Administration of the Common Mock Examination

As a data collection instrument, the common mock examination was administered to 200 pupils who were candidates for the 2003 MSCE History examination in the six participating schools. By terming this test "MSCE Mock examination," it had an added advantage in securing students' commitment to it because mock examinations in Malawi are regarded as the final teacher made tests that check on the adequacy of preparation of examinees prior to national examinations. Therefore, both teachers and students took this examination more seriously than would have been the case if it were treated as a revision test or just a research instrument.

All schools administered the common mock examination on the same day and time at the end of the second term of the school calendar. Such an arrangement was necessary to minimize leakages that could put the schools that would take the examination after the others at an advantage. Because the common mock would make comparison of student performance from different schools possible, it was feared that it would spark the spirit of competition that has sometimes led to teacher collaborated cheating (Chakwera et al., in press; Malunga, 2000). Therefore, throughout the testing time, security measures were undertaken to ensure fair assessment and objectivity of results from the duplicate test construction experiment.

The teacher participants scored the scripts of their students and reported examinee scores on each test form separately. The students' scripts were entered into Lertap program (Nelson, 1973) to analyze the tests for reliability and obtain item statistics that could be compared with the tryout data to assess item quality. The students' response data were also entered in SPSS program to obtain item score data to compute split-half reliabilities and standard error variances as needed in the analyses. Teachers kept copies of the scores of their students for use as mock examination results. Examination scripts were also returned to students to facilitate revisions, which teachers undertook with them. By returning copies of test items to schools, this study contributed to the assessment resources that would be useful to teachers during the classroom test construction. Teachers have a tendency to store multiple-choice items for future use, and so the sets must have been a great contribution to their capacity to adequately assess the MSCE syllabus in future.

3.4.4 Impact of Duplicate Test Construction Activities on Teacher Participants

The second teacher participant questionnaire was completed after the review exercise that led to the final form of the Common History Mock examination. Individually, teacher participants were asked to indicate how their involvement in the duplicate test construction experiment contributed to their professional growth. The teacher participants were given the questionnaire on the last day of the third workshop and were asked to return the questionnaire together with the scored scripts. More time was given to them so that they could think through all the activities and determine how they felt about how well each activity helped them professionally. A 100% return rate

was again experienced with this questionnaire because the sample of the participants was easily accessible and mutual relationship had been established.

3.5 Procedures for Data Collection with Teacher and Student Surveys

The first two questionnaires for teacher-participants and students were used to collect information about the effect of national examinations on instruction and assessment practices. Both questionnaires asked about the topics that were covered and the number and format of tests that had been written during the course of instruction.

3.5.1 Teachers' Survey on Adequacy of Student Preparation

Through their instrument, teacher participants gave reasons for the decisions they made about the topics they taught and assessed as they prepared students for the 2003 MSCE History examination. The teacher participants completed the questionnaire prior the test construction experiment and this arrangement ensured independence of teacher opinions about students' preparation. The researcher distributed the questionnaires during his first visits to the six schools to recruit teacher participants. All teacher participants returned completed questionnaires to the researcher during first training workshop. Since the teacher participants were meeting their partners in the experiment on the same day that they submitted their responses on this survey, the arrangement ensured the independence of the responses that were made. This was necessary to determine the effect of the experiment on the participants because the questionnaires sought pre-testing information that were contrasted with the information obtained after the experiment.

3.5.2 Students' Survey on Adequacy of Instruction They Received

The first students' questionnaire was meant to draw supplementary information to instructional and assessment decisions that teachers made in their schools. Specifically, it asked students to indicate the topics that they were instructed in and assessed on. They were asked to indicate their perceived adequacy of preparation for the anticipated MSCE History examinations. Through this information it was expected that levels of agreement between the students and teacher's decisions would be indicated by the frequency of the students' responses on each decision. In a way, the student questionnaire was a means of verifying the decisions that the teachers had indicated in their responses. For this reason, the completion of the students' questionnaires was made independent of the teachers that taught the subject. This was considered necessary to guarantee respondents' confidentiality of their responses, some of which reflected their perception about teacher's competence on some topics. For the sake of anonymity, students were advised not to write their names on the questionnaires.

To ensure a high return rate, student questionnaires were distributed through the Headteachers who were instructed to ask one teacher other than the teacher participant to administer the questionnaire to the students. Although the questionnaire was administered in a group setting, its response rate dropped to 62.5% from what was expected and various reasons were given for the low return rate. In one school the low response rate was explained as resulting from failure to find time during the normal school time while in another, students did not just cooperate. As per the letters of consent that the headteachers signed on behalf of the students, this experience clearly demonstrated that students were not forced to complete the questionnaires.

The student questionnaires were not exposed to the teacher participants so that they could not influence their students on what to say. Completed questionnaires were returned to the researcher through the Headteachers in sealed envelopes. Such security measures in data collection ensured authenticity of the student information on decisions that guided their preparation for the national examinations which otherwise would have been confounded with teachers' input in a bid to appear effective. In addition, the procedure enforced the confidentiality that respondents were promised in the directions that were given to them on the questionnaires.

3.5.3 Collection of Item-Topic and Item-Cognitive Relevance Ratings

The fifteen raters were given pairs of item sets on two occasions. Sets 1 and 2 that originally were subsets of the common mock examination were delivered to SMEs first in September 2003 while sets 3 and 4 of Class test and the 2003 MSCE History Paper 1A were given to SMEs in November 2003 after the MSCE examinations had been administered. In each case, the sets were presented without any information that referred to their origin, such as test name or directions to prevent raters' bias or prejudice that would result from familiarity with the origin of the test set.

Two research assistants helped to deliver packages of data collection instruments to the SMEs at their work places in five education divisions. These packages included: a letter of appointment, a consent letter; two sets of items to be rated; two rating scales for each set (one for content and another for cognitive skills); Guidelines for the rating exercise and a copy of the description of Bloom's Taxonomy of the cognitive domain as presented by Linn and Gronlund (2000, pp. 550-551); and

questionnaire for the SMEs' impressions on overall quality of the tests, a copy of which is presented in Appendix L.

The use of assistants to personally deliver the questionnaires and explain the instructions on the rating scales increased the response rate and helped to obtain the kind of data that were expected. The researcher conducted a thorough briefing of the assistants so that they could respond to queries that the SMEs could have. In the process of collecting the information, it was observed that some SMEs had problems in using the rating scales and their queries were handled by phone during follow up conversations with the researcher. Three types of data were obtained from each SME on each set of 60 multiple-choice items. Firstly, SMEs provided content relevance ratings by rating the degree to which an item represented each topic of the MSCE syllabus. Secondly, they also rated the extent to which each item was relevant to each category of the cognitive skill areas as described by Bloom et al. (1956). And finally, the SMEs provided their opinions of the overall quality of the tests they rated as described in next section.

3.5.4 Collection of Data on Test Quality

The test quality questionnaire sought to obtain the SMEs' opinions about their impressions on item clarity, level of difficulty and appropriateness of the language vocabulary used. One questionnaire was used for each set of items for them to judge how well the test as a whole reflected those qualities. The questionnaire was expected to be completed immediately after the rating tasks to ensure that the SMEs made relevant

reference to what they had just rated. The study derived comparisons of the test sets from what all judges said about each paper.

Similar information was obtained from students who completed a second questionnaire after writing the MSCE examinations in November 2003. As was the case with the information about students' preparation, this questionnaire asked for students' perception on test quality to compliment the ratings of the SMEs. Because the survey was taken after the national examination the researcher asked the subject teachers to persuade their students to take the questionnaire otherwise they would be less cooperative. This fear was based on the experience with the first survey that had 62.5 % response rate. The use of the subject teachers could not necessarily affect the results of the second survey because the focus of the survey was on overall quality rather than individual items on the test. However the response rate to this questionnaire was even lower than the first. Only 37% return rate was obtained.

3.6 Data Analysis Procedures

Several procedures were used to analyze the triangulated data that were collected in this study. Test scores, surveys, and SMEs' ratings were collected using different methods because Denzin (1978) as quoted by Chimwenje (1992) stated:

> No single method ever adequately solves the problem of rival causal factors....because each method reveals different aspects of empirical reality, multiple methods must be employed. I now offer as a final methodological rule that multiple methods must be used in any investigation. (p. 96)

The data analyses that were carried out were therefore, consistent with the type of data and appropriate for the research questions. As was the case with methodological

triangulation, different analyses were meant to strengthen interpretation of the findings when more than one source of data and analyses produce similar results.

3.6.1 Mean of Squared Difference Scores and Sum of Error Variances of Mock A and Mock B

Following the proposal of Cronbach (1971, pp. 455-456) test scores from Mock A and Mock B were treated as score distributions of X_1 and X_2 to compute the mean of difference and Standard Errors of Measurement for the two tests. The hypothesis to be investigated was whether the mean of the squared difference scores would be no larger than the sum of the error variance for the two tests, expressed as

$$\frac{\sum (X_1 - X_2)^2}{N} \leq \sigma_{e1}^2 + \sigma_{e2}^2 \tag{3.1}$$

where, X_1 represents observations for Mock A, X_2 represents observations for Mock B, σ_{e1}^2 is the error variance for the Mock A distribution, and σ_{e2}^2 is the error variance for the distribution of scores in Mock B. For the purpose of this analysis, the two error variances were derived from the split-half analyses of the two distributions using the odd-even procedure. The Spearman-Brown formula was used to compute split-half reliability estimate as follows:

$$\text{Split-half Reliability} = \rho_{XX'} = \frac{2(\rho_{AB})}{1 + \rho_{AB}}, \tag{3.2}$$

where, ρ_{AB} is the correlation between "even" and "odd" halves of each mock examination. And the standard error of measurement for the split-half was computed by using the following formula:

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{XX'}} \tag{3.3}$$

where, σ_X is the standard deviation of the half test; $\rho_{XX'}$ is the estimate of the split-half reliability of the half test.

The second analysis of the mean difference that was done on the same data was to test the hypothesis of no difference between the two means using the t –test for dependent samples. The hypothesis was stated as: $Ho: \delta = \mu_1 - \mu_2 = 0$ (3.4)

and for samples this translates to $\bar{d} = \bar{X}_1 - \bar{X}_2$. (3.5)

The t-statistic was computed to determine the significance of the difference between the means of the two test forms. This required estimation of the population variance and standard deviation by using the following formulae

$$sd^2 = \frac{\sum (d - \bar{d})^2}{n-1} \quad \text{and} \quad sd = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} \tag{3.6}$$

The standard error of the difference was estimated by the following formula:

$$s\bar{d} = \frac{sd}{\sqrt{n}} \tag{3.7}$$

and the dependent t- statistic was estimated by the following formula:

$$t = \frac{\bar{d} - \delta}{s\bar{d}} \tag{3.8}$$

Cohen's delta effect size was also computed using the formula:

$$\frac{\bar{d}}{sd} \tag{3.9}$$

3.6.3 Computation of Descriptive Statistics and Percentages of Agreements

For most of the survey data, it was sufficient to compute means, percentages and proportions of agreement to interpret the findings. Some interpretations were based on

the frequencies and an average frequency was sufficient to establish comparison between tests. The SME relevance ratings and survey responses were generally subjected to this kind of analysis.

In addition to describing content ratings as frequencies and percentages, Aiken's (1980) Validity index (V) was computed for each item to determine the extent to which the rating was not due to chance. The equation for the Aiken's Validity Index is:

$$V = \frac{\sum_{i=1}^{c-1} in_i}{N(c-1)} \tag{3.10}$$

where

c is the number of categories on the item relevance rating scale

i is the weight given to each category,

n_i is the number of judges who rated the item in the i th category, and

N is the total number of the SMEs.

As Sireci and Geisinger (1995) put it, "The lowest category is given a weight (or i-value) of 0, the next category is given a weight of 1, and so forth, and the highest category is given a weight of c-1" (p 243). In view of a large number of raters (14 SMEs), the computed Aiken values were tested for significance by using a formula, which provides a normal deviates (z) for the index as follows:

$$z = \frac{N(c-1)(2V-1)-1}{\sqrt{\frac{N(c-1)(c+1)}{3}}} \tag{3.11}$$

The normal deviate values were then used to determine the probability of obtaining the z from a standard normal z table at two levels of significance, $p < .05$ and .01.

3.7 Summary of Data Collection and Analysis Procedure and Conclusion

Activities undertaken in this study are summarized in Table 3.2 as follows:

Table 3.2. Summary of Data Collection and Analysis Procedures

Validity Question	Data Collection Method	Data Type	Data Analysis
To what extent do independent exams developed from the same syllabus equally represent the domain of tasks as described by the educational objectives contained in the syllabus?	Scoring students' responses on mock tests and MSCE 1A; Rating of items on the mock subtests and MSCE 1A for content and cognitive relevance.	Paired scores from mock subtests and MSCE 1A, SME Item relevance ratings of content topic and cognitive skill area.	Comparison of mean of the squared difference scores & sum of error variances; Mean Difference t-test & RMANOVA; Aiken Validity index, V and its significance test.
How well do items on each mock subtest represent items on national exam in terms of content and cognitive levels of achievement?	SME rating of item relevance to topic and cognitive area; inspection of test specifications	Relevance Ratings; number of items per topic in the specification	Comparison of percentage of coverage and correlation
How well do mock subtests, end of term test, and MSCE 1A compare in terms of clarity and difficulty?	SMEs and Student questionnaires/opinion survey	Ratings	Percentage and frequencies; RMANOVA
How well do end of term represent the MSCE History examination content?	SMEs' Judgment of item quality	Rating impressions on overall quality	Frequencies of agreement between judges and RMANOVA.
To what extent do classroom practices reflect exam influence?	Teacher and students' judgment about what was covered and how it was covered in class.	Opinion/impressions of teachers and students	Frequencies and percentage averages

3.7.1 Conclusion

This chapter has described the instruments and procedures that were used to collect the data of the study. A description of approaches to data analyses that were selected has also been outlined. The next chapter attempts to describe how the analyses were carried out and also presents summaries of the findings following the analyses.

CHAPTER 4

PRESENTATION OF RESULTS AND ANALYSIS

This chapter presents results and analyses that were carried out in this study in four sections. The first section, 4.1 presents content validity evidence of the duplicate tests derived from the comparison of the mean of the squared difference scores and the sum of the split-half error variances. It also presents results from the analysis of the mean difference of the dependent samples (dependent t-test). Similar comparisons were made with the MSCE 1A scores and the results are also been reported in this section. Section 4.2 contains further evidence on content validity derived from results of SMEs' item relevance ratings on content and cognitive domain representation. Section 4.3 presents summaries on the student and SMEs' survey about overall quality of the examinations; and finally, section 4.4 presents survey results on the influence of examinations on instruction and assessment practices in the classroom.

4.1 Content Validity Evidence from the Duplicate Constructed Tests

This section presents results from analysis of students' scores on Mock A, Mock B, and MSCE 1A. Equivalence of the duplicate tests was assumed because they were developed and reviewed by people with similar competencies, under the same conditions and they used the same content domain and test construction guidelines. According to Cronbach (1971) this equivalence depends on an ideally refined description of the universe and sampling of the items that go into the test. Once such conditions are fulfilled then "any person's score will be the same on both tests, within

the limits of sampling error" (p. 456). This means that the score of a particular person in either test represents the persons' actual performance regardless of the test sample from the same universe of admissible observations. For this reason, Cronbach (1971) stated:

> To be precise, if the scores are X_1 and X_2 and the corresponding standard errors of measurement are σ_{e1} and σ_{e2}, then the mean of $(X_1 - X_2)^2$ should be no larger than $\sigma_{e1}^2 + \sigma_{e2}^2$. For this purpose σ_{e1} and σ_{e2} can be derived from split-half analyses of X_1 and X_2, respectively. (p. 456)

To sum up, this contention proposes that the mean of the squared difference scores should be no greater than the sum of the split-half error variances of the two equivalent test forms.

4.1.1 Computation of Mean of the Difference and Standard Errors of Measurement

To verify Cronbach's proposition, the first part of the analysis of the study was to compute the mean of the difference between Mock A and B and their corresponding standard errors of measurement for both the full tests and their split halves. This necessitated the inclusion of estimates of reliability coefficients for each test. It was then possible to compare the values of the mean of the squared difference scores and the sum of the split-half error variances. Only the two mock examinations have been included in these computations because the MSCE 1A had neither item score data nor reliability estimates. These would have to be computed for the schools that were involved because the national score reporting does not include reliability estimates. The results of these analyses are reported in Table 4.1 as follows:

Table 4.1. Means, Standard Deviations, and Standard Errors of Measurement of Mock Tests

Descriptive Statistic	Mock A		Mock B
Mean	26.71		27.33
Standard Deviation	7.37		6.92
Reliability (Coefficient alpha)	0.89		0.88
Standard Error of Measurement	2.44		2.40
Error Variance	5.97		5.76
Split half Standard error of Measurement	3.85		3.12
Error Variance (split half)	14.82		9.77
Mean of the Difference = $\frac{\sum(X_1 - X_2)}{N}$		=	-0.63
Mean of Squared Difference = $\frac{\sum(X_1 - X_2)^2}{N}$		=	21.47
Sum of Split-half Error Variance ($\sigma_{e1}^2 + \sigma_{e2}^2$)		=	24.59
N	200		200

The results support the contention that tests that have equally sampled the domain of interest would be statistically equivalent. The difference in means is extremely small and the mean of the squared difference scores is less than the sum of their split-half error variances. In this case, the sum of split-half error variance was found to be 24.61 while the mean of the squared differences was 21.47. On this basis it is concluded that the independently constructed duplicate tests were equivalent and that the test designers essentially interpreted the MSCE History syllabus in the same

manner. This conclusion was further supported by a high and significant ($p < .01$) correlation of .80 between the duplicate tests scores.

4.1.2 Paired Dependent Sample Mean Difference t-Test and 1 Way Repeated Measures ANOVA

The computation of the paired dependent samples' t-test for the two duplicate tests was not significant. This finding is consistent with the mean difference and error variance comparison analysis in the preceding paragraph and supports the conclusion that the independently constructed duplicate tests were equivalent. Therefore, the hypothesis that there is no difference between the mean of Mock A and Mock B was not rejected because the two tests were deemed to be statistically equivalent.

In the mean difference analysis, the classical true score theory assumptions that parallel tests enable examinees to attain the same true score and that they have equal variances (Crocker & Algina, 1986) were made. Using the same assumption, each duplicate test was compared with the MSCE 1A on which the two forms were modeled in both coverage and level of difficulty. This meant extending the same hypothesis of no difference between the mean of either of the mock tests with that of the MSCE 1A. Therefore, scores from the duplicate tests were analyzed together with the MSCE scores using the 1 Way Repeated Measures ANOVA.

The results presented in Table 4.2 show no significant differences among the means of the three tests, Mock A, Mock B and the MSCE 1A as indicated by the repeated measures F- tests. In addition, the 95% confidence intervals for the three means overlapped while the effect size of .008 and .024 for the F tests were negligible.

This means that the two mock subtests were individually statistically equivalent to the MSCE 1A.

Table 4.2. Summary of 1-Way Repeated Measures ANOVA of the Three Tests

	Factor				
Statistic/Test	Mock A	Mock B	MSCE 1A	p	eta-squared
Mean	26.71	27.33	27.83		
SD	7.37	6.92	7.41		
95% CI	25.68-27.73	26.37-28.29	26.81-28.85		
Pillai's Trace		F = 2.476		.087	.024
Within Subject Effects		F = 1.663		.191	.008

$p < .05$

It was pleasing to note these results because they were an indication that it was possible to construct tests that would be comparable with the national examinations when test constructors have been trained. The results meant that the teacher participants who claimed not being able to construct their own multiple-choice items had gained some competency in assessment through the experiment.

4.2 Judgmental Evidence for Content and Cognitive Skills Relevance of the Items

Fourteen SMEs rated content and cognitive relevance of the items in four tests against the MSCE syllabus topics and the Cognitive skill taxonomy of Bloom et al (1956)

on separate 1 to 6 rating scales. The data from the two scales were averaged for each item to establish agreement of the SMEs on the allocation of an item to either a topic or cognitive skill. A mean of 3 was determined as a cut-off point for an item to be

considered as relevant but the final placement of the item was decided in favor of the topic or skill to which the item had the highest average rating. The raters' judgments were compared to the test specifications to determine how well the tests represented the specifications that were set. By comparing the number of items on each topic or skill domain across the tests, the degree to which each test covered the MSCE History content and cognitive skills was established. In this analysis, four tests including the class test were considered.

4.2.1 Test Designers' Item Allocations to Topics and Cognitive Domain Areas

Although the SMEs were not shown the test specifications, their ratings were interpreted in relation to what was included by the test specifications. Table 4.4 presents test specification tables for the two duplicate tests and MSCE 1A. The test specification for the Class Test was not available because its purpose was mainly to show the type of items that teachers constructed and the number of topics on which students were assessed. Being a class test, coverage of the whole syllabus was not expected because teaching was still going on by the time of the experiment. It should also be noted that the MSCE test blueprint was constructed after the examination had been written. It was not possible to obtain the original specifications from MANEB, and so the researcher constructed the blueprint using his knowledge of the subject matter to facilitate interpretation. These specifications are presented in Table 4.3 as follows:

Table 4.3. Test Specification Tables for Mock A, Mock B, and MSCE 1A

	Mock A				Mock B				MSCE 1A			
Content Areas	*K	C	H	T	K	C	H	T	K	C	H	T
Iron Age	0	1	2	3	1	1	2	4	1	1	0	2
Pre-Colonial Kingdoms	1	1	1	3	1	2	1	4	4	1	0	5
Trade in gold, ivory, slaves	1	1	1	3	1	2	1	4	3	1	2	6
19th Century Immigrants	1	2	2	5	2	1	1	4	2	3	0	5
Missionary Factor	1	2	2	5	0	2	2	4	2	2	1	5
European Occupation of CA	0	2	3	5	1	2	2	5	2	1	0	3
Political and Economic dev.	2	2	2	6	1	2	2	5	3	1	0	4
Causes and Results of WW1	2	2	2	6	1	3	2	6	6	4	2	12
The Inter War Period	1	2	3	6	0	6	3	9	10	1	1	12
Causes of WW	2	2	3	7	1	3	1	5	1	4	0	5
Decolonisation in Asia, Afr.	2	2	1	5	1	2	2	5	0	1	0	1
Post-Colonial Alliances	2	2	2	6	1	2	2	5	0	0	0	0

*Note: K = Knowledge, C= Comprehension, H = Higher Order Abilities; T = Total

As it can be observed from Table 4.3 both test construction teams planned for100 percent coverage of the content domain while the MSCE 1A covered 92 percent of the same MSCE syllabus. This means that all tests ensured content representation by drawing items from all content areas, but they were different in terms of the weighting of the topics. This must have reflected the belief systems of the test designers regarding the importance of the content areas. They also differed considerably with respect to the allocation of the items according to the cognitive domain areas.

It is also observed that while the two mock tests attempted to have an even distribution of the items over all content areas, they both had relatively higher weighting over three topics of Causes of First World War, The Inter War Period, and Causes of the Second World War. The MSCE 1A also concentrated on the same topics, especially the Causes of the First World War and the Interwar Period in which there were 12 items each. This obviously shows that both teachers and MANEB test developers had same perception regarding the importance of these topics. However, it is of concern that the MSCE 1A had allocated 40 percent of items to two topics only and yet there was no question for the Post Colonial Crises and Challenges. The MSCE 1A did not match well with the duplicate constructed tests in those areas because of overrepresentation and underrepresentation.

In responding to why they included those topics two reasons were advanced. Firstly, it was because they wanted to ensure wide coverage of the syllabus and secondly because the topics were relatively new in the syllabus. As such, they felt that national examiners would most likely find something new to ask by focusing items on those topics. This was one indication of how examinations influence assessment decisions at

classroom level. As long as the topic is considered likely to be examined in the national examination, teachers are likely to include items from that topic in their final class assessment as a way of showing students what to concentrate on as they make final preparations for the national examinations.

4.2.2 Content Relevance Ratings

Relevance ratings were averaged over the 14 SMEs and mean relevance ratings for each content area was computed (Hambleton 1984). An item was considered relevant to a topic when its highest mean was ≥ 3. By identifying the match to the topic with the highest mean it meant that an items could be allocated to one topic only. The cut off point of the mean rating of 3 was considered appropriate because it is the mid point of the six-point relevance rating scale. Table 4.4 presents the results as follows:

Table 4.4. SMEs' Average Rating Determination of the Number of Items Per Content Domain Area

Content Topic	Mock A	Mock B	MSCE	Class Test
Central Africa Section				
Iron Age	3	4	3	3
Pre-colonial Kingdoms	3	4	7	1
Trade in Gold, Ivory and Slaves	3	2	4	3
19th Century Immigrants	5	2	3	7
The Missionary Factor	5	4	7	4
European Occupation In Cent. Afri.	5	8	4	10
Political and econ. Develp in C.A*.	6	6	2	14
World History Section				
Causes and Results of World War I	8	10	11	14
Inter-war Period (1919 – 1939)	7	10	12	0
Causes & Results of World War II	5	2	6	0
Decolonisation in Asia and Africa	4	4	1	0
Post Colonial Crises and Challenges	6	4	0	0

In terms of total coverage of the syllabus, it is noticed that both Mock A and Mock B had 100 percent coverage while the MSCE1A covered only 92 percent of the same MSCE History syllabus. The last two topics on the Central Africa section and the first two topics on the World History section received greater weighting than the other topics. Similar topic concentration is noticed on both the MSCE 1A and the Class Test. For Mock B and MSCE 1A the item concentration on the first two topics of the World History section accounts for 33 and 38 percent of the items in the tests, respectively. But MSCE 1A is grossly underrepresented in the last two topics of the syllabus with one item on Decolonisation of Asia and Africa and no item on the Post Colonial Crises and Challenges. Yet both teams of the duplicate test constructors included up to 17 and 13 percent, respectively of the items to the last two topics of the World History section.

The difference in areas of item concentration between the MSCE 1A and the mock tests was further illustrated by differences in correlation coefficients derived from item allocation in the test specifications. The correlation coefficients were .71 between Mock A and B; .61 between mock B and MSCE 1A; and .46 between Mock A and MSCE 1A. This means that MSCE 1A had considerable different item content weightings than the two duplicate tests.

Apart from analyzing the match of the items according to topics from the relevance ratings of SMEs, Aiken's (1980) validity index was computed for each item by using equation 3.10. This index helped to ascertain that the rating of the items in the categories was not by chance and the higher the values of the Aiken index, the lesser the effect of chance in the rating. According to Sireci and Geisinger (1995) the "index accounts for the number of categories used to rate each item and the number of judges

that responded to each category" (p. 243). In addition the Aiken indices were also tested for significance by using equation 3.11. The mean Aiken values for items in each topic across the tests were analyzed and are presented in Table 4.5 as follows:

Table 4.5. Mean Aiken Values for Items Per Topic Across Tests

Topic	Mock A	Mock B	MSCE 1A
Iron Age	.95*	.94*	.98*
Pre-Colonial Kingdoms	.90*	.89*	.98*
Trade in Gold, Ivory and Slaves	.84*	.80*	.94*
19th Century Immigrants	.95*	.94*	.96*
The Missionary Factor	.90*	.87*	.96*
The European Occupation of C Africa	.94*	.92*	.92*
Political and Economic Develop in Africa	.85*	.93*	.88*
Causes and Results of 1st World War	.84*	.92*	.89*
Interwar Period (1919-1939)	.87*	.85*	.87*
Causes and Results of 2nd World War	.87*	.89*	.86*
Decolisation in Asia and Africa	.97*	.96*	.95*
Post Colonial Crises and Challenges	.83*	.95*	N/A
Mean Aiken Value for the Test	.88*	.90*	.92*

Note: * $p < .05$; N/A = no item was included on the topic.

The range of the mean Aiken validity indices for all the items representing the topics of the content domain was from .80 and .98 and they were all statistically significant. This means that all items were accurately rated according to the topics that they represented in the tests. For each test the mean Aiken index values were .88, .90, and .92 for Mock A, Mock B, and MSCE 1A, respectively. These results are consistent with those reported earlier in Table 4.4 and they demonstrate that the three tests had high content validity.

4.2.3 Inclusion of Items Not Related to the MSCE History Syllabus

Content validation is traditionally concerned with the relevance of items to the domain of interest and focuses on representativeness to describe test validity. Sometimes

however, the problem of construct irrelevance can undermine content validity of a test when it includes items that do not belong to the content domain of interest. When such items are discarded in an evaluation that may cause underrepresentation of the content domain. Therefore, SMEs were asked to identify items that were drawn from History content that is not prescribed in the MSCE syllabus. An item was considered drawn from outside the syllabus when at least three SMEs, which represents 21% of the evaluators, marked it as such. Using this criterion, it is observed that only the classroom test did not have items that were identified as drawn from outside the MSCE History syllabus. The results of the analysis are presented in Table 4.6 as follows:

Table 4.6. Items Identified as Drawn From Outside the Content Domain Across Tests and their Aiken's Validity Indices (V)

Mock A		Mock B		MSCE	
Item	V	Item	V	Item	V
6	.814	11	.814	11	.942
7	.856	19	.542	17	.942
48	.857	53	.814	42	.628

Note: V = Aiken's Validity index

The results in Table 4.6 show that there was confusion regarding the limits of certain topics in the syllabus. Items 6, 7, and 11 (for Mock B and MSCE 1A) dealt with the Portuguese and the gold trade in Africa. The current MSCE History syllabus removed a topic on Portuguese Influence in Central Africa but retained the topic on Trade in Gold Slaves and Ivory. Since the Portuguese were players in the Central African trade it is not clear to teachers as to what factors of the Portuguese were removed with the topic their influence. This confusion explains the high validity indices despite labeling the items as drawn from out of the syllabus. Item 17 was on the Bemba conflict with Europeans but it

was being confused with the topic of the Bemba kingdom, which has also been removed from the current syllabus. There was however, agreement in rating it as being relevant to European Occupation of Central Africa. Item number 19 had the lowest validity index because it touched on the Missionary factor, Pre-Colonial Kingdom of Mwenemutapa, and the Portuguese involvement in the gold trade. Both the Mwenemutapa Kingdom and the Portuguese influence topics are no longer in the History syllabus. Therefore, SMEs' used personal discretion as to what topic they had to associate with the item. For the World History section, item 42 was on Italian dictatorship while 48 was on the Mandatory system of the League of Nations. While Italy is indeed out of the syllabus, details on the Mandatory system appear to be a disputable extension of the topic.

The fact that the three tests had the same problem of stray items points to looseness of the syllabus rather than sheer carelessness of test constructors. This is a serious problem for an achievement test, which should be based on a well-defined domain for it to be content-valid. When the interpretation of the limits of the content domain is not clear, agreements between instructors and examiners cannot be guaranteed.

4.2.4 Cognitive Coverage Relevance Ratings of Items Across Tests

The SMEs' ratings of the items according to cognitive skills that they measured were tabulated as frequencies and the results are presented in Table 4.7. The results were compared with the test specifications to determine how well the test designers followed their plans when coming up with the tests.

Table 4.7. Summary of Item Cognitive Rating Frequency Per Domain Area Per Test

Cognitive Level	Number of Items In a Test			
	Mock A	Mock B	MSCE	Class Test
Knowledge	39	45	50	50
Comprehension	21	15	10	10
Application	0	0	0	0
Analysis	0	0	0	0
Synthesis	0	0	0	0
Evaluation	0	0	0	0
Other	0	0	0	0

It is observed that all items measured up to comprehension level of the cognitive domain. The MSCE 1A and the Class Test had more items assessing knowledge than comprehension while the duplicate tests had relatively higher number of comprehension items. These results clearly contradict what the test designers had indicated in their tables of specifications. What the test constructors considered being items of higher order abilities did not receive similar ratings from the SMEs. It is however, not clear whether this indeed meant absence of such items or failure of the raters to recognize items measuring higher order abilities. It may not have been easy for the SMEs to recognize the items as measuring anything beyond comprehension because they might not have been familiar with such distinctions in assessment. The raters were basically teachers with the same experience as those who constructed the tests. Although they were given guide notes on how to recognize such items there is no evidence that they really used the notes adequately to confidently rate the items according to cognitive skills.

For History, the raters' inability could be hidden in the fact that the subject is basically factual in nature. It would therefore, be expected that by focusing on such facts items render themselves to knowledge and understanding of the facts. It is however, psychometrically worrying in that even the national examination board is not able to train persons who can construct higher order items because the 2003 MSCE 1A had none rated for that level. Among the teachers, the art of constructing multiple-choice items is not common, as many had confessed not able to do so (Selemani-Mbewe, 2003). Participants in this experiment made similar sentiments after the short test construction training, which they recognized as an "eye opener."

4.3 Perceived Overall Quality Impressions on the Mock Tests and MSCE 1A.

The ratings of content and cognitive skills relevance, which dominate test validation, tend to ignore other item qualities that can affect examinee performance. Such qualities as item clarity, test difficulty, and language level are important in test evaluations. Both examinees and SMEs were asked to rate the four tests in terms of overall impressions about quality of the items in the three tests with respect to the specific features mentioned in this section.

4.3.1 Appropriateness of Language and Item Clarity

The English language is a medium of instruction and assessment for MSCE History but it was a second language to all examinees. For this reason, it was important to ensure that the language vocabulary used in a test would be appropriate otherwise it could lead to errors that have nothing to do with the content itself. Percentages of students'

responses on the ratings of item clarity and language appropriateness are presented in Table 4.8.

Table 4. 8. Percentage of Students' Ratings for Overall Test Clarity of Language Vocabulary, N = 76

Test Paper	Item Clarity and Language Appropriateness				
	1	2	3	4	most rated category
Mock A	01	14	44	36	3
Mock B	02	13	41	36	3
MSCE	05	23	38	32	3

Note: 1 = Poor; 2 = Marginal; 3 = Satisfactory; 4 = Excellent

It is observed from the results in Table 4.8 that the majority of examinees rated the tests as being satisfactory in terms of clarity and language appropriateness for their academic level. The most favorable category was 3 in each case, which means that examinees considered the tests to have been clear and that items used appropriate level of language vocabulary. This observation was supported by the results of the one way repeated measures ANOVA as presented in Table 4.9 follows:

Table 4.9. Summary of 1-Way Repeated Measures ANOVA of Student' Ratings on Item Clarity and Language Vocabulary Appropriateness Across Three Tests

Statistic/Test	Factor Mock A	Mock B	MSCE 1A	p
Mean	23.75	23	24	
SD	19.77	18.56	14.39	
Pillai's Trace		F = .398		.715
Within Subject Effects		F = .144		.868

$p < .05$

As can be seen in Table 4.10, both Pallai's Trace and within-subjects effects F-tests were not statistically significant at both the .05 and .01 α levels. This means that the

mean ratings for clarity and language appropriateness across tests were not significantly different from each other. Students rated all the three tests as having same level of clarity and language vocabulary. For independently constructed tests this finding further demonstrated that the mock tests were not only equivalent in content coverage, but also that they were produced to a required MSCE level.

Additional evidence on the clarity and language appropriateness was obtained from SMEs whose frequency ratings were transformed into percentages on each rating category and are reported in Table 4.10 as follows:

Table 4.10. Percentage of SMEs' Rating on Clarity of Test Items, N = 14

	Judgment Categories				
	1	2	3	4	$\bar{X}$
Mock A	0	0	43	57	3.57
Mock B	0	0	57	42	3.43
MSCE	0	21	43	35	2.71
Class Test	0	14	57	24	3.00

Note: 1 = not at all clear; 2 = less clear; 3 = clear enough; 4 = very clear

In general, both examinees and SMEs rated all the four tests as clear enough and that the items used appropriate level of language for the form 4 examinees. Of the three tests, it is also observed that in both Tables 4.8 and 4.10, Mock A is rated the most clear and appropriate while the MSCE 1A is the least clear. However, the test of the mean differences in the rating percentage using the one-way repeated measures ANOVA was not significant as presented in Table 4.11.

Table 4.11. Summary of 1-Way Repeated Measures ANOVA of the SMEs' Ratings on Item Clarity Across Four Tests, N = 14

Statistic/Test	Factor Mock A	Mock B	MSCE 1A	C/Test	p
Mean	25	24.75	24.75	24.75	
SD	29.43	29.23	18.83	24.02	
Pillai's Trace		F = 1.000			.609
Within Subject Effects		F = .001			1.000

$p < .05$

This shows that all the four tests were equally understood by the examinees because no significant differences were observed in SMEs' ratings of item clarity. It is also noted that even the classroom test items were also rated as being written in appropriate language for the level of examinees. This is surprising because some of the teachers' indicated that they were not able to develop multiple-choice items.

4.3.2 Overall Test Difficulty

Students were also asked to rate the mock examinations and MSCE 1A in terms of perceived difficulty. Their responses were summarized and are reported in Table 4.12.

Table 4.12

Percentage of Students' Ratings on Overall Test Difficulty, N = 74

Test Paper	Ratings of Test Difficulty				
	1	2	3	4	$\overline{X}$
Mock A	14	43	14	3	2.05
Mock B	10	43	19	2	2.18
MSCE	13	35	15	6	2.05

Note: 1 = Easy; 2 = somewhat difficult; 3 = difficult; 4 = very difficult

Examinees rated the papers as somewhat difficult and the MSCE 1A had a relatively higher rate for very difficult than the mock papers. This again demonstrates that examinees were more comfortable with the tests that their teachers developed than what MANEB had constructed. This is not surprising because the teachers constructed the mock examinations with their students' ability at the back of their minds. In general however, the average rating was around category 2 for each of the tests, which means that students considered all the three tests as somewhat difficult. The three tests were therefore, perceived as being of equal difficulty. This conclusion is consisted with the results of repeated measures ANOVA, which showed that there were no significant differences between the mean ratings of the students over the difficulty of the tests. The results are presented in Table 4.13 as follows:

Table 4.13. Summary of 1-Way Repeated Measures ANOVA of Student's Ratings on Item Difficulty Across Four Tests

Statistic/Test	Mock A	Factor Mock B	MSCE 1A	p
Mean	18.5	18.5	17.25	
SD	17.14	17.75	12.45	
Pilai's		F = .092		.916
Within Subject Effects		F = .179		840

$p < .05$

As it can be seen from Table 4.14, no significant F test was obtained in the repeated measures ANOVA. Therefore, the tests were not significantly different in terms of item difficulty. Students felt that the three tests were of the same difficulty.

4.4 Equivalence of Independently Constructed Mock Examinations

The two teams that independently developed Mock A and Mock B began the exercise by constructing a test specification grid in each case. As presented in Table 4.3 earlier, test specifications were noted as the beginning point for establishing equivalence of the duplicate tests. Comparing the weighting that each team placed on each topic and cognitive category in each test did this. In addition, Teacher participants were asked to express their opinions about what they learned from the duplicate test construction experiment.

4.4.1 Participants' Impressions about the Test Construction Experiment

Apart from obtaining the duplicate test construction data that would be useful in content validity analysis, the experiment also served the purpose of equipping teachers with appropriate testing skills. Prior to the experiment teacher-participants had indicated that they were not familiar with test construction techniques. It was therefore, important to check how much the training had benefited them in view of the fact that one key objective for the experiment was to equip teachers with assessment skills that could be useful in their practice. The training was necessary to make the teachers more confident in their attempts to prepare students for national examinations than had been the case before. Unless teachers have received extra training in test construction or scoring national examinations, what MANEB does remains a mystery to them. Participants' responses to the survey about whether they considered the experiment worthwhile are presented as frequency ratings in Table 4.14 as follows:

Table 4.14. Frequency Ratings of Teachers' Impressions about Duplicate Test Construction Experiment, N = 6

1. Mock Construction Activities	not useful useful	somewhat useful	useful	very useful
Multiple-choice construction techniques	0	0	1	5
Construction of Higher order ability items	0	0	1	5
Construction of Test Specification grid	0	0	2	4
Item Analysis Techniques	0	0	2	4

2. Lessons from Presentations	not at all informative	somewhat informative	informative	very informative
Selection of Topics For the Test	0	0	4	2
Construction of HOA items	0	0	3	3
Selection of Test Items	0	1	1	4
Editing Items	0	0	1	5
Moderating Items	0	0	1	5
Using Item Analysis information	0	1	3	2

Information in Table 4.14 demonstrates that teachers found their participation in the duplicate test construction experiment enriching as evidenced by the high rating of the activities that were involved. In responding to whether the exercise made them more confident about testing, they all responded with an emphatic "yes" and highlighted multiple-choice construction, item editing, and review as activities that they found most useful and informative. One of them said,

> This exercise has been helpful to me because for the first time I have been able to construct my own multiple-choice questions. Although at first I doubted the standard of my items I was happy when my team accepted my items. Activities such as MC test construction, editing and moderation were very important.

While another one said: "The exercise reminded me of things I learned but never practiced such as, test blueprint. I now realize that a comprehensive test can come out of a well focused test specification." It is clear from comments such as these, that the exercise had met its objective of helping teachers to acquire new skills in testing. Their

confidence was boosted because they felt able to do what they were not capable of doing before participating in the experiment.

The participants also indicated that the exercise helped them to better understand the link between instruction and assessment in that, the latter informs about how the former met objectives for teaching and learning. One of the participants said, "I have come to know better the link between instruction and assessment and appreciate the role of editing and moderation to minimize errors that would result from ambiguities in the items." The knowledge and skills they gained from participating in this experiment would be supportive in their career as teachers because assessment competence would supplement the teaching competence they already had. When they were asked about further training they would require in testing, there was an agreement about having more time spent on item analysis and how to interpret its data to improve test items It was felt that the activity had not been given adequate time during the training, yet it was the least known to the teachers.

4.5. Evidence of Examination Influence on Instruction and Assessment Practices

In high-stakes achievement testing such as the MSCE examinations, it would be expected that instructional decisions would be driven by the need to articulate the content and objectives that best meet the focus of the examinations. The choice of topics for instruction and assessment would be an expression of what teachers think best represent the test designers' belief system. It was felt necessary to evaluate instructional and classroom test decisions in the process of student preparation to understand the relationship between teachers and test designers' belief systems regarding the topics of

the MSCE History syllabus. In view of school differences and teacher belief system differences in choosing instructional methods and content, both teachers and students were asked to evaluate adequacy of instruction in preparation for the 2003 MSCE examination. This was important to derive evidence for the background of the examinees and the influence of national examination in the instructional process.

4.5.1 Teachers' Instructional Emphasis Prior to Mock Examination

In a survey that was administered to teachers prior to their participation in the construction of the common mock examinations, teachers were asked to complete a questionnaire that covered several aspects of classroom instruction. Two of the questions asked about the extent to which they regarded the topics as important and how well they covered the topics they taught their students. The teachers' responses to the questions on the survey are presented in Table 4.15.

The results show that the only subtopics that received an average rating of less than 3, which was equivalent to the rating of somewhat important, were Iron Age, the League of Nations, and Failure of the League. Otherwise teachers felt that the rest of topics, which represented 92 percent of the domain were important although they differed in the weighting of various subtopics. Such differences explain why the weighting of what was actually covered differed considerably from the weighting of what was considered important. The topic coverage columns of Table 4.15 show that only 38 percent of the subtopics were adequately covered with an average rating of 3 and above across schools.

Table 4.15. Mean and Standard Deviation of Teachers' Rating on the Importance and Coverage of Syllabus Topics

Syllabus Topic	Importance $\bar{X}$	SD	Coverage adequacy $\bar{X}$	SD
Iron Age	2.8	.4	3.5	.5
Pre-colonial Kingdoms				
Maravi	3.5	.5	3.3	.8
Tumbuka Nkhamanga	3.2	.8	3.5	.5
Kazembe	3.2	.4	3.7	.5
Trade in Gold, Ivory and Slaves	3.2	1.2	3.2	1.2
19th Century Immigrants				
The Yao	3.3	.8	3.3	.8
The Ndebele	3.3	.5	3.3	.5
The Ngoni	3.2	.8	3.0	1.1
The Missionary Factory				
Spread of Islam	3.7	.5	3.0	1.1
Spread of Christianity	3.3	.8	3.0	1.1
Impact of Missionaries	3.2	1.0	2.7	1.0
European Occupation of C. Africa				
Nyasaland	3.7	.5	2.7	1.4
Southern Rhodesia	3.3	.8	3.2	1.2
Northern Rhodesia	3.0	.9	2.8	1.2
Political & Economic dev. in CA				
African reaction to colonial.	3.5	.8	2.7	1.7
Agriculture and Mining	3.5	.5	2.3	1.5
Central African Federation	3.7	.5	2.3	1.5
National Political Parties	3.7	.5	2.2	1.5
Causes & Results of First World War				
Alliances	3.5	.5	3.8	.4
Arms Race	3.7	.5	3.7	.8
Nationalism	3.0	.9	3.5	.8
International Crises	3.3	.5	3.7	.5
Versailles Treaty	3.5	.5	2.5	1.4
League of Nations	2.5	1.0	2.2	1.2
Inter-war Period				
Economic problems	4.0	0	2.8	1.5
Russian Revolution	3.0	.9	2.3	1.2
German Dictatorship	3.0	1.0	2.2	1.5
Causes of the 2nd World War				
Problems of Versailles Treaty	2.8	1.2	1.8	1.3
German Aggression	3.0	1.3	1.8	1.3
Appeasement Policy	3.2	1.2	2.0	1.5
Failure of the League	2.7	1.2	2.2	1.5
Decolonisation in Asia and Africa				
India	3.5	.8	1.8	1.3
Kenya	3.7	.8	1.5	.8
Post Colonial Challenges and Crises in Africa				
African Expectations	3.7	.5	1.5	.8
Mozambican Civil War	3.7	.5	1.5	.8
Military Govts in Nigeria	3.7	.5	1.5	.8
Impact of crises on society	3.5	.5	1.5	.8

Note: 1 = not at all important or not at all covered; 2 = somewhat important or less well Covered; 3 = important or well covered; 4 = very important or very well covered

The fact that teachers rated 92 percent of the content topics as important and yet were not able to cover 62 percent of the same syllabus three months before the final examinations is a matter of concern for a two-year course. This would mean either that teachers were not working hard enough or that the syllabus is to long for a two-year instruction time. In responding to a focused interview questions that followed up the responses from the questionnaire, reasons like "the syllabus is too long" and "frequent teacher change over" were common. In fact, 67 percent of the participants indicated that they took over the classes in form 4 and they claimed that the teachers who taught the first parts of the syllabus were too slow and that affected the pace of instruction in form 4. Students' responses about adequacy of syllabus coverage were compared to what teachers had said and results are also presented in this section. This was important to establish the relationship between instruction and assessment because it was assumed that classroom practices were intended to match assessment practices at national level.

4.5.2. Evidence for Assessment of Examinees During Instruction

It was also felt appropriate to find out the topics on which the teachers assessed their students in preparation for national examination. It is often recognized that teachers who assess their students more frequently than not help them to gain appropriate skills for more demanding but similar curriculum based examinations. Classroom assessment often provides feedback on adequacy of examinee preparation when national examinations are being expected. Table 4.16 summarizes teachers' responses on the topics on which the students were assessed and the topics they wanted to include in the Mock Examination.

Table 4.16. Percentage of Teacher Agreement on Topics on which students were assessed and What To Include In Mock Examinations Across Schools, N = 6

Syllabus Topic	Did you assess this topic? % Yes	Should topic be assessed in Mock Exam? % Yes
Iron Age	83	67
Pre-colonial Kingdoms	100	83
Trade in gold, ivory and slaves	83	83
19th Century immigrants	66	67
Missionary Factor	66	100
European Occupation of CA	66	83
Political & Economic Develop. CA	17	83
Causes and Results of 1st World War	50	100
Inter-war Period	50	83
Causes & Results of 2nd World War	50	67
Decolonization in Asia and Africa	33	67
Post Colonial Challenges and Crisis	33	67
Mean	58	79

More topics in the Central Africa section were assessed than the World History section. This is not surprising because schools often begin teaching the syllabus with the Central Africa section in form 3 and teach most of the topics of World History in form 4. The least assessed topics were the Political and Economic Developments in Central Africa and the last two World History topics. It is also observed that there was less agreement on what was assessed in the classroom than what was to be assessed in the Mock Examination, with mean percentages being 59 and 79, respectively. In general,

teachers agreed that Mock examinations should assess 79 percent of the syllabus yet classroom tests only assessed 59 percent. A paired score dependent sample t-test of the difference of the two means was found to be significant at the .05 alpha level. This suggested different weightings in assessment between class tests and mock examinations.

The fact that teachers agreed on higher percentage of coverage for the mock examination than on what was assessed during instruction is not surprising. The mock examinations are considered to be the final test to tease out the students on how well they are prepared for the national examinations. Even when teachers have not completed teaching the topics of the syllabus, the tendency for mock test is to include questions from the uncovered areas to ensure that students prepare adequately. In response to why they wanted to include those topics in the mock examination, 67% of the teachers indicated that they wanted to ensure that all areas anticipated to be covered by the national examinations were covered by the test. This was seen necessary because the mock examinations are often adequately revised to help students to update themselves before the national examinations. It is during such revisions that teachers highlight important aspects of the topics that have not been taught in class. It is clear from this evidence that national examinations have some influence on content coverage and type of assessment that teachers conduct in the classroom.

4.5.3 Students' Perception of Instruction and Assessment Prior to National Examination

Student perspective about the preparation they went through during instruction over the MSCE History syllabus was obtained through a questionnaire that repeated more or less what the teachers were asked. This also served as a way of verifying the teachers

self reported data by hearing from those with the greatest interest in classroom practices. Students' responses to the question of how well the topics of the syllabus were covered were summarized as a percentage over schools as presented in Table 4.17 below:

Table 4.17. Percentage of Students' Frequency Rating on Extent of Topic Coverage Across Schools, N = 124

Topic	1	2	3	4	$\bar{X}$
Iron Age	03	12	36	48	2.63
Pre-col. King.	01	09	35	55	2.77
Trade	06	11	40	41	2.52
Immigrants	07	14	37	41	2.40
Missionary Factor	16	11	31	41	2.38
Europ. Occup.	16	25	37	20	2.07
P & E Dev. CA	36	17	35	10	1.73
C & R of WW1	04	03	35	58	2.80
Inter-war period	22	13	23	41	2.27
C & R of WW2	35	07	23	34	2.04
Decol. Asia & Af.	57	12	19	13	1.53
Post Col. CC	73	08	19	03	1.27

Note: 1 = did not cover at all; 2 = poorly covered; 3 = covered well; 4 = covered very well

The results in Table 4.17 are consistent with teachers' information in Table 4. 15 in terms of topic coverage. The same topics that were indicated not covered by most teachers were also indicated as such by the majority of students. This is true for topics like Political and Economic Development of Central Africa, Decolonisation in Asia and Africa, and Post Colonial Crises and Challenges in Africa. It is also noted that students were generally satisfied with the coverage of the topics that were taught, as the "Very well covered" category seems to have been the most favored rating for the topics that were covered. However, poor coverage and non-coverage was rated by about 34 percent of the total respondents. This appears to agree with the observation in the teacher's data that about 38 percent of the syllabus was not taught by the time of the Mock examination.

Students were also asked to identify topics on which they felt most adequately and least adequately prepared for the MSCE examinations. Their responses were analyzed as percentages and were reported as means and standard deviations of the ratings about most and least adequately prepared topics as presented in Table 4.18.

Table 4.18. Mean and Standard Deviation of a Distribution of Student Percentages of Choice of the Most and Least Adequately Prepared Topics, N = 124.

Topic	Student Perception of Adequacy of Preparation Per Topic			
	Most Adequate		Least Adequate	
	Mean	SD	Mean	SD
Iron Age	59.7	17.6	20.7	18.4
Precolonial Kingdom	55.2	15.0	10.3	7.9
Trade in Gold, Ivory and Slaves	47.5	16.2	18.5	12.6
19th Century Immigrants	43.5	15.1	21.2	12.1
The Missionary Factor	39.3	22.0	20.7	19.4
European Occupation of C Africa	31.7	21.8	37.7	18.7
Political & Econ Development	16.7	17.0	37.0	13.7
Causes and Results of WW 1	64.8	14.49	10.5	11.4
Interwar Period	39.2	23.57	23.7	12.8
Causes and Results of WW 2	26.5	20.75	25.8	16.0
Decolonisation in Asia and Africa	13.3	16.3	29.0	11.9
Post Colonial Crises & Challenges	6.3	7.9	32.2	13.9

Over 50 percent of the students indicated that they were most adequately prepared in three topics only. This was again consistent with the topics of Iron Age, Pre-colonial Kingdoms, and Causes of the First World War, which all teachers indicated to have taught. Similarly, the least covered topics of Political and Economic Development in Central Africa and Post Colonial Crises and Challenges were also ear marked by students as the topics in which they were least prepared. Topics like the Missionary Factor, European Occupation of Central Africa, Interwar Period and the Causes and Results of the Second World War had large standard deviations because they had greater variation in coverage just as teacher coverage ratings in Table 4.15 had indicated. Similar variation

was noted in the least prepared question. Fewer numbers of candidates identified each topic as different teachers covered different topics differently.

Only 115 students responded to the question that required them to indicate the grades they expected to obtain in the MSCE examination. Forty three percent indicated grade 1 and 2, which are distinction pass category, and 56 percent expected between 3 and 6 in the credit pass category. Only one student indicated a failing grade. The question was intended to find out whether students had confidence in the preparation they received during instruction. Despite rating most topics as not satisfactorily covered, students were expecting to get good grades in national examination. Such optimism was surprising but it should have derived from the fact that students most often count on last minute activities that teachers perform in preparation for examinations. It is common for both teachers and students to expect a review of the mock examination as the determining factor of students' success in the national examination. This also explains why teachers were willing to include into the mock examination items from the topics that were not taught. By going through the items during the revision lessons, additional information on the least or uncovered topics would be given to students.

Examinees were also asked to indicate the type of items that frequently appeared in the classroom tests and the frequency of such tests. This information was sought to establish the influence of national examinations in classroom assessment practices. In responding to these questions, students indicated an average of 7 classroom tests which they had taken since they started getting instruction in the MSCE History syllabus. This translated to about two tests per term. The most popular item format in class testing was an essay type. The next in popularity was the short-answer type, and the multiple-choice

type came last in the student ranking. Such ranking confirmed what the teachers who participated in the duplicate test experiment had indicated that they found the multiple-choice item construction difficult. Apart from this fact, the ranking also demonstrates the influence of examinations on classroom practices. Essays and short-answer items are weighted more than the multiple-choice items in national examinations. By focusing more on these item formats, teachers must have believed to offer a better preparation of their students than when they concentrated on multiple-choice items.

4.6 Conclusion

The results in this section have provided some information that the duplicate test forms generally covered the same content as evidenced by non-significant mean difference between the two tests. The variance and mean difference computations also confirmed what Cronbach (1971) had expected for the duplicate test construction experiment. Evidence for the impact of assessment of instruction and classroom practices has also been presented. The next chapter attempts a detailed discussion of these findings in relation to research questions.

CHAPTER 5

DISCUSSION AND CONCLUSION

This chapter draws together the results presented in the previous chapter in view of the research questions that were investigated. In doing this, the chapter attempts to establish common links for the data of different sources to address the major concerns of the validation study. The most important objective of this evaluation was to investigate the extent to which independently constructed MSCE History mock and national examinations favorably compared in content and cognitive representation of the MSCE syllabus and their test specifications. Specifically, this called for the comparison of the tests in terms item topic weighting as provided in test specifications, examining the mean difference relationship with the sum of split-half error variances, and evaluating the mean difference t-test for dependent samples to reject or not to reject the hypothesis of the difference being equal to zero. In addition, the items in the tests were rated for clarity, difficulty, and language appropriateness by SMEs and examinees the results of which were compared to determine the quality of each test in relation to the others. Information about student preparation provided an evaluation of the extent to which national examinations influenced classroom practices.

The chapter is organized in four sections, with each section answering a particular research question or proving plausibility of an assumption. The first section, 5.1 discusses equivalence of the independently constructed tests in content representation by examining statistical evidence, SMEs' ratings and test specifications, and the problem of cognitive ratings is also discussed in section; section 5.2 discusses overall quality issues that SMEs

and students rated, and the impact of examinations on classroom practices is also discussed in this section; 5.3 discusses implications and recommendation to policy makers and future research. Finally, section 5.4 provides conclusion to the validation study.

5.1 Statistical Equivalence of Independently Constructed Duplicate Tests

The duplicate test construction experiment assumed that test constructors who have similar skills and use same guidelines to construct items from the same frameworks would draw up equivalent tests. To satisfy this assumption, the participants' test construction skills were boosted through training in multiple-choice construction, item analysis, and item review. This training standardized the skills for the test constructors to enable them construct items that would be comparable in quality. Student preparation data were also analyzed to establish common levels on which the score comparison was based and determine the impact of examinations on instruction and classroom assessment practices.

5.1.1 Content Validity Evidence from Statistical Equivalence of Duplicate Tests

The traditional method of judging the fit of items to objectives of the content being measured as specified in the test blueprint does not answer the question of whether the stated objectives are representative of the domain of performance to which the test user wants to make inference (Crocker & Algina, 1986). It is this problem that led Cronbach (1971) to propose a duplicate test construction experiment as a more suitable approach to content validity to evaluate both item and objective representativeness. Using

this approach the study found that the History mock tests that were independently developed were content valid. The mean of their squared difference scores was 21.47, which was less than the sum of their split-half error variances of 24.61. These results were consistent with the true score theory's expectation for parallel tests which must have the mean of squared difference scores equal to the sum of their error variances, illustrated as $\sum(X_1 - X_2)^2 / N = \sigma_{e1}^2 + \sigma_{e2}^2$ where X_1 and X_2 are scores on the two parallel forms. According to Crocker and Algina (1986) evidence that "the objectives measured by the test are not idiosyncratic to some perception of the domain that is uniquely held by a single team of test developers" (p. 222) is established when this expectation is satisfied. Crocker and Algina further recommend that when sample statistics are used a ratio to estimate the relationship of parallel tests could be expressed as $\frac{\sigma_{e1}^2 + \sigma_{e2}^2}{\sum(X_1 - X_2)^2 / n}$. When this ratio approach approaches 1.00 then the independently constructed tests could be as similar in terms of error variance as the split-half tests constructed by a single team. In the current study, this ratio for the full tests was found to .80, it was therefore justified to proceed with the split-half error variance computations that have been reported here.

These results are particularly important for the Malawi context because teams of test constructors change quite frequently and common test specifications are not available. Because the two teams independently constructed the test specifications by using the same syllabus and common test construction guidelines, the equivalence of the tests suggest agreement over the aspects of the History content that were considered important for assessment. It is also an indication that the MSCE History syllabus is a

clear operational definition of the content domain because independent teams generally interpreted the important objectives in a similar manner.

The results were supported by the t- test for the dependent sample mean, which indicated that the difference between the means of the scores of the independently constructed tests was not significantly different from zero. This meant that the hypothesis $Ho: \mu_1 - \mu_2 = 0$ could not be rejected because the two tests essentially would rank persons in a similar manner, as the persons would obtain scores that were within the limits of the standard error. The two tests also had very small differences in coefficients alpha and standard errors of measurement, which were .01 and .04, respectively. Therefore, the independently constructed tests were considered of equal difficulty. This would give compelling reason to hold decisions as comparable when they are based on the results of either of the tests. However, considering the limitations that were cited earlier in chapter 1, these results may only hold for the multiple-choice section of the MSCE History examination. It should also be mentioned that accurate data would be required for the analyses of this section to provide correct findings. This point is further discussed in the recommendation section of this chapter.

5.1.2 Test Specifications and SMEs' Content Ratings Validity Evidence

From both the test specifications and SMEs' relevance ratings additional evidence for content validity of the two duplicate tests was found. The test specifications for the independently constructed duplicate tests ensured 100 percent coverage of the MSCE History syllabus. This was confirmed by the SMEs' content relevance ratings that showed that the test constructors followed their specifications when building the tests.

Similar observations were made on the MSCE 1A, which was found to have covered 80 percent of the syllabus. Despite this slight difference, it was concluded that the three tests adequately covered the MSCE History content domain. However, the MSCE 1A's failure to cover the whole syllabus is a matter of concern for the national examination of high stakes nature. It is even of greater concern that MSCE 1A concentrated 40 percent of the items on two topics only. Such selective syllabus coverage should account for the narrowing of the curriculum that teachers actually teach. This also means that students would pass the subject by concentrating their attention on fewer topics only than expected. Nitko et al. (1998) observed that a student's score could easily be influenced by the few topics on which most of the questions tend to focus. Promoting the attainment of a pass score from a limited coverage of content can undermine the need for content valid assessments.

Such a practice can also adversely affect the curriculum as teachers only pay attention to those topics that are assessed (Mucheru, 2003; Steeves, et al., 2002; Shepard, et al., 1992). The practice should therefore, help to explain why teachers did not cover most of the syllabus by the time of the duplicate test experiment but all teachers reported to have covered the very topics in which the MSCE 1A had the largest allocation of items. It was also observed that the same topics had relatively higher weighting in the test specifications of the duplicate tests. This observation suggests that over the years MANEB has been concentrating items on those topics and teachers have learned that pattern, hence its replication in their tests. These results are worrying because they demonstrate the negative effects of high stakes examinations on classroom practice. For both teachers and students, the History syllabus of instruction was getting reduced to

those topics that are frequently assessed. This explains why students remained confident of getting good passes despite confessing that they were not well prepared in terms of syllabus coverage.

However, the fact that the teams of test constructors followed the test specifications that they had set is pleasing for Malawi. It would suggest that if common specifications were used, more comparable content coverage across tests would emerge. Although there were agreements among the teams about what to include in the tests, the presence of stray items that accounted for 5 percent of the total number of items in each of the three tests points to looseness of the syllabus in some parts. This problem was common on items that dealt with the gold trade and Portuguese in Central Africa. The confusion seems to have arisen because the curriculum review that was recently carried out removed a topic on the Portuguese Influence in Central Africa and retained the topic on Trade in Gold, Ivory and Slaves. Unless clarification of objectives for the topic of Trade in Gold, Ivory and Slaves is made, this part of the content domain would remain contentious for the teachers and test designers because the Portuguese were among the players in the gold and slave trade in Central Africa. For the items that were completely out of syllabus as the question on Italy on the MSCE 1A, it could be avoided if test reviewers did more than just editing of the items. A serious quality review is needed before a test is passed for administration.

Looseness of the syllabus can lead to serious variations in the samples of items drawn by the tests on the same domain even in the presence of written test construction guidelines. Cronbach (1971) observed that when the content domain is not accurately

defined only those aspects that both test constructors and reviewers consider to be important would be assessed. And he writes

> Test construction is no better than the reviewers of items. If there is a tacit assumption, on the part of everyone who participates in the test construction, that a certain topic is of negligible importance, items on that topic will not appear even though it falls within the universe definition. A common blind spot is almost impossible to detect. (p. 456)

This means that despite standard guidelines that can be given for test construction, test constructors' belief system would influence decisions about what goes into the test more than the written guidelines. One way to control the designers' influence on the weighting of topics is to use a common table of test specification for different test construction teams. Otherwise, it is reasonable to conclude that item/topic representation cannot be ensured unless the content domain is precisely defined and clear construction guidelines are provided.

5.1.3 Lessons from the Cognitive Skills Relevance Ratings

Serious disagreements between test designers and SMEs were observed regarding cognitive skills that the items in the tests measured. The SMEs did not identify any item for the higher order categories of Application, Analysis, Synthesis, and Evaluation. This raised three issues that are equally important for advancement of testing in Malawi. Firstly, the lack of higher order items in examinations promotes rote memory and can direct instruction to simple drill and memorization of facts. This is likely when all that is assessed is indeed a reflection of factual knowledge presented in instruction or books. Considering the nature of History as a subject, this is a possibility because history is dominated by facts of what actually happened and why it happened. It is possible that the

SMEs' saw all items as simply imploring examinees to recall and demonstrate some interpretation of what they learned. This explanation however, is contrary to the argument that you can assess higher order abilities with multiple-choice in any subject matter. Gronlund (1968) argued;

> Objective test items can be used to measure a variety of complex learning outcomes ranging from the lowest level of understanding… to the higher levels of intellectual skills. Both simple and complex items are useful… and various adaptations can be designed to fit a particular learning outcome. (P.51)

It would therefore, be worrying if SME's rating on cognitive domain genuinely meant absence of items measuring higher order abilities considering the amount of training that MANEB conducts for its examiners. Lack of such items would seriously undermine the educational objective of producing a creative and self-reliant person as advocated in the goals of education in Malawi (Ministry of Education, Science and Technology, 2000). The problem requires further investigation so that examinations could be improved to make the required impact on curriculum implementation.

Secondly, absence of higher order abilities items may simply be an indication of SMEs' failure to recognize such items in the tests. Because the majority of the SMEs were practicing teachers they must have found the rating scale to be too detailed and too specific that it looked divorced from reality. It is possible that if the scale lumped the higher order abilities together some items would have been properly classified. But SMEs' inability to distinguish higher order abilities could be a serious limitation in their own teaching of the subject for which they are considered experts. When both the national examinations and classroom tests put emphasis on knowledge and comprehension only, teaching experiences may routinely focus on transmission of factual

information. Should this speculation be the right explanation of what was experienced in this study, then something needs to be done to improve teacher and test developers' performance otherwise examinations will continue to influence instruction and classroom assessments negatively. An investigation involving curriculum specialists and the testing agency could be the starting point toward finding a solution to this problem. This is important because one syllabus is used for teaching and assessment. Teachers may need to be guided by spelling out objectives that target higher order abilities and the examination should be seen to assess such abilities.

Finally, SMEs' inability to recognize items that measured higher order abilities could be a reflection of a general assumption that multiple-choice items do not measure higher older abilities. The multiple-choice items have often been resisted in Malawi on grounds that they do not measure students' mastery of knowledge appropriately. There is also the general belief that objective type items are inferior to constructed response items. It is therefore possible that what was observed in the cognitive relevance ratings was an expression of teachers deep conviction of multiple-choices' inability to assess higher order abilities. Had it been that the rating was carried out in the United States where multiple-choice items are common in assessment, different results would most likely have emerged.

5.2 Construction of Equivalent Duplicate Tests

Content validity is often ensured by the decisions that test constructors make when they select items from different topics to build a test. For this reason, evidence of such decisions from the two independent teams would meaningfully be the basis for

comparison on how well the two tests tapped from the same content domain. While the MSCE syllabus described the content domain well enough, the decisions of what to include in the instruction and assessment so much depends on the belief systems of the instructor and test developer (Nitko, 1989). Because the duplicate tests were constructed in teams, comparing of SMEs and students' comments on the quality of the tests, evaluated the extent to which the belief systems affected test quality.

Evaluation of test quality also demonstrated the extent to which teacher participants used the item development skills they had acquired from the test construction experiment. This finding is important in view of the fact that a policy of school based assessment has been proposed, yet teachers' poor testing knowledge and skills are among major limitations for its successful implementation (Selemani-Mbewe, 2003). The results demonstrated that it is possible to upgrade assessment skills in teachers through a well-focused short workshop as was conducted in the study.

5.2.1 SMEs and Students' Perceptions about Test Quality

Considerable agreements were observed between SMEs' and students' ratings over the quality of items in terms of clarity, language appropriateness, and test difficulty. The general impression was that all the three examinations were able to meet the expected standard of the MSCE examinations in terms of overall quality. This was pleasing because it supported the proposition that if test developers were well trained they would independently produce tests that could be comparable in quality. It is therefore, concluded that the independently constructed duplicate tests were not only content valid but also of acceptable quality for the MSCE examination level. Since

examinees rated them as satisfactory in terms of clarity and level of difficulty it is also concluded that the duplicate tests, like the MSCE 1A satisfied face validity.

It was surprising however, that even the items that were compiled from the classroom tests were rated as good as those constructed the trained writers. This surprise is in view of the fact some of the duplicate test developers indicated that they were not able to construct their own multiple-choice items prior to participating in the study. Therefore, the samples of items in class tests seem to suggest that there is a lot of recycling of previously administered items in the schools. Most of such items must have been drawn from the national examinations past papers. Such a dependence on national examination items could be a serious threat to the security of MANEB's item bank. Item exposure can negatively affect test validity as fairness in assessment could be violated when some examinees have prior knowledge of the items. Future research could investigate the extent to which previously used items have been exposed in the schools.

The use of students in evaluating test quality was unique to this study. Test content validation quite often involves only SMEs in matching items to content and cognitive skills domains. Such a limited evaluation leaves out important characteristics of items that affect student performance. In Malawi where English is a second language but is used in assessment, item clarity and language vocabulary can affect validity when respondents are not able to grasp what the item is assessing. The characteristics of individual items in each test could determine the way students would perform in relation to a given criteria. Sireci, Wiley, and Keller (2002) stated:

> Items that violate item guidelines may undermine test validity in ways that do not show up in item statistics. For example, some items may facilitate anxiety. Others may take longer to answer. If these items appear earlier on the test, the smarter test takers may still

> answer them correctly, but their performance on later items may be affected. (p. 25)

A better understanding of these item characteristics could be reached if information is obtained from examinees who responded to the items in an examination condition. For the validity of independently constructed duplicate tests, these characteristics were considered important to discern differences that could have led to score variations in examinees. Therefore, information was obtained from both SMEs and examinees regarding such qualities as level of language vocabulary used in the test, clarity of the items in terms of describing the tasks, and overall test difficulty.

These results added value to the conclusions that have been reached regarding the validity of the duplicate constructed tests. By rating the items as satisfactory in terms of clarity, language vocabulary level, and difficulty it meant that no factors other than the test content explained the results that were obtained. With the duplicate test construction approach to content validity, it is not only the selection of the materials and the universe description of the content that matters, but also the persons whose scores would be used to compute the statistics that provide evidence for the content validity.

5.2.2 Impact of Examinations on Classroom Practices

Consideration of what teachers believe to be important is necessary whenever teacher made tests are evaluated. Unlike external tests, teachers develop tests that would link with instruction as a way of getting feedback on the learning that has taken place. Ensuring this goal would be the utmost expectation of the teachers who developed the Mock examination because the examination was meant to inform them about how well their students were prepared for the national examination they expected. Failure to link

assessment to instruction would lead to undesirable consequences such as inappropriate information about students' learning progress and difficulties, reduction of student motivation and unfair decisions regarding effectiveness of instruction (Nitko, 1989). As the last internal examination before the national examinations, mock examination does not only check on student readiness and instructional adequacy, but also helps learners' switch to "intention to remember mode" (Hopkins, Hopkins, & Stanley, 1990), as the national examination approaches.

While teachers rated as "very important" all topics except the subtopics of the League of Nations, Problems of Versailles Treaty, and Iron Age, it was surprising that they did not cover the whole syllabus in their instruction with the same belief system. It was however, observed that some topics were more popular than others because all teachers covered such topics regardless of the amount of topics that had been taught. This showed the tendency for teachers to prioritize some topics during instruction. On the whole, fewer topics than would have been expected were covered by the time of the mock examination. Despite this inadequate coverage, students were optimistic about getting good grades in national examination. Similar inadequacy was reported for topics in which class tests were administered across schools. This seems to suggest that teachers did not feel pressurized to ensure adequacy of syllabus coverage during instruction. Such an attitude looks strange in view of the high-stakes nature of the MSCE examination. There seem to be no motivation for teachers to meet syllabus obligations for their students to pass national examination well enough.

As noted earlier, however, teachers spread allocation of items on the mock examinations throughout the syllabus to match with their expectation of the national

examination. This was seen necessary to achieve equivalence with the national examination and to accord them an opportunity to touch on untaught topics of the syllabus when revising the mock tests. The fact that teachers could not care about completing the syllabus during instruction and still expected their students to do well in the national examination is another indication of the looseness of the syllabus and assessment regulations in History.

Apart from the MSCE History 1A, which covers the whole syllabus because of multiple-choice items, the MSCE History 1B comprises short answers and essays, that are topic focused. In this paper, examinees are given options on equally weighted questions. Such an assessment outline would enable testwise teachers to drill their students on the topics that seem to be favored by the test designers. which could make their students pass the examination without necessarily attaining the expected mastery of the content. For a high-stakes achievement examination as the MSCE, such looseness in assessment regulations and the predetermined overloading of items on favorable topics can be detrimental to effective instruction. Mucheru (2003), Jones and Johnson (2000), and Steeves et al. (2002) observed the tendency for teachers to ignore curriculum stated outcomes and concentrate only on those objectives that are more frequently assessed. It is likely that examinations in MSCE History are contributing to negative consequences on instruction, which might have reduced to drill and coaching. This should explain why both teachers and students did not seem anxious that the syllabus had not been fully covered. As long as the popular topics were taught their expectations for high achievement were certain.

It would appear that over the years MANEB test designers have selected more items from topics like Causes and Results of the First World War and Interwar Period so that all teachers now ensure their coverage. This is an indication of an assessment driving instruction. And when the MSCE History examinations have an overload on the most popular topics, as was the case with the MSCE History 1A, then the examination drive would lead to curricular reduction in the manner observed in the study. For a high-stakes examination on which selection to University is based, such a reduction is a serious matter because those who pass the test may not attain the assumed mastery of the subject matter. A mismatch with an external criterion that would determine future success in the subject would most likely result in a predictive validity study for university or any other future success that assumes mastery of MSCE History syllabus content as a required prerequisite.

5.3 Implications and Recommendations of the Study

The results of this study have touched on a number of issues that have implications on assessment and curriculum implementation. The following are key issues that may need attention to improve education quality through assessment:

i. The MSCE History syllabus is well defined for different test construction teams to produce statistically parallel tests. However, the intrusion of 5 percent of items from outside the syllabus requires tightening up demarcations of some content domain areas to make them more understandable than is currently the case. Otherwise, it can undermine efforts to produce equivalent examinations even when same test construction guidelines are used because test designers' would be

directed mostly by their belief systems in such situations. For the MSCE History syllabus, the problem of poorly defined content areas is aggravated by the directions of the MSCE History 1B which limits syllabus coverage by offering options to examinees. As a result of this problem, persons with different mastery of the subject matter may end up with similar grades from one year to another. This raises a validity question of whether the MSCE examination correctly classifies persons in various categories of mastery and whether success in MSCE can be used to predict future success despite the year the exam was taken.

ii. Test designers' concentration of items on specific topics can have negative effects on curriculum implementation in that teachers would earmark such topics and ignore other topics that do not seem favored by MANEB. This is harmful because it can guarantee success out of little effort as testwise teachers would read the examination content patterns and concentrate instruction on the popular topics only. It appears that such a practice has already entered the education system in Malawi as fewer than half of the topics had been covered by the time of the test construction experiment.

iii. Training of teachers in test construction can improve the quality of classroom assessment practices as teachers gain competence to link assessment to instruction. Mkhonta (2003) observed that success and superior achievement of students would result when teachers become competent in effective use of continuous assessment. The quality of the tests from the duplicate construction experiment was high in terms of clarity, and item distribution over the syllabus. It was also evident that teachers felt confident to do better in assessment after the

training session because they felt that useful skills had been acquired. Teachers' confidence in assessment would be boosted, if teachers were exposed to assessment skills through in service activities. This could reduce teacher anxiety, which has been the cause for teacher collaborated cheating in some schools.

iv. The presence of errors in teacher-produced scores was more than a sign of carelessness on the part of the teachers. It was an indication of how serious such errors could affect the decisions that are reached from examination results. Although MANEB has always tried to minimize scoring errors by deploying script checkers during live marking sessions, MANEB (1999) observed that some errors go unspotted. This calls for the use of machine scoring for multiple-choice items and thorough supervision of constructed response scoring so that accurate scores become the basis on which decisions are made.

v. Failure to produce or recognize higher order abilities poses a threat to well intended educational objectives aimed at promoting creativity and problem-solving techniques in schools. It appears that there is emphasis in only knowledge and comprehension skills in both instruction and assessment. The curriculum may not have treated higher order abilities adequately or the training of the teachers has not brought the cognitive skills to the practical level for the teachers in practice. This is another area that requires further investigation so that a suitable solution to the problem could be devised.

In view of these policy implications, it should be reasonable to make the following recommendations:

i. To improve content coverage and ensure comparable content representation, MANEB should use a pre-determined test specification, which would provide weighting that could systematically sample the content domain. Such specifications should focus on important course objectives to ensure a reasonable balance of cognitive skills. In addition, MANEB should provide the same and clear guidelines to trained test constructors. When each test construction team draws up its own test specification, it is difficult to come up with tests that can be equivalent, because each team will be directed by the belief systems that are shared by its members. Smisko, et al. (2000) observed that a systematic test development procedure ensures strong evidence for both content and curricular validity.

ii. MANEB should critically review items to ensure that test papers contain a proportionate number of items as stipulated by a test specification grid. When a national examination includes higher order items it is most likely that the examination would have positive impact on education quality. Future research should investigate teachers' awareness of the different categories of the cognitive domain and how much attention they give to those categories in instruction and assessment. Such a study, could inform on the type of training that teachers need.

iii. Considering the high-stakes nature of the MSCE examinations and observed teachers' lack of competence in assessment skills, it is appropriate that some in-service training in test construction and use of assessment in instruction be given to teachers to improve curricular implementation. Such training should emphasize the role of assessment in instruction so that teachers are able to see a link between

classroom practices and examination expectations. In this way, assessment and curriculum will be integrated to the extent that learners would see the results of their performance as a springboard for further learning (Mchazime, 2003).

iv. There is need to upgrade assessment skills in teachers to ensure improvement in the quality of education being offered. Failure to use higher order abilities of the cognitive domain is both a pedagogical and psychometric concern. An education system that ignores thinking or problem-solving skills cannot produce a self reliant or productive citizen as stipulated in the education policies (Malawi Ministry of Education, 2000). Teacher training institutions need to ensure practical translations of the Bloom's taxonomy for teachers to gain useful understanding of the cognitive domain.

v. A follow up duplicate construction experiment would be required for the MSCE History 1B to evaluate the extent to which different teams could come up with similar tests when constructed response items are used. This would be necessary to check the extent to which the results of the current study could be generalized to all examinations in MSCE History. Another content validity study in MSCE History should be on tests over a number of years to further examine the problem of curriculum looseness. Such a study should focus on topics that are often left out and those that are often responsible for stray items. Information from such a study would guide appropriate reforms in curriculum and assessment. An immediate action would necessitate the review of the objectives that guide both instruction and assessment. The objective of such a review would be to remove the ambiguities that could have been responsible for stray items.

5.4 Conclusions and Summary

The major purpose of this study was to evaluate content validity by comparing results on two independently constructed tests. The teams of test constructors worked independently in drawing up test specifications, constructing multiple-choice items, and selecting the items for the final test. Test constructors were trained in item construction editing and reviewing, and the use of item analysis data to improve items. The two teams comprised of teachers who were teaching the same form in six different schools and they used the same syllabus to construct items for MSCE History. These characteristics in the experiment established a basis for comparison of the two tests whose scores from the same students were analyzed for significance of the mean difference, and correlation difference.

The hypothesis $\mathrm{H}o : \delta = \mu_1 - \mu_2 = 0$ was not rejected because the computed t-statistic of the mean difference was less than the t-critical. The difference between the mean of Mock A and Mock B was found to be not significantly different from 0. A comparison of the mean of squared difference scores and the sum of the split-half error variances also supported the rejection of the hypothesis. This means that the two tests would yield similar results within the limits of the standard error of measurement when exchanged. It was therefore, concluded that the duplicate constructed tests were statistically equivalent because they tapped similar content from the MSCE History syllabus.

Information on cognitive rating was quite unsatisfactory because the rating departed greatly from the propositions set in the table of specifications. Should this be a result of raters' inability to distinguish the cognitive skills, it means that the evaluation

did not effectively articulate the levels of difficulty of the tests. The distinction of the tests based on the skills measured would have provided more information about test difficulty to supplement what students said about the quality of the tests.

This study has brought up a number of revelations that should be followed up to improve assessment practices. Rather than providing answers it has pointed out where problems could be and how they could be dealt with to improve assessment procedures. One suggestion is the need for research to find solutions that can appropriately direct improvements in examinations. While this need can be appreciated, research is often not among the priorities of policy makers in developing countries, such as Malawi. Even in examinations agencies like MANEB, administrative functions are given more attention than psychometric research activities. Apart from problems of funding research, data collection from testing agencies can be difficult because of confidentiality and secrecy obligations. Researchers from outside the organization are often treated with skepticism regarding the use of the data from examinations. It is the researcher's view that greater collaboration of those involved or interested in assessment could bring out valuable information that can help MANEB to improve examination practices.

APPENDICES

APPENDIX A

SET 1 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MOCK A)

1. The discovery of glass beads, and cowrie shells in the Iron Age sites of Central Africa tell historians that Iron Age settlers practiced
   A. internal trade.
   B. iron smelting.
   C. fishing.
   D. long distance trade.

2. Archaeological remains in the Iron Age sites of Central Africa CANNOT tell us about the people's
   A. migration routes.
   B. political organization.
   C. trading activities.
   D. material culture.

3. How do archaeologists tell that there was a general North to South movement of Iron Age people in Central Africa?
   A. through appearance of pottery
   B. through skeletal remains excavated
   C. through the dating of artifacts
   D. through trading activities

4. The following were duties of the councilors in the Maravi Empire EXCEPT
   A. acting as messengers.
   B. collecting tribute.
   C. consulting ancestral spirits.
   D. acting as judges.

5. Why did Nkhamanga kingdom decline?
   It was because of
   A. sorcery powers.
   B. slave trade.
   C. vastness of the empire.
   D. the arrival of missionaries.

6. Why did King Manuel of Portugal send Francisco D'Almeida to the East African Coast?
   A. to improve the gold trade
   B. to preach the gospel to the Portuguese and local people
   C. to win the support of the coastal rulers
   D. to avenge the death of Father Silveira

7. Why did the influence of the Portuguese decline in Central Africa after 1600?
   A. Portuguese were few in number.
   B. Swahili-Arabs defeated the Portuguese.
   C. There was competition between the Portuguese and the British.
   D. Their government did not assist them.

8. Why were some Europeans in Northern Rhodesia opposed to the Hilton Young Commission?
   A. Because of Southern Rhodesia's racial policies.
   B. Because Sir Young wanted Northern Rhodesia's mineral resources.
   C. Because they feared loss of large tracts of fertile land.
   D. Because they feared an influx of Europeans from Southern Rhodesia.

9. Which of the following is NOT true about why Africans in Southern Rhodesia supported the federation?
   The federation
   A. was more liberal than their own territorial governments.
   B. brought them closer to the Africans in other territories.
   C. profited the Africans from economic expansion in their own country.
   D. promised to share out land to Africans.

10. What led to the dispersal of the Jere Ngoni at Mapupo?
   A. succession disputes
   B. lack of pastureland
   C. attacks from neighbouring ethnic groups
   D. effects of slave trade

11. Why was Nkulumane enthroned as a Ndebele king at Gibihxegu?
   A. Because he had bribed the Indunas.
   B. Because Mzilikazi was feared dead.
   C. Because he was Mzilikazi's eldest son.
   D. Because Lobengula was still young.

12. Why did the Ndebele abandon Mosega?
   A. prevalence of tsetse flies
   B. inadequate water
   C. frequent raids
   D. fear of Tchaka's attacks.

13. What important event occurred at Thabasinduna?
    A. Ndiwere declared himself king.
    B. Mzilikazi killed some indunas.
    C. Ndiwere installed Nkulumane as king.
    D. Inyati was declared the capital.

14. Why did the Chewa in Dedza-Ntcheu area welcome back the Maseko Ngoni?
    A. The Chewa admired the Ngoni culture.
    B. The Chewa admired Ngoni political system.
    C. They wanted to be protected against Yao slave traders.
    D. The Chewa wanted to learn Ngoni fighting tactics.

15. What factor explains the reason why the Yao accepted Islam?
    A. The Yao's traditional belief in a supreme god was similar to Islamic beliefs.
    B. Both used charms in their religion and hence attracted to each other.
    C. The Yao supplied slaves to the Arabs hence were close to each other.
    D. Both emphasized on cleanliness and therefore were more connected to each other.

16. Which areas in Malawi were greatly influenced by Islam before 1800?
    A. Dedza, Karonga, and Mangochi
    B. Nkhotakota, Salima, and Karonga.
    C. Mangochi, Salima, and Mulanje
    D. Nkhotakota, Mulanje, and Dedza

17. The main reason why the Livingistonia Mission was shifted from Bandawe to Khondowe was to
    A. stop slave trade at Khondowe
    B. convince the Ngoni to accept colonial rule
    C. fight against Ngoni raids.
    D. settle in an area free of mosquitoes

18. Which reason explains best why the Catholic Church abandoned Nyasaland in 1889?
    A. The area they occupied was unhealthy.
    B. There were few missionaries.
    C. They were disturbed by slave traders.
    D. They were accused of pursuing political goals.

19. What is the relationship between the statement and the reason given below?

Statement	Reason	
Chief Mazengera invited the DRC mission at Nkhoma	because	he wanted protection against slave raids.

A. The statement is true but the reason is false.
B. The statement is false but the reason is true.
C. Both the statement and the reason are true and the reason is the best explanation of the statement.
D. Both the statement and the reason are true but the reason does not explain the statement.

20. What was agreed in the Rudd Concession of 1889?
A. Europeans promised to grab African land.
B. Lobengula granted exclusive mineral rights.
C. Europeans gave Lobengula trade monopoly.
D. Lobengula allowed Europeans to control the Zambezi.

21. The following were results of the 1893 Ndebele War EXCEPT
A. The Ndebele submitted to the B. S. A. Company.
B. Lobengula was captured.
C. Gwaai and Shagaan were declared reserves.
D. The B. S. A. Company took Lobengula's cattle.

22. Why did the whites in the 1896-97 revolts easily defeat the Ndebele?
A. The Ndebele were fewer in number.
B. The Ndebele used inferior weapons.
C. The Ndebele army commanders were disorganized.
D. Their religious leaders deserted them.

23. Why was the British occupation of Nyasaland regarded purely humanitarian.
A. Nyasaland had many slave traders.
B. Nyasaland had many poor people.
C. The British wanted to protect their missionaries.
D. The British wanted to end ethnic conflicts in Nyasaland.

24. Why was the B. S. A. interested in occupying Matabeleland?
A. It wanted to extend the rail line to Mafeking.
B. It needed more land for mining.
C. It wanted to gain more pasture land.
D. It had experienced hardships in Mashonaland.

25. Which of the following statements explains the reasons for African opposition to the Federation of Rhodesia and Nyasaland?
    A. Africans feared an extension of the labour migration policies.
    B. Africans feared that the federation would encourage cross border trade.
    C. Africans feared that the federation would encourage racial segregation policies.
    D. Africans feared that they would not be allowed to participate in economic activities.

26. Which of the following is a positive impact of labour migrations in Nyasaland?
    A. It improved the health of the immigrants
    B. The immigrants began to use mining skills gained from abroad.
    C. It led to the spread of rural economies.
    D. It strengthened marriages of the immigrants.

27. Why did Kaphwiti run away from Mankhamba?
    A. He was accused of witchcraft.
    B. He quarreled with local priests.
    C. He was encouraged to leave by the Portuguese.
    D. He had the desire to control long distance trade.

28. How were the voluntary associations and independent churches similar?
    A. They both organized mass rallies.
    B. They both had same leaders.
    C. They both had political objectives
    D. They both used force against whites.

29. What did the Chewa of Nkhotakota benefit from their acceptance of the Jumbe?
    A. He taught them modern methods of fishing.
    B. He protected them from the Ngoni raids.
    C. He signed a trade agreement with Malenga Chanzi.
    D. He introduced maize to the area.

30. Why did Charles Chinula breakaway from the Livingistonia Mission to establish his own church?
    A. He was not ordained as a church minister.
    B. He was accused of speaking ill of the colonial government.
    C. He embezzled church funds.
    D. He was against the teachings of the church.

31. The building up of an empire by seizing territories overseas is called
    A. nationalism
    B. communism
    C. imperialism
    D. capitalism

32. What was the name of the gunboat that Kaiser William II sent to the Moroccan port of Agadir?
    A. The U-boat
    B. The Panther
    C. The Cruiser
    D. The Dreadnaught

33. How did the Dual Entente make a two-frontal war against Germany almost certain?
    A. Russia and France were now on one side.
    B. Britain and France had come together.
    C. Austria and Russia had buried their differences.
    D. Serbia and French troops were advancing towards Germany.

34. Why did Germany choose Austria-Hungary as an ally?
    A. Austria was easier to control.
    B. Austria would help Germany to expand trade.
    C. Austria contained some German-speaking people.
    D. Russia had refused to become an ally.

35. Why did the Schlieffen Plan fail to materialize?
    1. Belgium put up a gallant fight.
    2. Transportation of heavy weapons was difficult.
    3. Russian mobilization was too fast.
    4. Rains made roads through Belgium impassable.
    A. 1 and 2 only.
    B. 1 and 3 only
    C. 2 and 4 only
    D. 3 and 4 only.

36. The League of Nations partly failed to solve international disputes because
    A. The British government rejected it.
    B. The Ambassadors Conference overruled its decisions.
    C. Its council had powers to veto decisions.
    D. France never contributed troops for operations.

37. Why did the British government partition the sub-continent of India prior to granting of independence in 1947?
    A. It was because they considered it too vast to be one state.
    B. It was because of clashes between Hindus and Moslems.
    C. The British enjoyed the policy of divide and rule.
    D. The British were trying to make each race rule itself.

38. Which other country emerged out of the portioned subcontinent of India in 1947?
    A. Kashmir
    B. Punjab
    C. Bangladesh
    D. Pakistan

39. Which of the following contributed to the growth of nationalism in Kenya?
    A. labour migration
    B. serious shortage of food
    C. massive unemployment
    D. land alienation

40. What was the immediate result of Bismark`s dismissal by the Kaiser William II in 1890?
    A. The Dual Alliance became stronger.
    B. Austria- Hungary became happy.
    C. The Reinsurance Treaty was not renewed.
    D. Britain and France were happy, as Germany had become weak.

41. Why was Serbia interested in the Balkan region?
    A. She wanted to assist other Slav regions and create a Slav nation.
    B. She hoped to crush Turkey and have access to the Mediterranean Sea.
    C. She wanted an alliance with Romania in order to crush Bulgaria.
    D. She wanted Russia to have colonies in the Balkan region.

42. Why is the Versailles Settlement blamed for the outbreak of the Second World War?
    A. The defeated countries were not allowed to attend the Versailles Conference.
    B. Victorious powers were allowed to increase their armament.
    C. The Treaty terms were very harsh to Germany and other defeated nations.
    D. The unfairness of the Versailles Treaty encouraged the defeated nations to go against it.

43. Which of the following events contributed to the outbreak of the Second World War?
    1. Spanish Civil War
    2. Italian invasion of Ethiopia in 1936
    3. German invasion of Poland
    4. German re-occupation of the Rhineland
    A. 1, 2, and 3
    B. 1, 3, and 4
    C. 2, 3, and 4
    D. 2, 1, and 4

44. Which of the following statements is NOT true about the results of the Second World War?
    A. The U. S. A. continued with the policy of isolation.
    B. Communism spread in Eastern Europe and South East Asia.
    C. Korea and Germany were both divided.
    D. There was destruction of life and infrastructure.

45. Why were there many military takeovers in African countries soon after the countries got their independence?
    A. The soldiers believed that they were more disciplined and better leaders than the civilian leaders.
    B. The people and soldiers were dissatisfied with the inefficiency and corrupt practices of the civilian leaders.
    C. The soldiers believed that the elections for the civilian leaders were rigged.
    D. The soldiers had strong desire to rule their countries after they had fought for their countries' independence.

46. Which of these statements is NOT true about the expectations of Africans at independence?
    A. Africans would be economically independent.
    B. Independence would bring national unity.
    C. Africa would become one nation.
    D. Africans would dominate white supremacy.

47. Which of the following factors are true of the Great Depression of 1929?
    1. More goods were produced than could be consumed.
    2. Foreign trade was killed by the high tariffs set by the American government.
    3. Easy Credit facilities had led to wild speculation.
    4. Agriculture was ignored thereby resulting in declining output.
    A. 1, 2, and 3
    B. 1, 2, and 4
    C. 1, 3, and 4
    D. 2, 3, and 4

48. What was the policy of the League of Nations over the mandated territories?
    A. To be governed by the European mandatory powers forever.
    B. To be ruled directly by the League's Secretariat.
    C. To be ruled by European powers until the indigenous people were ready to take over power.
    D. To be governed by the local people under the supervision of the European mandatory powers.

49. What was the aim of the Dawes Plan of 1924?
    A. To make USA pay reparations for Germany
    B. To stop reparations payments
    C. To ease Germany's reparations payments
    D. To reduce the reparation amount

50. Read the following and answer the questions that follow:
    Dear Mr. President,
    This is just to tell you everything is alright now. The man you sent found our house alright and we went down to the bank with him and the mortgage can go for a while longer. You remember I wrote you about losing the furniture too. Well your man got it back for us. I never heard of a president like you…..(Source: L. E. Snellgroove, *Modern World History*).
    What is the speaker referring to?
        A. End of First World War
        B. Germany's praise to Hitler after Reparations
        C. The New Deal and Fredrick Roosevelt
        D. Appeasement to Hitler

51. What sparked the Russian Revolution of November 1917?
    A. People were dissatisfied with Rasputin.
    B. Food shortage
    C. Failure of the Provisional government
    D. Russian defeat in the Russo-Japanese War

52. Which of the following was Hitler's domestic policy?
    1. To undo the Treaty of Versailles
    2. To ill-treat the Jews
    3. To reduction of unemployment
    4. To ban political parties
    A. 1, 2, and 3
    B. 1, 2, and 4
    C. 1, 3, and 4
    D. 2, 3, and 4

53. The following countries financed the FRELIMO military equipment and training EXCEPT
        1. Algeria
        2. Russia
        3. China
        4. South Africa
    A. 1, 2, and 3
    B. 1,2, and 4
    C. 1, 3, and 4
    D. 2, 3, and 4

54. Why was the Nkomati Accord signed between Samora Machel and Fredrick de Klerk?
    1. To allow the logistical flow of goods to neighbouring Zimbabwe.
    2. To call for immediate elections in a multiparty democracy in Mozambique.
    3. To prevent ANC from using Mozambique.
    4. To make Pretoria sever economic assistance to RENAMO.

    A. 1 and 2
    B. 1 and 3
    C. 2 and 4
    D. 3 and 4

55. What was Joachim Chissano hailed for on 3rd November 1990?
    A. His call for immediate end to his country's civil war.
    B. Organized and ready to compete in democratic elections.
    C. Accepted RENAMO demands for direct elections.
    D. Free enterprise economy and freedom of religion.

56. Why did US Congress prohibit the provision of development aid in Mozambique in 1977?
    A. She was concerned with alleged human rights violations.
    B. Her struggle against neo colonialist policies.
    C. She was against remnants of colonialism.
    D. She was for peace and universal security

57. What factor describes the fall of the economy of Zambia before 1990?
    A. Zambia had a heavy burden to host thousands of refugees.
    B. Zambia had been severely taxed in foreign trade.
    C. Zambian government challenge multinational corporations that wanted to invest and create jobs.
    D. Zambian economy depended on copper only at the world market.

58. Which of the following events was the immediate cause of the First World War?
    A. Political alliances
    B. Arms race
    C. International Crises
    D. The assassination of Franz Ferdinard

59. Which countries formed the group of "Allies" that fought during the Second World War?
    A. Austria, Japan, Germany, and France
    B. U. S. S. R, France, U. S. A, and Britain
    C. Italy, Germany, U. S. S. R. and Japan
    D. Japan, Germany, Italy, and Austria

60. The Appeasement Policy resulted in the following, EXCEPT
    A. The renewal of Germany's union with Austria
    B. Germany's occupation of Poland
    C. Germany's occupation of Sudetenland
    D. Germany's invasion of Czechoslovakia.

APPENDIX B

SET 2 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MOCK B)

1. What changes did farming bring about during the Iron Age period?
   1. Permanent settlement
   2. Making of pottery
   3. Hunting of animals
   4. Building of huts.
   A. 1 and 4 only
   B. 1 and 3 only
   C. 2, 3, and 4
   D. 1, 3, and 4

2. Which of the following statements is TRUE about the Tumbuka-Nkhamanga Kingdom?
   A. Immigrants from Zambia founded it.
   B. The founders were known as Balowoka.
   C. The founders traded with the Maravi kingdom.
   D. The prominent clans were Msiska, Mzumara and MKandawire.

3. The factor that prolonged the Shona-Ndebele rebellion of 1896-7 was
   A. the killing of cattle that were attacked by rinderpest.
   B. the African loss of good and fertile land.
   C. the emphasis on religious interpretation of the disaster.
   D. the presence of famine, drought and locust in the region.

4. What was the factor that prohibited Nyasaland African Congress (NAC) from registering mass political support?
   A. Lack of clearly defined political objectives
   B. The activities of the Young Turks
   C. Most of the leaders were uneducated
   D. It was centred on one ethnic group.

5. Why did the Jere Ngoni fail to defeat Mwase of Kasungu?
   A. Because Mwase was assisted by the Jumbe of Nkhotakota.
   B. Because Mwase had adopted the Zulu type military tactics.
   C. Because Mwase had firearms obtained through trade with the Jumbe.
   D. Because Mwase had a very strong stockade at Linga.

6. Where was the headquarters of Makewana's religious shrine?
   A. Mankhamba
   B. Nsinja
   C. Khulubvi
   D. Mbewe wa Mitengo

7. What was the earliest Iron Age site in Central Africa?
   A. Gokomere
   B. Kalambo Falls
   C. Nkope Hills
   D. Ingombe Ilede

8. What evidence is there to show that there was interaction between Iron Age people and the East Coast of Africa?
   A. Fish skeletal remains
   B. Burnt dagga
   C. Slag heaps
   D. Sea shells

9. Which of the following leaders was the FIRST to be titled "Kalonga" in the Maravi Kingdom?
   A. Mazizi
   B. Chinkhole
   C. Chidzonzi
   D. Sosola

10. Why was the 1873 Jon-Barghash Treaty signed in 1873?
    A. To stop the Sultans from exporting slaves to Europe.
    B. To let Britain open clove plantations.
    C. To close slave markets and end slave trade.
    D. To stop the Sultans from exporting slaves to America.

11. Where did the Portuguese obtain gold in Central Africa?
    A. Maravi Kingdom
    B. Mwata Kazembe Kingdom
    C. Tumbuka-Nkamanga Kingdom
    D. Mwenemutapa Kingdom

12. Why did the Yao migrate into Malawi between 1830 and 1860?
    A. They were looking for good fishing area.
    B. They were running away from conflicts with Makuas.
    C. They were running away from ill treatment by Portuguese.
    D. They were running away from Ngoni attacks.

13. Why did San Luiz Catholic Mission fail in its activities at Mpondas in Mangochi around 1891?
    A. Conflicts between the Yao and Mang'anja.
    B. The Missionaries were few in number.
    C. Slave trade and restrictions to meet people.
    D. Poor transportation and food.

14. How did Lobengula realize that he had given away too much in the Rudd Concession of 1888?
    It was through
    A. letter from the Queen Victoria
    B. Dr. Leander Starr Jameson
    C. advice from missionaries
    D. Cecil John Rhodes

15. What name was given to Elliot Kamwana's Independent Church?
    A. Providence Industrial Mission
    B. Zambezi Evangelical Mission
    C. Mpingo Wa Afipa Wamu Africa
    D. The Watch Tower Movement

16. Why was the Nyasaland African Congress (NAC) formed in 1944?
    A. To replace the independent churches
    B. To welcome Kamuzu Banda
    C. To unite Africans for Independence
    D. To resist nationalism in Nyasaland

17. At which Iron Age site were 46 remains of dead bodies excavated?
    A. Kalambo Falls
    B. Ziwa
    C. Ingombe Ilede
    D. Gokomere

18. Which of the following factors led to the decline of the Kazembe kingdom?
    1. Succession disputes that followed the death of Kazembe Kaleka
    2. Ngoni invasion
    3. Loss of control of trade
    4. Colonisation
    A. 1, 3, and 4
    B. 1, 2, and 3
    C. 1, 2 and 4
    D. 2, 3 and 4

19. Why was Fr. Goncalo da Silveira killed in Mwenemutapa kingdom?
    A. He bewitched the king.
    B. He robbed the Mwenemutapa of his gold.
    C. He was accused of being a sorcerer.
    D. He converted the king to Christianity.

20. Which of the following were strategies used by Dr. David Livingstone to open up Central Africa for missionary work?
    1. Raiding slave traders
    2. Opening up mission stations
    3. Establishing the Universities Mission to Central Africa
    4. Writing books about Central Africa.
    A. 1 and 4
    B. 2 and 4
    C. 2 and 3
    D. 1 and 3

21. One of the problems which the missionaries of Blantyre Mission faced was
    A. Few missionaries.
    B. Malaria and the resulting loss of life.
    C. Sparse population in the area.
    D. Slave raids in the area.

22. Why did the B. S. A. Company extend its influence from Mashonaland to Matabeleland?
    A. The Shona were troublesome.
    B. There was need to extend the rail line to Mafeking
    C. The whites believed there was gold.
    D. The Shona needed their land.

23. What influenced the British to occupy Nyasaland?
    A. Mineral resources
    B. Cheap labour
    C. Desire to develop Nyasaland
    D. Desire to stop slave trade.

24. Why were the Bemba easily defeated by the British forces?
    A. Lack of unity among the Bemba chiefs
    B. Some Bemba chiefs had befriended the British
    C. The Bemba used traditional weapons
    D. The Bisa and Mambwe fought alongside the British.

25. Which commission demanded the release of Dr. H. Kamuzu Banda?
    A. Bledsloe Commission
    B. Monckton Commission
    C. Devlin Commission
    D. Hilton Young Commission

26. Why did the colonial government in Nyasaland buy land and incorporate it into the Crown land?
    A. To distribute it to poor Africans.
    B. To give it to the European commercial farmers.
    C. To incorporate it into the Native Purchase Areas.
    D. To establish reserves for future allocations.

27. Why was the British government initially not interested in occupying Nyasaland?
    A. They feared conflicts with slave traders.
    B. The area was not accessible.
    C. The area was not economically viable.
    D. They feared its strong ethnic groups.

28. Which statement is TRUE about the Lawley Treaty?
    A. It was signed between Arthur Lawley and Lobengula.
    B. It was signed in 1890.
    C. It opened up trade opportunities.
    D. It increased the annual subsidy to Lewanika to £2000.

29. Why did recruiting agencies prefer Nyasaland as a source of labour?
    A. Nyasaland offered cheaper labour than other countries.
    B. People of Nyasaland were skilled mine workers.
    C. Nyasaland was close to the mining centers.
    D. Nyasaland workers could be exported without a pass.

30. In what way did labour migration result in the stagnation of rural communities?
    A. Returning migrants preferred to stay in urban areas.
    B. Absence of most men resulted into increased insecurity.
    C. Agricultural production at family level declined.
    D. Migrants did not send financial support to their families.

31. Which countries were signatories of the Dual Entente?
    A. Germany and Austria-Hungary
    B. France and Russia
    C. Britain and Russia
    D. France and Italy

32. Which European countries formed the Triple Entente in 1907?
    A. Austria-Hungary, France and Russia
    B. Britain, USA and Germany
    C. Italy, France and Austria-Hungary
    D. Britain, France and Russia

33. What was the immediate cause of World Economic crisis?
    A. low industrial production
    B. huge debts to pay back to USA
    C. great speculation
    D. heavy dependence on imports

34. Which of the following contributed to March 1917 Russian Revolution?
    A. Deportation of Lenin
    B. Influence from Fr. Gapon
    C. People's dissatisfaction with Rasputin
    D. Peasants' dissatisfaction with redemption payments

35. Why was the North Atlantic Treaty Organization (NATO) formed in 1949?
    A. To recover Europe's economy
    B. To react to Warsaw Pact
    C. To check USSR aggression
    D. To check Cuban attacks

36. What made India and Pakistan exist as independent states from August 1947?
    A. Japan liberated them from Britain
    B. Muslim-Hindu conflicts
    C. Death of Mahatma Gandhi
    D. Disagreement between Nehru and Jinnah

37. How was nationalism fostered in Kenya against the British colonization?
    1. Patriotism by Sir Evaly Baring
    2. Pressure to from KANU and KADU parties
    3. Resistance by Ronald Ngala
    4. Mau Mau terrorist attacks
    A. 1 and 2
    B. 2 and 4
    C. 1 and 3
    D. 3 and 4

38. Which statement is TRUE about the Angolan Civil War?
    A. Russia and Cuba co-worked with Savimbi
    B. Americans and Mobutu assisted UNITA
    C. Agosto Neto led MPLA
    D. MPLA and UNITA fought over diamond

39. Why was the Common Market for Eastern and Southern Africa (COMESA) formed in 1994?
    A. To create economic prosperity.
    B. To end economic conflicts in the region.
    C. To plan for single currency in the region.
    D. To avoid dependence on South Africa

40. Why has Zambia been facing severe economic hardships since the 1980s?
    A. Fluctuating prices at the world market.
    B. Inflation and underdevelopment
    C. Deaths of potential leaders
    D. Economic sanctions from rich countries.

41. The objective of the Triple Alliance in 1882 was to
    A. wage war on Britain.
    B. defend its members against the Triple Entente.
    C. enhance friendship amongst its members.
    D. protect Germany and Austria-Hungary from Russia and France.

42. What caused the Second Balkan War?
    A. The division of Macedonia
    B. The attack of Bulgaria on Greece
    C. The signing of the Treaty of Bucharest
    D. Rumania was dissatisfied with the Balkan League

43. What reasons led to the Versailles Conference in 1919?
    1. The horror of the First World War.
    2. The feeling to punish Germany.
    3. The need to create the world body.
    4. To solve the world economic crisis.
        A. 1, 2, and 4
        B. 1, 2, and 3
        C. 1, 3, and 4
        D. 2, 3, and 4

44. Why did the League of Nations fail to preserve world peace?
    1. The US Senate rejected it
    2. Socio-economic problems in Europe
    3. Lack of a standing army
    4. Any member had veto power
        A. 1 and 2
        B. 3 and 4
        C. 1 and 3
        D. 2 and 4

45. What is the significance of the "Black Thursday" in American History?
    A. Millions of jobs were lost.
    B. Americans sold millions of shares.
    C. American Black demonstrators were killed.
    D. American farmers presented a petition to the White House.

46. What was the main agreement of the Nazi-Soviet Pact of 1939?
    A. Non-alignment to Britain and France
    B. Economic assistance in time of trouble
    C. Neutrality if one of is at war
    D. To defy the principles of the League of Nations

47. Why did the Western countries consider the spread of communism a threat to world peace?
    1. It promoted dictatorship governments.
    2. It encouraged the manufacture of military hardware.
    3. It discouraged private ownership of industries.
    4. It encouraged the use of chemical and biological weapons.
    A. 2 and 4
    B. 1 and 3
    C. 2, 3, and 4
    D. 1, 2, and 3

48. Why was the Reinsurance Treaty signed in 1887?
    A. Germany and Russia agreed to help each other in times of any attack.
    B. Germany wanted to prevent Russia from signing a treaty with France.
    C. Russia and Germany planned to attack France.
    D. Russia wanted Germany to help her to gain territories in the Balkans.

49. Why did Gandhi walk to the Sea Coast to collect salt?
    A. To help local inhabitants.
    B. To conduct mass rallies along the way.
    C. To see the Sea Coast.
    D. To protest against British discriminating laws.

50. Why did Britain abandon the policy of "Splendid Isolation"?
    A. Britain was defeated by Germany in 1900.
    B. Britain wanted to create a large army.
    C. Britain had so many colonies in Africa and Asia.
    D. Britain wanted to protect her empire.

51. Which territories were annexed by Germany after the Franco-Prussian War?
    A. Alsace and Lorraine
    B. Ruhr and Saar region
    C. Rhineland and Alsace
    D. Danzig and Ruhr

52. Which of the following are some of the permanent members of the Security Council of the UNO?
    A. Britain, Japan, China, and Russia.
    B. China, Italy, USA and France
    C. Britain, USA, Russia, and Japan
    D. China, France, Russia, and USA

53. Which organ of the UNO replaced the League of Nations Mandates Commission?
    A. The Trusteeship Council
    B. The General Assembly
    C. The Economic and Social Council
    D. The International Labor Organization

54. Which religions played a prominent role in the politics of India?
    A. Buddhism and Islam
    B. Islam and Hinduism
    C. Buddhism and Buddhism
    D. Buddhism and Christianity

55. What happened when Germany failed to pay the reparations in 1921-22?
    A. France occupied the Ruhr coal fields
    B. Germany was allowed not to pay for five years.
    C. Germany got a loan from USA to pay for five years.
    D. There were revolts in France because Germany stopped paying.

56. Why did businessmen support Adolf Hitler in Germany?
    A. The businessmen feared communism.
    B. Hitler promised them business opportunities.
    C. Hitler hated the Jews who hated businessmen.
    D. Hitler promised to create many job opportunities.

57. Which part of Germany was to be administered for 15 years after the Versailles Treaty?
    A. The Rhineland
    B. The Ruhr
    C. The Saar
    D. The Upper Silesia

58. Which countries benefited from the reparations?
    A. Belgium, France, and Britain
    B. Britain, USA, and Belgium
    C. Belgium, Russia, and USA
    D. USA, Britain, and France

59. Which of the following international disputes was successfully resolved by the League of Nations?
    A. The Japanese invasion of Manchuria in 1931
    B. The Polish invasion Lithuania conflict over Vilnn in 1920
    C. The German/Polish conflict over of Upper Silesia in 1920
    D. The Italian invasion of Abyssinia in 1935

60. Which of the following were causes of the Great Depression of 1929?
    1. Domestic overproduction
    2. American people were lazy
    3. Drop in the volume of exports
    4. Speculations
    A. 1 and 4 only
    B. 2, 3, and 4
    C. 1, 3, and 4
    D. 2, and 3 only

APPENDIX C

SET 3 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (CLASS TEST)

1. Which part of Nyasaland had the most land problems?
   a. Northern province of Nyasaland
   b. Shire highlands
   c. Southern province of Nyasaland
   d. Central Province of Nyasaland

2. Who controlled most of the land in the Northern Province of Nyasaland?
   a. B. S. A. Company
   b. African Lakes Company
   c. Livingstonia Mission
   d. Blantyre Mission

3. Why did the colonial government buy land and incorporate it into the crown land?
   a. to be distributed to the Africans
   b. for the Europeans to start commercial farms
   c. to be incorporated into the Native Purchase areas
   d. to be allocated for future use

4. What was the use of Nature Trust Lands
   a. was to be held exclusively for African tribal use.
   b. Was exclusively distributed to the immigrants from Mozambique.
   c. Was to be alienated for Europeans to open farms.
   d. Was reserved for future use

5. Name the cash crop that was first introduced in Nyasaland.
   a. Tobacco
   b. Tang oil
   c. Coffee
   d. Tea

6. What was the cash crop that was grown by Europeans in Central Africa?
   a. Tobacco
   b. Tang oil
   c. Coffee
   d. Tea

7. What was the major problem to African farming in Central Africa?
    a. congestion in the reserves
    b. keeping too much animals
    c. soil erosion
    d. overgrazing

8. Why did Elliot Kamwana leave Livingstonia Mission in 1901?
    a. He was dissatisfied with the teachers and their teachings.
    b. Ordinations of African priests was taking too long to be conducted.
    c. He was against the introduction of school fees.
    d. To form Watch Tower movement

9. Why did most European farmers in Malawi settle in the Shire Highlands?
    a. It had the best soils for agriculture.
    b. It had favorable climate for settlement.
    c. It had good communication system.
    d. Most of the land was not occupied.

10. Why did the Bemba refuse to accept British authority in their area?
    a. They had already been approached by the Portuguese.
    b. They considered the British to be harsh.
    c. The British came with anti-slave campaigns.
    d. They were afraid of losing their powers to the whites.

11. In what did labor migration results in the stagnation of rural communities?
    a. Returning migrants preferred to stay in urban area.
    b. Absence of most men resulted into increased insecurity.
    c. Production of agricultural of family level declines.
    d. Migrants did not send financial support to their families.

12. The advice to bring together all the Native Associations into one body came from
    a. Dr. Hastings Kamuzu Banda
    b. Symon Chiukepo Mhango
    c. Levi Zililo Mumba
    d. Chief Amon Mtwalo

13. Name two countries where Malawians went to work.
    a. Mozambique and Zambia
    b. South Africa and Botswana
    c. Zambia and Zimbabwe
    d. Zambia and Zaire

14. Which Ngoni group settled at Fort Jameson?
    a. Mpherembe
    b. Mlonyeni
    c. Mpezeni
    d. Chiwere Ndhlovu

15. Why was the British government initially not interested to occupy Nyasaland?
    a. There were too many slave trading activities in the area
    b. The area was not accessible
    c. They were afraid of resistance from strong tribes
    d. The area was not economically viable.

16. What problems did independent churches face in Nyasaland?
    1. Lack of followers
    2. Frequent arrests of its leaders
    3. It leaders were not well educated
    4. Limited funds
    a. 1 and 2
    b. 1 and 3
    c. 2 and 4
    d. 3 and 4

17. Why did recruiting agencies prefer Malawi as a source of labor?
    a. Labor from the country were cheaper
    b. Malawians were more skilled in mine work.
    c. Malawi was nearer to where labor was needed.
    d. Malawians could be exported without a pass.

18. During the colonial period Nyasaland was unable to achieve much economic development mainly because
    a. her neighbors were industrially advanced
    b. she lacked mines and factories
    c. she did not have sufficient farming land
    d. the colonial government did not plan for the governed.

19. Charles Chinula believed that African customs had to be
    a. Modernized.
    b. Left as they were.
    c. Studied and understood.
    d. Purified and preserved.

20. What was the Independent church that was formed by Hannock Msokera Phiri?
    a. Ana a Mulungu
    b. Watch Tower
    c. African Methodist Evangelical church
    d. Sazu Mission

21. People associated with the early Iron Age sites of Central Africa were
    a. Bantu
    b. Akafula
    c. Negroes
    d. Bushmen

22. What evidence is there to show that there was interaction between the Iron Age people and the Stone Age people
    a. Iron Age people depended on meat from Stone Age people.
    b. Stone Age people lived on cereals produced by Iron Age people.
    c. Each group had to exchange what they had for what they lacked.
    d. Pottery was found in Stone Age sites.

23. Who were the Amaholi?
    a. People who could not speak Ngoni at Linyati.
    b. The original Ngoni settlers in Mashonland.
    c. The lowest class of people in the Ndebele Society.
    d. They were the young Ngoni dancers in Matabeleland.

24. What was an Impi?
    a. A group of Ngoni warriers from Zululand.
    b. A segment of Chewa fighters from Nyasaland.
    c. A group of white settlers in Mashonaland.
    d. A leader of the garrison in Matabeleland.

25. What was the main port for exporting slaves in East Africa?
    a. Sofala
    b. Kilwa
    c. Zanzibar
    d. Beira

26. Which Ngoni group settled at Fort Jameson till the colonial administration of Northern Rhodesia?
    a. Mpezeni
    b. Mpherembe
    c. Mlonyeni
    d. Chiwere Ndlovu

27. How did the Kololo come to the Shire Valley in 1861?
    a. They were looking for fertile land.
    b. They came as traders.
    c. They were escaping Lozi attacks.
    d. They came as porters of Dr. David Livingstone.

28. The earliest Protestant Mission in Nyasaland was the
    a. Universities Mission to Central Africa
    b. London Missionary Society
    c. Roman Catholic Mission
    d. Free Church of Scotland

29. Some Yao chiefs who helped to spread Islam in Malawi were
    a. Jumbe and Makanjira
    b Mponda and Jumbe
    c. Mlozi and Jumbe
    d. Mponda and Makanjira

30. Which two European nations were interested to colonize Malawi?
    a. Britain and France
    b. Britain and Germany
    c. France and Portugal
    d. Britain and Portugal

31. What was the capital of Northern Rhodesia after 1911?
    a. Kalomo
    b. Livingstone
    c. Lealui
    d. Lusaka

32. Which concession seeker representing the rival companies escorted Babyana and Mshete to London?
    a. Maund
    b. Lotje
    c. Umgadan
    d. Dawson

33. The Mlimo Cult priest and the two mhondoros who backed up the Shona-Ndebele Rebellion of 1896/97 were
    a. Mwari, Umgadan and Nehanda
    b. Lomagundi, Mkwati and Kakubi
    c. Gomalla, Mwari and Chibi
    d. Mkwati, Kakubi and Nehanda

34. Which Boer leader attacked the Ndebele at Mosega in 1837?
    a. Hendrix Potgieter and Louis Tricharidt
    b. Hartly Mauch and Paul Kruger
    c. Silva Porto and Maund
    d. Miguel Bernandes and Caiada

35. Who united the Livingstonia Free Church of Scotland, Blantyre Mission, and the Dutch Reformed Church to form the Central African Presbyterian in 1904?
    a. William Murray
    b. Dr. Robert Laws
    c. Dr. Robert Moffat
    d. Robert Blake

36. When was Domasi C.C.A.P. mission in Zomba established as a branch of Blantyre Mission?
    a. 1878
    b. 1884
    c. 1889
    d. 1892

37. Which of the following countries formed the entente cordiale in 1904?
    a. Germany and Russia
    b. Britain and France
    c. Russia and France
    d. France and Belgium

38. The Central Powers of Europe during the first world war included
    a. Germany, Britain, France and Belgium
    b. Russia, Italy, Turkey and Bulgaria
    c. Austria-Hungary, Germany, Turkey and Bulgaria
    d. Germany, Romania, Bulgaria and Japan

39. Why was the Dual Alliance made in 1892?
    a. As a defence against Russia and France by Germany and Austria.
    b. Russia and France invited to defeat Germany and her allies.
    c. As a defence against entente allies by Britain and USA.
    d. To threaten Britain to stop her policy of Splendid Isolation.

40. How did the political alliances contribute to the outbreak of the First World War?
    a. They ended up in dividing Europe into two rival camps.
    b. Bosnia-Hezegovina was annexed by Austria-Hungary.
    c. By oppressing all European countries that were weak.
    d. By sparkling the immediate outbreak of the First World War.

41. The naval race before the outbreak of the First World War was between
    a. Austria-Hungary and Russia
    b. USA and Germany
    c. Germany and France
    d. Britain and Germany

42. Why was German Ruhr district occupied by France and Bulgaria in 1923?
    a. To stop her from fighting against Russia.
    b. To force German pay the reparation.
    c. To get rid of all terrorists in Germany.

43. Why did Germany declare war on Russia in 1914?
    a. Russia had refused to demobilise.
    b. Russia attacked Austria and Prussia.
    c. Russia declared support for Serbia.
    d. Russia was an ally of Britain.

44. Which Germany leader scheduled a plan to attack France through Belgium and defeat her within six weeks?
    a. von Papen
    b. Alfred Schlieffen
    c. Otto von Bismarck
    d. Kaiser William II

45. What forced America to join the First World War under the allies?
    a. The bombing of Pearl Harbour by Japan.
    b. Germany bombed American nationals.
    c. The sinking of American ship by Germany.
    d. She was requested by the allies to assist them.

46. The Balkan League was formed by the following members
    a. Bulgaria, Greece, Montenegro, and Turkey
    b. Bulgaria, Serbia, Greece and Montenegro
    c. Albania, Bulgaria, Montenegro and Greece
    d. Serbia, Albania, Greece and Turkey

47. Why was Archiduke Franz Ferdinand of Austria killed by the Serbs?
    a. He was ill-treating the Serbs.
    b. The Serbs were too racists
    c. The Serbs wanted their independence.
    d. The Archduke befriended their enemies.

48. The four political leaders during the Versailles Peace Treaty were
    a. Lloyd George, Orlando, Clemenceau and Woodrow Wilson
    b. Woodrow Wilson, Jan Smuts, Leon Bourgeois and Orlando
    c. Leon Bourgeois, Clemenceau, Woodrow Wilson and Lloyd George
    d. Robert Cecil, Orlando, Widrow Wilson and Jan Smuts.

49. How did the Moroccan Crises contribute to the outbreak of the First World War?
    a. They encouraged Germany to make more war weapons.
    b. They made European countries to make more political alliances.
    c. They made the enmity between Germany and Russia to intensify.
    d. They strengthened bitterness between two alliance groups.

50. What caused the Second Balkan War to break up in 1913?
    a. The idea of nationalism
    b. The act to contain the Young Turks.
    c. A move to expand the Turkish Empire.
    d. A problem in sharing the spoils of war.

51. Why was the Turkish Empire nick-named the "the sick man of Europe"?
    a. It was conquered by Russia in 1914?
    b. It isolated itself from all conflicts.
    c. There were many HIV/AIDS victims in the empire.
    d. It was economically and militarily weak due to conflicts.

52. a. What was meant by
        i) Voluntary Associations
        ii) African Independent churches
    b. Why were voluntary associations formed in Nyasaland. Give two reasons.
    c. Why were such associations important in the development of African politics in Malawi.
    d. Mention three independent churches in the Northern Malawi and their founders.
    e. State any two problems, which the African Independent churches faced.

53. a. Define Labor migration.
    b. Give three factors that led to labor migration.
    c. State three ways how colonial government tried to exercise control over migration system.
    d. State three ways how each of the following were affected by the labour migration system.
        i) Nyasaland
        ii) South Africa, Zambia and Zimbabwe.

54. a. What is Iron Age culture?
    b. Identify any two Iron Age sites in Zimbabwe.
    c. Explain how important is Iron Age culture to civilization.

55. a. Which three factors led to the rise and expansion of the Maravi kingdom?
    b. Who were the founders of the Kazembe kingdom?
    c. Name any two economic activities of the Kazembe kingdom.

56. a. Name three importances of religion in the Maravi Kingdom.
b. What contributed to the organizational structure of the Nkhamanga Kingdom.

57. a. What event led to the migration of the Ndebele into Central Africa?
b. State three reasons why slave trade expanded in the East African coast after 1750.
c. Name the three groups of people who helped to suppress the slave trade in Central Africa.

58. During 1893 Jameson gave hope of incorporating Matabeleland peacefully under company rule.... He therefore made preparations in Salisbury for a force of volunteers to invade Matebeleland ..... Rhodes gave his famous answer. Read Luke, Chapter 14, vs. 31 "If any king goes out with ten thousand men to fight another king who comes against him with twenty thousand men, he will sit down first and decide if he is strong enough to face that king"._ PEN TINDALL _ History of Central Africa
    a. Relate the Bible quotation to the situation at that time.
    b. Describe events, which indicated that the Ndebele were preparing for war.
    c. What was the Pioneer Column.

59. a. When did Northern Rhodesia become a British Protectorate?
b. State five effects of mining and labor migration.

60. a. Briefly describe the role played by Tippu-Tipu in the East Africa slave trade.
b. Explain three negative and two positive effects of the slave trade.
c. What made Dr. David Livingstone start early journeys from Kuluman?

APPENDIX D

SET 4 OF MSCE HISTORY MULTIPLE-CHOICE ITEMS (MSCE 1A)

1. What major factor contributed to the establishment of the Tumbuka-Nkhamanga Kingdom?
   a. trade in ivory and slaves
   b. marriage into the Tumbuka royal clans
   c. peaceful co-existence with the local people
   d. wars of conquest by the immigrants

2. In what way did the loss of trade weaken the position of the Kazembe Kingdom in the later part of the 18th Century?
   a. loss of control of the army which strengthened the position of the kingdom.
   b. loss of control of subject peoples who no longer paid tribute
   c. loss of religious powers that had been uniting the people in the kingdom
   d. loss of control of economy which had backed up the king's position

3. Why was the Nkhamanga kingdom unique in the way it was founded?
   a. Its leaders did not establish a permanent dynasty.
   b. Local leaders were involved in running the affairs of the kingdom.
   c. Its leaders did not lead a life of seclusion.
   d. It was founded through peaceful means.

4. What was the major factor for the decline of the Kazembe Kingdom?
   a. slave trading activities
   b. loss of monopoly over long distance trade
   c. succession disputes which led to civil wars
   d. weak kings who were only interested in tributes

5. One effect of slave trade which could be seen here in Malawi is
   a. graves of people who died in raids.
   b. depopulated areas.
   c. establishment of towns.
   d. remains of guns and gun power.

6. Which of the following was the effect of farming and domestication of animals during the Iron Age?
   a. population increase
   b. civil wars
   c. political states
   d. abandonment of hunting

7. What major role did religion play in the Maravi Kingdom?
   a. It gave the Kalong religious powers.
   b. It strengthened the position of the rain cult.
   c. It made people loyal to their ancestors.
   d. It united the people in the Kingdom.

8. Gold beads, glass beads, copper and iron bangles, Mpande shell and cloths were excavated at Ingombe Ilede Iron Age site. This suggest that Ingombe Ilede
   1. was very important trading center.
   2. was the main port used by Africans and Arabs in the East African Coast.
   3. practised foreign trade.
   4. was on the route from the East to the West Coast of Africa.

   a. 1, 2, and 3
   b. 2, 3, and 4
   c. 1, 3, and 4
   d. 1, 2, and 4

9. The immediate importance of the Iron Age Culture to Central Africa was the
   a. formation of pre-colonial kingdoms.
   b. beginning of voluntary associations.
   c. formation of nationalist movement.
   d. beginning of resistance movements.

10. The Kazembe Kingdom prospered more easily because
    a. of rich soils and abundant natural resources.
    b. they were successful in raiding other tribes.
    c. the kingdom was united as a result of centralization.
    d. of Kazembe Lukwesa's political success.

11. Portuguese influence in the interior of Africa declined during the later part of the 17th Century because
    a. gold trade in the interior declined.
    b. they were expelled from the interior by the Zimba.
    c. the Arabs who established plantations on the island of Zanzibar took over control of the gold trade.
    d. their position as a leading trading nation was challenged by other European nations.

12. In What ways did the Yao culture facilitate the quick spread of Islam in their areas?
    a. They came from same area as the Arabs.
    b. They had been trade partners with Muslims for a long time.
    c. Their language was similar to that of the Arabs.
    d. They had similar religious beliefs with Arabs.

13. The major benefit brought by Jumbe to the Chewa of Nkhotakota was
    a. the introduction of new crops such as rice.

b. protection against Yao and Ngoni raids.
c. numeracy and literacy.
d. the Arabic language.

14. Which of the following strategies used by Dr. David Livingstone was a success in opening up Central Africa to missionary work?
a. his writings and public speeches
b. his public speeches and expeditions
c. his expeditions and writings
d. his writings and opening up of mission stations

15. What major effect of Islamic culture can still be witnessed today in Malawi?
a. Arabic language
b. formal education
c. circumcision
d. mode of dress

16. Which of the following best explains why the British delayed in occupying North-Eastern Rhodesia?
a. The area had a lot of hostile chiefs.
b. The BSA Company wanted to consolidate its position in Southern Rhodesia first.
c. North-Eastern Rhodesia was not economically viable.
d. The BSA Company was waiting for permission from the British government to occupy the area.

17. The Bemba failed to resist British occupation of their territory after 1899 because
a. of internal conflicts.
b. their weapons were inferior to the British.
c. some of their chiefs supported the British.
d. British forces were helped by the Bisa and the Mambwe.

18. The Roman Catholic missionaries withdrew from Mponda's area because
a. the area was already under British protection.
b. their teachings attacked the people's traditional beliefs and customs.
c. most of the missionaries died of repeated attacks of malaria.
d. The people of Mponda's area were already converted to Islam.

19. What major factor led to the end of the East African slave trade?
a. work of the Scotish missionaries
b. intervention of the British Government
c. legal trade started by the African Lakes Company
d. Africans' realization of the evils of slave trade

20. The successes of the Ndebele and Ngoni invasions were mainly due to
    a. organizational abilities of their kings.
    b. use of firearms they seized from weak tribes.
    c. the assistance they got from the incorporated Amaholi.
    d. their traditional war tactics and bravery.

21. What made the Ndebele finally move away from Mosega in 1837?
    a. death of their leader Mzilikazi
    b. their defeat by Boers
    c. need to find a better place to cultivate crops
    d. attacks by Dingani

22. The Balowoka were accepted as leaders by the Tumbuka speaking people in Northern Malawi in the 18th Century because they were
    a. honest businessmen.
    b. at a time of power vacuum.
    c. very good fighters.
    d. very good elephant hunters.

23. Why did the British Government administer Southern Rhodesia through the British South Africa Company?
    a. It wanted to avoid extra administrative expenses.
    b. The Matebele preferred to be ruled by the British South Africa Company.
    c. The British South Africa Company made a lot of profit for the British Government.
    d. The British South Africa Company officials were already within the area.

24. In what way did labour migration result in the stagnation of rural communities?
    a. Returning migrants preferred to stay in urban areas.
    b. The absence of most men resulted into increased insecurity.
    c. The production of agriculture goods at family level declined.
    d. Migrants did not send financial support to their families.

25. What was the 1949 Victoria Falls Conference important in the history of Central Africa?
    a. It was the first time Africans were allowed to attend.
    b. It decided to form a Federation of Rhodesia and Nyasaland.
    c. It introduced a new concept of partnership among races.
    d. It initiated economic co-operation among the countries in Central Africa.

26. The Ngoni resisted the Protectorate government because
    a. they hated of paying tax.
    b. they hated the government because it encouraged the Livingstonia Mission to condemn polygamy.
    c. they thought their leaders would lose power to the whites.
    d. they wanted to protect their fertile land.

27. What actions of the Ngoni of Mmbelwa influenced the Tonga of Nkhata Bay to welcome the Livingstonia missionaries?
    a. their policy of assimilation
    b. persistent raids on the Tonga
    c. forcing the Tonga to join their army.
    d. interference in Tonga politics.

28. Why did the Chiwere Ndlobvu group settle in Dowa?
    a. The area had good pasture for cattle.
    b. They were interested in the activities of the missionaries.
    c. The area had fertile soils for growing crops.
    d. The area was free of slave trade.

29. What made Dr. David Livingstone start early journeys of exploration from Kuruman?
    a. signs of slave trade which he saw for the first time
    b. encouragement which he received from Robert Moffat
    c. the trekking of the Boers northwards
    d. his desire to stop Muslim influence in the interior

30. Which of the following was Dr. David Livingstone's achievement in Central Africa?
    a. He established Livingstonia Mission.?
    b. He drew a map of a larger part of Central Africa.
    c. He established legitimate trade in the area.
    d. He introduced Ilala as a mode of transport on Lake Malawi.

31. Otto Von Bismarck started signing political alliances in an attempt to
    a. challenge British navy supremacy.
    b. prevent a war of revenge from France.
    c. isolate Russia.
    d. avoid a two front war.

32. The USA joined the First World War on the allied side when
    a. the Lusitania was sunk.
    b. Germany declared unlimited submarine warfare.
    c. Russia pulled out of the First World War.
    d. Woodrow Wilson was elected President pf USA.

33. Which of the following factors **best** explains why Britain and France became allies after being enemies for a long time?
    a. Germany had rejected British attempts of friendship.
    b. German aggression had threatened British supremacy.
    c. British felt isolated by alliances.
    d. The British monarch paid a state visit to France.

34. The immediate consequences of the Wall Street Crash in USA included
    1. the closure of the banks.
    2. the change of government.
    3. unemployment.
    4. reduction in industrial production.

    a. 3 and 4
    b. 1, 2, and 3
    c. 1 and 3
    d. 1, 3, and 4

35. Russia wanted to expand into Balkans because she wanted to
    a. avenge her defeat at the hands of Japan.
    b. gain access to the Mediterranean sea.
    c. prevent the Bolsheviks from infiltrating into Russia.
    d. show solidarity with members of the Triple Entente.

36. Why did French troops invade the Ruhr industrial region of Germany in 1923?
    a. Germany stopped paying reparations.
    b. The Ruhr was originally a French region.
    c. French was revenging on her loss of Alsace and Lorraine.
    d. Most people of the Ruhr were French speaking.

37. During the First World War the countries, which made up of the Central Powers were
    1. Germany.
    2. Austria-Hungary.
    3. Italy.
    4. Russia.

    a. 3 and 4
    b. 2, 3, and 4
    c. 1 and 2
    d. 1, 2 and 3

38. For Stalin to rebuild Russia after the Second World War he had to
    a. relax economic control.
    b. give priority to heavy industry.
    c. conscript workers into industry.
    d. encourage large families.

39. Why did the Versailles Settlement create the Polish Corridor in 1919?
    a. To punish Germany for attacking Poland.
    b. To divide Germany into two.
    c. To give Poland access to the sea.
    d. To create a free city of Danzig.

40. Which one of the following was **not** one of the Woodrow Wilson's fourteen points?
    a. Alsace and Lorraine were to be given back to France.
    b. Russia had to be left free to form her type of government.
    c. No free navigation at sea for nations during war or peace.
    d. There were to be no trade barriers between nations.

41. What caused the civil war of 1918 to 1920 in Russia?
    a. nationalization of industries
    b. abolition of Kulaks system
    c. dissolution of the assembly
    d. requisition of peasant grain

42. What result did the program of the "Battle of Wheat" produce in Italy?
    a. Peasant wheat was requisitioned by the government.
    b. Italy produced enough wheat to feed its population.
    c. Wheat became the economic backbone of Italy.
    d. Italy imported a lot of Wheat.

43. Why did Lenin embark on the New Economic Policy?
    a. to allow peasants to own some land and sell food grown for profit
    b. to encourage private enterprise among businessmen
    c. to cool down the Kronstadt naval uprising against his government
    d. to encourage economic prosperity which war communism had failed to achieve.

44. Which event was used by Adolf Hitler to destroy the Communist party in Germany?
    a. the night of Long Knives
    b. the passing of Nuremberg Laws
    c. the Reichstag fire
    d. the passing of Enabling Law.

45. The establishment of the Comecon in 1949 was primarily to
    a. establish trade with the capitalist countries.
    b. coordinate propaganda programmes within Russia.
    c. encourage communist revolution abroad.
    d. integrate Soviet and East European economies.

46. One lasting impact of the Treaty of Versailles is
    a. the existing states of Eastern Europe.
    b. activities of the United Nations.
    c. reunion of West and East Germany.
    d. existence of democracy in the World.

47. "I can predict with absolute certainty that within another generation there will be another war." What prompted Woodrow Wilson to make this statement?
   a. The other allied powers did not appreciate the importance of the League of Nations.
   b. He was not satisfied with the punitive measures, which were outlined for Germany.
   c. His countrymen rejected USA's membership at the League of Nations.
   d. He noted that some big powers were not ready to forgive Germany.

48. The Dreadnaught was
   a. Germany's war strategy on two fronts.
   b. The steam-boat that Germany sent to Agadir.
   c. An alliance between Germany, Austria-Hungary and Russia.
   d. A super battleship that the British launched in 1906.

49. Local Germans supported the Weimar Republic because it
   1. supported terrorism.
   2. arrested the people without trial.
   3. had signed the Treaty of Versailles.
   4. failed to operate a democratic government.

   a. 1 and 2
   b. 2 and 4
   c. 1 and 3
   d. 3 and 4

50. The following statements on Gregory Rusputin are true except that he
   a. was murdered by members of the Russian ruling class.
   b. tried to get rid of the corruption in Russian government.
   c. acquired the reputation of a holy man.
   d. exercised great power over government policy during the First World War.

51. What did Adolf Hitler mean by the "stab in the back"?
   a. that Germany was unfairly treated by the Versailles Treaty.
   b. that France caused the First World War
   c. that the German army was betrayed.
   d. that Germany's problems were caused by the storm troopers.

52. The collectivization programme embarked upon in Russia in the 1920s meant that
   a. everyone was given free land.
   b. tractors were used freely.
   c. food production would be enhanced.
   d. land would be owned communally.

53. The Young Plan of 1929
   a. drew up a new plan for reparation payments.
   b. guaranteed the frontiers of Germany, France and Belgium.
   c. guaranteed polish neutrality.
   d. led to the Wall Street crash in the USA.

54. By the policy of "containment," the United States of America wanted to
   a. provide loans to European nations.
   b. attract European nations to join its defensive pact.
   c. prevent further spread of communism.
   d. destroy the Iron Curtain developed by Stalin.

55. Which policy helped Russia to achieve heavy industrialization?
   a. war-communism
   b. five-year plans
   c. collectivization
   d. new economic policy

56. Serbia was very bitter with Austria-Hungary soon after the First Balkan War because
   a. Serbia was attacked by Austria-Hungary
   b. Serbia was taken over by Austria-Hungary.
   c. Austria-Hungary supported Ottoman Empire.
   d. Serbia was prevented by Austria-Hungary from getting a coastline on the Adriatic sea.

57. Just before being granted independence, the sub-continent of India was partitioned by the Government because of
   a. the vastness of the subcontinent.
   b. the desire of the British Government to continue their policy of divide and rule.
   c. the serious conflicts between the Muslims and the Hindus.
   d. the need to make each race rule itself.

58. Which leader in Germany took advantage of harshness of the Treaty of Versailles to rise to power in 1933?
   a. Gustav Stresseman
   b. Franz von Papen
   c. Adolf Hitler
   d. Paul Von Hindenburg

59. The Cuban Crisis of 1962 would not have occurred, had
   a. the USA not helped General Batista to invade Cuba.
   b. Fidel Castro not declared Cuba a Communist State.
   c. Russia planted missiles on Cuba.
   d. Cuba not nationalized foreign industries.

60. At the Yalta Conference an agreement was reached that

1. Germany be divided into zones.
2. the German terms of surrender be taken into account.
3. Russia should declare war on Japan.
4. democratic governments be formed in liberated countries.

a. 1, 2, and 3
b. 1, 3, and 4
c. 1, 2, and 4
d. 2, 3, and 4

APPENDIX E

QUESTIONNAIRE ON TEACHERS' PERCEPTION ON EXAMINEE PREPARATION

Thank you for agreeing to complete this survey.
The survey is about what you considered important during the period you were teaching your students who are about to write this year's national MSCE examination. The main interest of the survey is on the topics that were given priority and emphasis as you anticipated the coming examination.
Your responses are strictly confidential and so you should feel free to fill out this questionnaire.

1. How did you view the following topics of the MSCE History syllabus as you instructed your students in preparation of this year's national examination?

Central Africa

- Iron Age

  not at all important __ somewhat important__ important __ very important __

- Pre-Colonial Kingdoms
  Maravi

  not at all important __ somewhat important__ important __ very important __

  Tumbuka-Nkhamanga

  not at all important __ somewhat important__ important __ very important __

  Kazembe

  not at all important __ somewhat important__ important __ very important __

- Trade in Ivory, Gold and Slaves

  not at all important __ somewhat important__ important __ very important __

- 19th Century Immigrants into Central Africa
  Yao

not at all important __ somewhat important__ important __ very important __

Ndebele

not at all important __ somewhat important__ important __ very important __

Ngoni

not at all important __ somewhat important__ important __ very important __

- The Missionary Factor

Spread of Islam

not at all important __ somewhat important__ important __ very important __

Spread of Christianity

not at all important __ somewhat important__ important __ very important __

Impact of Missionary work on the society

not at all important __ somewhat important__ important __ very important __

- European Occupation and administration of Central Africa
  Nyasaland (Malawi)

not at all important __ somewhat important__ important __ very important __

Southern Rhodesia (Zimbabwe)

not at all important __ somewhat important__ important __ very important __

Northern Rhodesia (Zambia)

not at all important __ somewhat important__ important __ very important __

African Reaction to Colonial rule

not at all important __ somewhat important__ important __ very important __

- Political and economic developments in central Africa up to independence

Agriculture, mining and labor migration

not at all important __ somewhat important__ important __ very important __

The central African Federation

not at all important __ somewhat important__ important __ very important __

Development of nationalism

not at all important __ somewhat important__ important __ very important __

Development of mass political parties

not at all important __ somewhat important__ important __ very important __

World History

- Causes and results of the first world-war

Political alliances

not at all important __ somewhat important__ important __ very important __

Arms race

not at all important __ somewhat important__ important __ very important __

Nationalism in the Balkans

not at all important __ somewhat important__ important __ very important __

International crises

not at all important __ somewhat important__ important __ very important __

Murder at Sarjevo

not at all important __ somewhat important__ important __ very important __

Treaty of Versailles

not at all important __ somewhat important__ important __ very important __

The League of Nations

not at all important __ somewhat important__ important __ very important __

- Inter-war period 1919-1939

Economic problems in Europe and the USA

not at all important __ somewhat important__ important __ very important __

The Communist Revolution in Russia

not at all important __ somewhat important__ important __ very important __

Development of dictatorships government in Germany

not at all important __ somewhat important__ important __ very important __

- Causes and Results of second world-war

Problems with the Versailles Treaty

not at all important __ somewhat important__ important __ very important __

Aggression of Germany

not at all important __ somewhat important__ important __ very important __

Appeasement Policy

not at all important __ somewhat important__ important __ very important __

Failure of the League of Nations

not at all important __ somewhat important__ important __ very important __

- Decolonization in Asia and Africa

India

not at all important __ somewhat important__ important __ very important __

Kenya

not at all important __ somewhat important__ important __ very important __

- Post-colonial crises and challenges in Africa

Popular expectations at independence

not at all important __ somewhat important__ important __ very important __

Civil war in Mozambique

not at all important __ somewhat important__ important __ very important __

Military governments in Nigeria

not at all important __ somewhat important__ important __ very important __

Impact on society

not at all important __ somewhat important__ important __ very important __

2. How would you rate your coverage of the following topics?

Central Africa

- Iron Age

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Pre-Colonial Kingdoms

Maravi

not at all covered __ moderately covered__ adequately covered __ most adequately covered__

Tumbuka-Nkhamanga

not at all covered __ moderately covered__ adequately covered __ most adequately covered__

Kazembe

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Trade in Ivory, Gold and Slaves

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- 19th Century Immigrants into Central Africa

Yao

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Ndebele

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Ngoni

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- The Missionary Factor

Spread of Islam

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Spread of Christianity

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Impact of Missionary work on the society

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- European Occupation and administration of Central Africa

Nyasaland (Malawi)

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Southern Rhodesia (Zimbabwe)

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Northern Rhodesia (Zambia)

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

African Reaction to Colonial rule

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Political and economic developments in central Africa up to independence

Agriculture, mining and labor migration

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

The central African Federation

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Development of nationalism

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Development of mass political parties

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

World History

- Causes and results of the first world-war

Political alliances

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Arms race

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Nationalism in the Balkans

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

International crises

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Murder at Sarjevo

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Treaty of Versailles

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

The League of Nations

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Inter-war period 1919-1939

Economic problems in Europe and the USA

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

The Communist Revolution in Russia

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Development of dictatorships government in Germany

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Causes and Results of second world-war

Problems with the Versailles Treaty

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Aggression of Germany

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Appeasement Policy

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Failure of the League of Nations

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Decolonization in Asia and Africa

India

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Kenya

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

- Post-colonial crises and challenges in Africa

Popular expectations at independence

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Civil war in Mozambique

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Military governments in Nigeria

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

Impact on society

not at all covered __ moderately covered__ adequately covered __ most adequately covered __

3. List the topics on which you assessed your students in the course of instruction?

4. List the topics on which you would like to concentrate your mock examination items? (use the back page if you need more space).

Thank you for completing this questionnaire!

APPENDIX F

QUESTIONNAIRE ON TEACHERS' IMPRESSIONS ABOUT THE IMPACT OF DUPLICATE TEST CONSTRUCTION EXPERIMENT

Thank you for accepting to complete this survey. The purpose of this survey is to get your opinion about your participation in the test construction experiment for the History Mock Examination. Your responses to this questionnaire will be held strictly confidential and you need not write your name on the questionnaire. You should also feel free not to fill out the questionnaire or any of its parts.

1. How did you like the following activities during the MANEB test construction training?

   - Basic Multiple-choice item construction techniques

   - Higher Order Multiple-Choice item Construction techniques

   not all liked __ somewhat liked ___ liked ___ very much liked ___

   - Test Plan/Test specification construction

   not all liked __ somewhat liked ___ liked ___ very much liked ___

   - Item Analysis techniques

   not all liked __ somewhat liked ___ liked ___ very much liked ___

   - Group formation

   not all liked __ somewhat liked ___ liked ___ very much liked ___

2. How enlightening were the following activities to you?

   - Selection of topics to include in the test

   Very enlightening___ Enlightening___Somewhat enlightening___ Not at all

   - Construction of high order abilities items

   Very enlightening___ Enlightening___Somewhat enlightening___ Not at all enlightening___

- Selection of items to include in the test

Very enlightening___ Enlightening___Somewhat enlightening___ Not at all enlightening___

- Editing of items

Very enlightening___ Enlightening___Somewhat enlightening___ Not at all enlightening___

- Moderation of item

Very enlightening___ Enlightening___Somewhat enlightening___ Not at all enlightening___

- Item Analysis techniques

Very enlightening___ Enlightening___Somewhat enlightening___ Not at all enlightening___

3. Did the exercise help you to be more confident about your test construction practices? (Briefly describe specific activities that helped you most and how)

4. Did the exercise help to better understand the link between instruction and assessment? (if so briefly explain how).

5. What further training would you like to have to meet your professional needs both as an instructor and examiner?

6. What did you NOT like about the whole exercise? (Be specific about what you would have liked to do differently).

Thank you for completing this questionnaire!

APPENDIX G

STUDENTS' PERCEPTION ABOUT THEIR PREPARATION FOR MSCE HISTORY EXAMINATION

Thank you for accepting to complete this survey. The purpose of this survey is to get your opinion about how well you feel prepared for the 2003 MSCE History examination. You do not need to write your name on this questionnaire because your responses will be confidential and will be used for the purposes of this research only. You are free to complete or not to complete the questionnaire or any of its parts.

1. Listed below are topics in the MSCE History syllabus and are often taught in secondary school history classes. Review each topic and indicate how well you think your teacher covered the topic. Indicate your opinions using the rating scale provided by ticking (ˇ) on the space next to the rating you select.
   - Iron Age

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Pre-colonial Kingdoms (Maravi, Tumbuka-Nkhamanga, Kazembe)

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Trade In Gold, Ivory & Slaves

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Yao, Ndebele, Ngoni (immigrants)

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Muslim and Christian Influence

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Colonial Occupation(Malawi, Zambia, Zimbabwe)

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Political and economic developments In C. A.

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

   - Causes and Results of 1st World War

   Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

- Inter-war period

Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

- Causes and Results of 2nd World War

Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

- Decolonization in Asia and Africa

Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

- Post-colonial crises & challenges in Africa

Did not cover at all__ Covered poorly___ Covered well___ Covered very well___

2. On which topics were you most frequently assessed during the course of instruction? (consider tests and assignments you did during instruction).

3. On which topic(s) do you feel more adequately prepared for the 2003 MSCE examination? List as many as you feel confident about.

4. On which topic(s) do you feel the least prepared for the MSCE examination? List as many appropriate.

5. How many tests have you taken in your course while preparing for the MSCE examinations? (Consider all tests since you entered form 3).

   _____ tests.

6. What type of items (questions) was most frequently asked in the tests you took during the course of instruction? (choose one of the test types below)
   - Short answer _____
   - Multiple-choice _____
   - Essay _____

7. What grade do you expect to obtain in your MSCE History examination?

   Between 1 and 2 ___ Between 3 and 6___ Between 7 and 8___ Grade 9 ___

Thank you for completing this questionnaire!

APPENDIX H

QUESTIONNAIRES ON STUDENTS' OPINIONS ABOUT TEST QUALITY

Thank you for accepting to complete this survey. The purpose of this survey is to get your opinion about the quality of the tests you have recently written (Mock Form A, Mock Form B and MSCE History Paper 1A). Your opinion will be held in strict confidence and will be used for the purposes of research only. You are therefore, free not to respond to this questionnaire or any of its parts. Do not write your name or your identification number on this questionnaire.

1. Please rate the overall appearance of each test in terms of how well it represented the topics of the history syllabus.

	Poor	marginal	Satisfactory	Excellent
Mock A	___	___	___	___
Mock A	___	___	___	___
MSCE Paper 1A	___	___	___	___

2. How well did each test represent the topics that were your teacher taught?

	Poorly	Marginally	Adequately	More Adequately
Mock A	___	___	___	___
Mock A	___	___	___	___
MSCE Paper 1A	___	___	___	___

3. Please rate the overall clarity of each test in terms of level of language vocabulary and sentence structure.

	Poor	marginal	Satisfactory	Excellent
Mock A	___	___	___	___
Mock A	___	___	___	___
MSCE Paper 1A	___	___	___	___

4. How do you rate overall difficulty of each of the test?

	Easy	Somehow difficult	Difficult	Very Difficult
Mock A	___	___	____	____
Mock B	___	___	____	____
MSCE Paper IA	___	___	____	____

5. Do you think your participation in the duplicative test construction helped you to prepare for the national examinations? Explain.

6. Did you find any of the questions repeated in the three tests? Yes___ No ____
   If yes, List the topics or parts of the topics that were repeated.

Thank you for completing this questionnaire

APPENDIX I

CONTENT COVERAGE RATING SCALE FOR MSCE HISTORY TEST ITEMS

Directions: Please indicate the relevance of each item to each of the content areas listed below using a six-point scale where a rating of "1" indicates "not at all relevant" and a rating of "6" indicates "highly relevant." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have twelve ratings for each item.

CONTENT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Iron Age															
Pre-colonial Kingdoms															
Trade in gold, ivory & slaves															
19th Century immigrants															
The missionary factor															
European occupation in Central Africa															
Political & economic development in CA up to independence															
Causes & results of WW 1															
Inter War period (1919-1939)															
Causes & results WW 2															
Decolonisation in Asia & Africa															
Post-Colonial Crises and Challenges															
Other/Specify															

1 2 3 4 5 6
not at all relevant highly relevant

APPENDIX I, "Continued next page"

CONTENT COVERAGE RATING SCALE FOR MSCE HISTORY TEST ITEMS

Directions: Please indicate the relevance of each item to each of the content areas listed below using a six-point scale where a rating of "1" indicates "not at all relevant" and a rating of "6" indicates "highly relevant." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have twelve ratings for each item.

	TEST ITEM														
CONTENT	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Iron Age															
Pre-colonial Kingdoms															
Trade in gold, ivory & slaves															
19th Century immigrants															
The missionary factor															
European occupation in Central Africa															
Political & economic development in CA up to independence															
Causes & results of WW 1															
Inter War period (1919-1939)															
Causes & results WW 2															
Decolonisation in Asia & Africa															
Post-Colonial Crises and Challenges															
Other/Specify															

1 2 3 4 5 6

not at all relevant highly relevant

APPENDIX I "Continued, next page"

CONTENT COVERAGE RATING SCALE FOR MSCE HISTORY TEST ITEMS

Directions: Please indicate the relevance of each item to each of the content areas listed below using a six-point scale where a rating of "1" indicates "not at all relevant" and a rating of "6" indicates "highly relevant." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have twelve ratings for each item.

		TEST ITEM													
CONTENT	31	32	33	34	35	36	37	38	39	40	41	42	33	44	45
Iron Age															
Pre-colonial Kingdoms															
Trade in gold, ivory & slaves															
19th Century immigrants															
The missionary factor															
European occupation in Central Africa															
Political & economic development in CA up to independence															
Causes & results of WW 1															
Inter War period (1919-1939)															
Causes & results WW 2															
Decolonisation in Asia & Africa															
Post-Colonial Crises and Challenges															
Other/Specify															

1 2 3 4 5 6
not at all relevant highly relevant

CONTENT COVERAGE RATING SCALE FOR MSCE HISTORY TEST ITEMS

Directions: Please indicate the relevance of each item to each of the content areas listed below using a six-point scale where a rating of "1" indicates "not at all relevant" and a rating of "6" indicates "highly relevant." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have twelve ratings for each item.

CONTENT	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
Iron Age															
Pre-colonial Kingdoms															
Trade in gold, ivory & slaves															
19th Century immigrants															
The missionary factor															
European occupation in Central Africa															
Political & economic development in CA up to independence															
Causes & results of WW 1															
Inter War period (1919-1939)															
Causes & results WW 2															
Decolonisation in Asia & Africa															
Post-Colonial Crises and Challenges															
Other/Specify															

1 2 3 4 5 6
not at all relevant highly relevant

APPENDIX J

COGNITIVE SKILL RATING SCALE FOR THE MSCE HISTORY ITEMS

Directions: Please rate the degree to which each item measures each cognitive skill area listed below using a six-point scale where a rating of "1" indicates "does not measure this skill at all" and a rating of "6" indicates "measures this skill very well." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have six ratings for each item.

ITEM #	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Other/ Specify .....
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							

1	2	3	4	5	6
Does not measure this skill at all					Measures this skill very well

APPENDIX J "Continued, next page"

COGNITIVE RATING SCALE FOR THE MSCE HISTORY ITEMS

Directions: Please rate the degree to which each item measures each cognitive skill area listed below using a six-point scale where a rating of "1" indicates "does not measure this skill at all" and a rating of "6" indicates "measures this skill very well." Use the numbers 2 through 5 to indicate intermediate degrees of relevance. You should have six ratings for each item.

ITEM #	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Other/ Specify .....
31							
32							
33							
34							
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							
50							
51							
52							
53							
54							
55							
56							
57							
58							
59							
60							

1	2	3	4	5	6
Does not measure this skill at all					Measures this skill very well

APPENDIX K

SAMPLE CONSENT LETTERFOR VOLUNTARY PARTICIPATION

I volunteer to participate in this study and understand that:

1. I will complete surveys drawn up by Mr. Elias Chakwera consisting of 4 and 6 questions, respectively.

2. The first survey will address my views about the instructional decisions that I made while teaching History to prepare students for the 2003 MSCE examination.

3. The second survey will seek my opinion about what I learned from my involvement in the activities of the common mock test construction.

4. I will voluntarily participate in the construction, administration and scoring of the common mock examination.

5. I will willingly submit scores of my students on the common mock examination to Mr. Chakwera for analyses in his study.

6. My name will not be used in reporting the results of this study.

7. I may withdraw from or part of the study at any time.

8. I will have the right to review the material that I submit prior to final publication of the study report.

9. I understand that results of this survey will be included in Mr. Chakwera's doctoral dissertation and may also be included in manuscripts submitted to professional journals for publication.

10. I am free to participate or not to participate without prejudice.

11. Because the Teacher sample is very small there is a possibility that an individual may have the risk of being identified although the focus of the study would be on shared opinion rather than that of an individual.

_____ _____

Researcher's Signature Teacher's Signature

Date_____ Date _____

APPENDIX L

SUBJECT MATTER EXPERTS' QUESTIONNAIRE ON OVERALL TEST QUALITY

Thank you for accepting to judge the relevance of the multiple-choice items in this set in terms of content coverage and cognitive skills involved. Please complete the questionnaire to express your impression about the quality of the items in terms of their suitability for the MSCE level.

1. How does set __ of questions meet the objectives of the MSCE syllabus?

does not meet objectives___ Less adequately___ adequately___ Very adequately___

2. How does the language and vocabulary used in the items of this set fit the level of understanding for a typical MSCE student?

Does not fit well at all ___ Fits less well ___ Fits well___ Fits very well ____

3. How would you rate the items in this set in terms of task specificity?

Not at specific___ less specific___ specific enough___ very specific____

4. How would you rate items in this set in terms of clarity?

Not at all clear___ less clear___ clear enough___ very clear ____

5. Were there items you think did not belong to the MSCE syllabus? Yes___ No__
   If YES, list the numbers of the items below:

6. Provide any other comment you might have regarding what you have learned from your involvement as a judge of this set of items?

BIBLIOGRAPHY

Aiken, L. R. (1980). Content validity and reliability of single items of questionnaires. *Educational and Psychological Measurement, 40*, 955-959.

American Psychological Association, American Educational Research Association, National Council on Educational Measurement, (1999). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Anastasi, A. (1982). *Psychological testing* (5th Edition). New York: Macmillan.

Angoff, W. H. (1988). Validity: An evolving concept. In Howard Wainer, & Henry I. Braun (eds.). *Test validity* (pp.19-32). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Banda, K. N. (1997). *A brief history of education in Malawi*. Blantyre: Dzuka Publishing Company Limited.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook 1, cognitive domain*. New York: Longman, Green and Company.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20,* 6-18.

Brennan, R. L. (2002). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measuremen,*. 24, 339-353.

Chakwera, E., Khembo, D., Sireci, S. G. (in press). High-stakes testing in the warm heart of Africa: The challenges and successes of the Malawi National Examinations Board. *Educational Policy Analysis Archives.*

Chimwenje, C. (1992). *Evaluation of the Primary School Leaving Certificate in Malawi: How the examination meets its educational and selection goals*. Unpublished doctoral thesis, University of Sussex.

Chulu, B. W., & Sireci, S. G. (2002). *Evaluating the content validity of the Malawi School Certificate of Education Physical Science examination*. Paper presented at the Northeastern Educational Research Association (NERA) Conference, Kerhonkson, N.Y., October 22-26.

Crocker, L (2002). Stakeholders in comprehensive validation of standard-based assessments: A commentary. *Educational Measurement: Issues and Practice, 21*, Spring, 5-6.

Crocker, L & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadworth Group.

Crocker, L., Llabre, M., & Miller, M. D. (1988). The generalizability of content validity ratings, *Journal of Educational Measurement, 25*(4), 287-299.

Crocker, L., Miller, M. D., & Franks, A. (1989). Quantitative Methods for assessing the fit between test and curriculum, *Applied Measurement in Education, 2,* 179-194.

Cronbach, L. J. (1971). Test validation. In Robert L. Thorndike (Ed.), *Educational measurement* (2nd Edition, pp. 443-507). Washington, DC: American Council on Education

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th Edition). New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validity argument. In Howard Wainer, & Henry I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Curreton, E. E. (1950). Validity. In E. F. Lindquist (Ed.). *Educational measurement.* (1st Edition, pp. 621-694). Washington, DC: American Council on Education.

Deville, C. W. (1996). An empirical link of content and construct validity evidence. *Applied Psychological Measurement, 20*, 127-139.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16,* 640-647.

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement, 22,* 15-25.

Ediger, M. (2002). *Assessing state mandates tests.* Opinion paper on National Assessment of Educational Progress. ERIC.

Frye, A. W., Szauter, K. M, Litwins, S. D., & Ofoegbu, E. N. (2002, April). *Validating processes for using curriculum objectives as standards in curriculum evaluation.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA., April 1-5.

Gronlund, N. E. (1968). *Constructing achievement tests.* Englewood Cliffs, N J.: Princeton- Hall.

Gronlund, N.E. (1977). *Constructing achievement tests.* Englewood Cliffs, N J.: Prentice-Hall, Inc.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art.* Baltimore: Johns Hopkins University Press.

Hambleton, R. K. (1984). Validating the test score. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199-230). Baltimore: Johns Hopkins University Press.

Henryson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd Edition, pp. 130-159). New York: American Council on Education & Macmillan Publishing co.

Hopkins, K. D., Stanley, J. C., & Hopkins, B. R. (1990). *Educational and psychological measurement and evaluation.* Englewood Cliffs, N J: Prentice-Hall, Inc.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology, 10*, 93-98.

Jones, B. D., & Johnson, A. F. (2000, April). *The effects of high-stakes testing on instructional practices.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Kane, M. (2002). Validating high-stakes programs. *Educational Measurement: Issues and Practice, 21*, 31-41.

Lane, S. & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*, 23-30.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology 28*, 563-575.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching.* Upper Saddle River, NJ: Prentice-Hall, Inc.

Malawi National Examinations Board. (1999). *Marking error analysis.* Paper presented to the Research and Test Development Committee of MANEB: Author.

Malunga, L. B. (2000). *Presidential Commission of Inquiry into the Malawi School Certificate of Education (MSCE) examination results.* Final Report presented to The President of the Republic of Malawi. Lilongwe: Ministry of Education Science and Technology.

Ministry of Education, Science, and Technology (2002). *Educational statistics.* Lilongwe: Author.

Ministry of Education, Science, and Technology (2000). *Policy implementation framework.* Lilongwe: Author.

Mbunge, J. C. F. (1986). *Content validity analysis of the Junior Certificate of Education (JCE) examinations in Geography and History in relation to school syllabuses and their relevance to everyday life in Malawi: 1975-1985*. Unpublished Master's Thesis, Chancellor College, University of Malawi.

Mchazime, H. (2003. July). *Integrating primary school curriculum and continuous assessment in Malawi.* Paper presented at the Sub-regional Conference on Continuous Assessment, Livingstone, Zambia.

Messick, S. (1988). The once and future of validity: Assessing the meaning and consequences of measurement. In H. Wainer, & H. I. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (3rd Edition, pp. 13 -103). New York: American Council on Education & Macmillan Publishing Co.

Miller, M. D. & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24,* 367-378.

Mkhonta, L.A. (2003, July). *Continuous assessment at primary in Swaziland.* Paper presented at the Sub-regional Conference on Continuous Assessment, Livingstone, Zambia.

Mucheru, O. (2003, July). *Good intentions, difficult challenges: Why comprehensive assessment fail to take root.* Paper presented at the Sub-regional Conference on Continuous Assessment, Livingstone, Zambia, June.

Mwanza, S. J. (1998, August). *The curricular relevance of MSCE Physical Science papers.* Paper presented at the annual meeting for the African Association for Educational Assessment. Lilongwe, Malawi.

Napior, D. (1972). Nonmetric multidimensional techniques for summated ratings. In R. N. Shepard, A. K. Romney & S. B. Nerlove (Eds.), *Multidimensional scaling:Theory and applications in the behavioral sciences, Volume 1Theory* (pp.157-178). New York: Seminar Press.

Ndalama, H. J. (1986). *Examination content analysis and course relevance to everyday life for Malawi secondary agriculture.* Unpublished Master's Thesis, Chancellor College, University of Malawi.

Nelson, L. R. (1973). *Lertap Program.* www.humanities.curtin.ed.au/pdf/edu/april.01.pdf.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In Linn, R. (Ed) *Educational measrement* (3rd Edition, pp. 447-474). New York: American Council on Education & Macmillan Publishing Co.

Nitko, A. J., Amedahe, F., Al-Sarimi, A., Wang, S. & Wingert, M. (1998, April). *How well are the Kentucky Academic Expectation matched to the KIRIS96 assessments, CTBS, and CAT?* (ED 420677). Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Oltman, P. K., Stricker, L. J., & Barrows, T. S. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology, 75*, 21-27.

O'Neil, T., & Sireci, S. G., & Huff, K. F. (2002, April). *Evaluating the content validity of a state-mandated science assessment across two successive administrations.* Paper presented at the annual conference of National Council of Educational Measurement, New Orleans, LA.

Popham, W. J. (1984). Specifying the domain of content or behaviors In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp.29-48). Baltimore: Johns Hopkins University Press.

Popham, W. J. (2003). Commentary: Seeking redemption for our psychometric sins. *Educational Measurement: Issues and Practice, 22,* 45-48.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290-296.

Ryan, K. (2002). Validation in the context of high-stakes assessment. *Educational Measurement: Issues and Practice, 21,* 5-6.

Selemani-Mbewe, C. M. (2003). *Knowledge, attitude and practice of classroom assessment: Implications on the implementation of school-based assessment in Malawi*. Center for Educational Assessment, Research Report Number 474. University of Massachusetts, Amherst, MA.

Shepard L., Schmit, J., & Pugh, R. (1992, April). *Factors influencing high school students change in plans for post secondary education: A longitudinal approach.* Paper presented at the annual conference for AERA, San Francisco.

Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.

Sireci, S. G. (1998b). The construct of content validity. *Social Indicators, 45*, 83-117.

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement 16*, 17-31.

Sireci, S. G., & Geisinger, K. F. (1995). Using subject matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

Sireci, S. G., Wiley, A., & Keller, L.A. (2002). An empirical evaluation of selected multiple-choice item writing guidelines. *CLEAR Exam Review*. 20-26.

Smisko, A., Twing, T. S., & Denney, P. (2000). The Texas model for content and curricular validity. *Applied Measurement in Education, 13*, 333-342.

Steeves, K. A., Hodgson, J., & Peterson, P. (2002). Are we measuring students' success with high-stakes testing? *Education Forum, 66*, 228-235.

Tindall, P. E. N. (1992). *History of Central Africa*. Blantyre: Dzuka Publishing Co.

Thorndike, R. M. (1997*). Measurement and evaluation in psychology and education (*6th Edition). Upper Saddle River, NJ: Prentice-Hall.

Yallow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher 12,* 10-14.