

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF HUMANITIES AND LAW

DEPARTMENT OF ENGLISH POST

GRADUATE PROGRAMME

CONTENT VALIDITY OF THE ETHIOPIAN UNIVERSITY

ENTRANCE ENGLISH LANGUAGE EXAMINATIONS OF

2011/12-2013/14

By

YilmaLegesse

(ID NO. GSR/6040/04)

Advisor

JeylanAman (PhD)

June, 2015 Adama, Ethiopia

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF HUMANITIES AND LAW

DEPARTMENT OF ENGLISH POST

GRADUATE PROGRAMME

CONTENT VALIDITY OF THE ETHIOPIAN UNIVERSITY

ENTRANCE ENGLISH LANGUAGE EXAMINATIONS OF

2011/12-2013/14

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE

STUDIES

**IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF
MASTERS**

IN TEFL

By

Yilma Legesse

(ID NO. GSR/6040/04)

Advisor

Jeylan Aman

Adama, Ethiopia

Acknowledgements

Above all,I am very much grateful and indebted to my advisor Dr Jeylan Aman for his valuable guidance, scholarly criticisms,suggestions and persistent encouragement.Without his constructive comments,timely feedback and dedicated guidance ,the completion of this work would have been un attainable at least and nightmare at most.

I extend my heartfelt thanks to my wife Sister Solyana kifle and my sister Amsal Legesse for the sacrifices they paid in sharing my pains and carrying all burdens during my study period.

A special word of thanks goes to the test experts of EUE English Examinations and my staff members ,particularly to English language teachers of Butajira Preparatory school who were volunteer to give me their precious time to be interviewed and give genuine information.

Finally,I owe a debt of gratitude to Butajira Preparatory school principals, who encouraged me by facilitating conditions to share my burden of duty in my work place, and providing stationery materials during the course of my study.

Table of Contents

ContentsPages

Acknowledgement.....	I
Table of Contents.....	II
List of Tables.....	III
List of Appendices.....	IV
List of Acronyms.....	V
Abstract.....	VI

CHAPTER ONE: INTRODUCTION

1.1. Background of the Study -----	1
1.2. Statement of the Problem-----	3
1.3. Objectives of the Study-----	5
1.4. Research Questions-----	5
1.5 . Significance of the Study -----	5
1.6. Delimitation of the Study -----	6
1.8.Limitations of the Study-----	6
1.7. Definition of Basic Terms-----	7
1.9.Organization of the Thesis-----	8

CHAPTER TWO:REVIEW OF RELATED LITERATURE

2.1. Overview of Educational Testing -----	9
2.2. Communicative Language Testing-----	10
2.3. Teaching and Testing-----	10

2.4. Contents of Language Test-----	14
2.5. The content validity of Tests -----	17
2.6. Language Testing Methods -----	18
2.7. Guidelines for Content Validity-----	20
2.8.Lack of Content validity and Backwash-----	23
 CHAPTER THREE: RESEARCH METHODOLOGY AND DESIGN	
3.1 Research Methodology-----	24
3.2. Research Design-----	24
3.2.1 Data Source -----	25
3.2.2. Sampling Technique-----	25
3.2.3 Data Gathering Instruments -----	26
3.2.3.1. Document Analysis-----	26
3.2.3.2. Interview-----	27
3.2. 4 Data Gathering Procedure -----	27
3.2.5. Methods of Data Analysis and Procedure-----	28
CHAPTER FOUR: DATA ANALYSIS AND INTERPRETATION	
4.1 Results of Document Analysis-----	30
4.2 Interview Results-----	30
 CHAPTER FIVE	
5. Summary, Conclusion and Recommendation-----	50
5.1 Summary of Major Findings-----	50

5.2 Conclusions-----	52
5.3Recommendations-----	54
Reference-----	I
Appendices-----	II

List of Tables

Tables	Pages
Table 1: Profile of Judges	32
Table 2: Major topics and periods Allotments of grades 11 and 12.....	32
Table 3:The Categorization of Test Items of 20011/2-20013/4 E.C in to Syllabus Contents.....	33
Table 4: Chi -square Test Analysis of the Expected and Observed Test Items.....	34
Table 5: Number of Expected and Observed Test Items from syllabus content	35
Table 6:Divission of Syllabus Objectives in to Learning Domains.....	37
Table 7:Division of Grades 11 and 12 Syllabus Objectives in to Learning Domain by judges.....	37
Table 8: Division of syllabus Cognitive Domain in to sub categories and their Rank orders by judges.....	38
Table9:Division of syllabus Cognitive Domain categories averaged by judges	39
Table 10: Categorization of Test Items in to Cognitive Outcomes and their Rank order.....	40
Table 11: The Categorization of Test Items in to Cognitive Sub Domains Averaged Across Judge....	40
Table12:The Number of Expected Test Items from the syllabus Cognitive Domain and the Chi –square values..	41
Table13:The classification of syllabus cognitive domain Categories averaged by Judges.....	43
Table14:Classification of the Expected and observed test items in to Syllabus Cognitive Domain.....	44

List of Appendices

Appendix A: A Guide for the Levels of Activities Based on the Cognitive Domain in Bloom's Taxonomy-I	
Appendix B: Domain Categories-----	II
Appendix C: Reliability Test of the Research Tool-----	XI
Appendix D: Major topics and periods allotments of grades 11 and 12-----	XII
Appendix E: Chi -square Test Analysis of the Expected and Observed Test Items-----	XIII
Appendix F: The Number of Expected and Observed Test Items from syllabus content---	XIV
Appendix G: The coefficient of concordance W and the Observed value of S on the categorization of test items in to the educational taxonomies-----	XV
Appendix H: Division of Syllabus Objectives in to Learning Domains-----	XVI
Appendix I: Division of Grades 11 and 12 Syllabus Objectives in to Learning Domain Averaged by Judges-----	XVII
Appendix J: The coefficient of concordance -W on the categorization of syllabus cognitive Outcomes-----	XVIII
Appendix K: Division of syllabus Cognitive Domain in to sub categories and their Rank orders by judges-----	XIX
Appendix L: The Number of Expected Test Items from the syllabus Cognitive Domain and the Chi – Square values-----	XX
Appendix M: Classification of the Expected and observed test items in to Syllabus Cognitive Domain-----	XXI
Appendix N: Interview for Exam Experts-----	XXII
Appendix O: Interview for Subject Teachers-----	XXIV

List of Acronyms

- **EGSECE:** Ethiopian General Secondary Education Certificate Examination
- **EUEEE:** Ethiopian University Entrance English Examination
- **NOE:** National Organization of Examination
- **SAT:** Standard Achievement Test
- **TOS:** Table Of Specification

ABSTRACT

The purpose of this study was to assess whether the Ethiopian University Entrance English Examinations of 2011/2- 2013/4 administered by the National Educational Assessment And Examinations Agency fairly represent the content coverage and adequate sampling of the objectives stated in the syllabus. To attain this, Mixed Method Approach was used. The qualitative methods was used to justify, clarify and interpret the data collected through semi structured interview and questionnaire while the quantitative method was used to show the outcomes of Pearson's chi-square test of independence and Cramer's coefficient of association. The raw data for the study was collected using syllabus and SAT content analysis, and semi structured interview with 2 testing experts to cross check the data obtained from the document analysis. Purposive sampling and availability techniques were used to select the sources of data and analysts of this study respectively. The content validity of the contents of the syllabus of grade 11 and 12; and EUEE (SAT) were analyzed using the standards of measuring content validity. The collected data were analyzed by using Chi-square Test of Independence and Cramer's Coefficient of contingency to test the goodness-of-fit and strength of association between the contents of the syllabus and SATs respectively. Conclusion was made both testing the null and alternative hypothesis. Finally, the data obtained from the content analysis was triangulated with that of the semi structured interview

The findings of the study reveals that the items of EUE English Examinations of 2011/2-2013/4 do not match with the syllabus content .They also show the existence of a significant difference between the syllabus outcomes and observed test items ,which means the test items do not adequately represent the syllabus outcomes.

Therefore, recommendations were suggested for different stake holders. For instance, the test constructors should note that test items must function as intended .They, the test constructors, were suggested to use not only the syllabus but also the time table of the satellite television program and the Students Text Books along with the syllabus.They should increase the number of items otherwise the curriculum developers reduce the number of syllabus contents .They were recommended to use TOS not only to determine the syllabus contents but allso to the learning out comes.What is more, they should present variety of literatures such as poems and novels for reading comprehension. The government was also recommended to equip secondary and preparatory schools with language learning materials. In order to assess affective and psychomotor skills, continuous assessment should be introduced in a formal way and value should be added on the result of EUEE.

III

CHAPTER ONE: INTRODUCTION

1.1 BACK GROUND OF THE STUDY

Teaching and learning go hand in hand with testing .This is because of different reasons. For instance, test is used to measure the performance, and achievement of students on what they have been taught, to monitor the teaching and learning process, etc. Above all, testing is used to facilitate the teaching and learning process. Therefore, the test should have validity, especially content validity .Regarding this, Hughes (1989) claims that despite the significance of every aspect of validity in teaching and learning, the content validity is the most important one due to its being a means to check the achievement of objectives of the syllabus contents.

If a certain test is valid in terms of its content coverage, it provides relevant information to the concerned bodies with regard to how much students have progressed in each content or syllabus.

In addition to this, tests which involve a proper content validity initiate learners to study committing to each content of the course. Otherwise, students will avoid studying and practicing in those language areas and skills which do not appear at all or appear less in tests. They give considerable attention to those language areas which are going to be tested. Hughes (1989) and Bachman (1990) said that practice for a test will have a harmful effect on learning and teaching so that learners learning style will be on the basis of the test and attempt to memorize knowledge in terms of the way they are going to be tested. Hence, quantity and quality of learning is affected by evaluation.

Therefore, tests should be prepared in a way the students have been taught and what they have been taught. Weir (1993,83) also says, “Students could work well in language areas or skills where tests are highlighted; that is to say, students study and practice more language areas and skills to which more emphasis is given during testing.” By this, Weir implies that considering all contents of the syllabus proportionally during testing facilitates language learning. Lastly, scores gained from tests which involve satisfactory sampling (content validity) can help

to approximate students' actual performance level. That is, they guarantee to draw acceptable statements about learners' proficiency status in a language in a particular grade level.

It is important, therefore, that tests should sample as widely as possible test items from each contents of the domain. As a result, decisions about students' performance level and the program or syllabus, will be more acceptable. By contrast, decisions based on scores gained from tests which are poor at content validity are likely to be less acceptable. Tests with appropriate content validity are helpful to increase quality and quantity of learning providing valid and reliable information about students' performance.

Hence, a study on content validity, especially in relation to standardized achievement test (SAT) is worthwhile and has paramount importance in the quality of education as it contributes to the improvement of the curriculum in general and language education program in particular. In line with this, Sideik (2009:138) asserts, <Achievement tests are directly related with language courses.>

Nevertheless, for years the role of testing in line with language teaching and learning had been given little attention. Baker (1989) justified that because of inability to draw clear distinction between teaching and testing, there was little emphasis to testing in the past decades. This has contributed to poor preparation of national exams though the due attention to testing in the communication era is a step forward for the advancement of good testing and teaching.

Besides, there are scarce studies on content validity in general and in the Ethiopian context in particular. Girma (1997:46) has substantiated that no one has given due attention to content validity of tests in the Ethiopian situation and this may indicate that the Ethiopian national exams were negatively influenced for the past few decades and this has affected the quality of national exams in particular and the education process in general.

Hence, preparing content valid test is always of the greatest significance in the assessment of achievement tests. Therefore, as a matter of importance and urgency, much work needs to be done

in line with the preparation of content valid tests in Ethiopian context and this problem makes the researcher embark on this area.

1.2. STATEMENT OF THE PROBLEM

As far as the researcher's experience is concerned and the experience of many of his colleagues is concerned, most of high school students, especially tenth and twelfth grade students are tending to have exam oriented learning style in language class. That is, their learning style appears to be influenced to a large extent by the contents in the previous Ethiopian University Entrance Examinations (EUEE) and Ethiopian General Secondary Education Certificate Examination (EGSECE). Therefore, the researcher claims that it has been serious problem to language teachers to arouse the interest of students to attend all language classes regularly and attentively. The students withdraw the normal class, especially when language skills are taught but learn best with regularly scheduled exams, EUEE and EGSECE. Due to the fact that the test items are all multiple choice and from cognitive domain, the students cannot present the skills that are not included in the test when the need arises. Regarding this problem, Hughes (1989:62) asserted, << Practice at multiple choice items will not usually be the best way for students to improve their command of a language.>> Hence, they miss the chance to practice and develop their productive skill. As a result, the majority of students tend to have lack of confidence in speaking and writing both in the classroom and outside classroom. It is unlikely to the majority of the students to be successful in subjective type of assessments, and this is called test wash back, particularly harmful or negative wash back. It is said that if the skill of writing, for example, is treated only by multiple choice items, then there is great pressure to practice such items rather than practice the skill of writing itself, and which is clearly undesirable (ibid). It is in order to solve such and the kind of problems that the UEEs and EGSECEs are expected to be valid, especially content wise. Questions should be representative of the whole subject or course to achieve all objectives.

Regardless of this, as different studies reveal, the EUEE and EGSECE, particularly, English exams often lack validity. They usually tend to focus on certain language skills and language elements under emphasizing other language areas and even neglecting skills such as listening. Concerning this, though they are few in number, some local studies were conducted on EGSEC

English examination about eleven years ago with special reference to the old syllabus and textbooks. For instance, Tamirat (2001) as cited in Nigussie (2002)), and Nigussie (2002) himself have conducted their studies on content validity of the first EGSEC English Examination administered in 1993 E.C as compared with the syllabus and instructional objectives of the then textbooks. They found that the content validity of the EGSEC English examination of 1993 E. C. was very weak and divergent from the contents of the syllabus.

The statistical results showed that the test items of the EGSEC English examination were not relatively valid and they did not match with the syllabus contents and their research was confined to only one year and only tenth grade test items, and it failed to see test items of other years and twelfth grade (UEE) to confirm the invalidity of the SAT English examination. Besides, as far as the researchers experience is concerned, majority of the students, i.e., before 2012 were heard saying some items were unexpected. The teachers, too, criticized the item sampling in UEE.

Moreover, as UCAS (2013:21) survey on International qualifications for entry to university announced, Ethiopian Higher Education Entrance Examination considered to be equivalent in standard to GCSE grade with the exception of English language exam.

To sum up, due to the above mentioned reasons and absence of studies on the area with special reference to EUEE in relation to the new syllabus and the Transmission Time table of the Satellite Television programs, the current study attempts to assess the content validity of EUEEs of 2011/12_2013/14.

1.3 OBJECTIVES OF THE STUDY

1.3.1 General Objective

The general objective of this study was to assess whether the Ethiopian University Entrance Examinations (EUEE) English exam administered by the National Educational Assessment and Examinations Agency fairly represents the content coverage and adequate sampling of the objectives stated in the syllabus

1.3.2 Specific Objectives

In specific terms, the study aimed to:

- assess the content coverage of the exam in comparison with the syllabus
- find out about the extent to which the test items match with the expected learning outcomes of the syllabus.
- evaluate the extent to which a test is subjectively viewed as covering the concept it purports to measure.

1.4 RESEARCH QUESTIONS

The purpose of this study was to assess the content validity of UEE English examinations of 2011/12_ 2013/14 with the new syllabus of 11 and 12 grades. Therefore, to attain these objectives, the following research questions would be answered:

- Which content areas of the syllabus have been covered in the UEE?
- Do the items of the exam match with the expected learning outcomes of the syllabus?
- What does the layout of the exam look like?

1.5 SIGNIFICANCE OF THE STUDY

Tests should always measure what is intended in the syllabus. Therefore, from the result of this study, the national exam experts, the language teachers and other researchers may be

benefited. The national exam experts get feedback on their quality and efficiency, while the language teachers get a better understanding of the need for a closer relationship between language teaching and testing.

As a result, tests could be prepared in a way that the items represent the course contents which makes the students give attention for the course contents-skills and language elements which provide opportunity for the learners to practice and develop their productive skills. Other researchers also get input for further study and insight on how to assess content validity of language tests.

1.6. DELIMITATION OF THE STUDY

The objective of this study was to check the content validity of EUE English Exam with the syllabus of the subject and the Transmission Time table of the Satellite Television programs.

Hence, though the generalization of the results of the current study would have been comprehensive if the study had assessed more than three years exam papers, but due to time constraint and the text books and syllabi being newly introduced, the study focused only on three years EUEEs of 2011/2012- 2013/14

What's more, the study would have been thorough if it had assessed the reliability and other validities; however, the current study dealt with only the curricular and face validities to evident content validity.

1.7. LIMITATIONS OF THE STUDY

If all the intended test experts had participated in both the content validity analysis and interview program, more reliable and rich data would have been obtained. However, due to time constraint

and load of work, two interviewees, out of three intended, participated in the interview while non in the analysis

1.8. Definition of Basic Terms

Wash back: The effect of testing on teaching and learning is known as backwash

effect; the proper relationship between teaching and testing is surely that of partnership. It is true that there may be occasions when teaching is good and appropriate and testing is not, we are then likely to suffer from harmful backwash. Backwash can be harmful or beneficial (Hughes,1989:1,).

Content validity:

It is the validity of a test that measures whether or not the EHEEE English items match proportionally to the language components such as the skills, structures, contents and the learning objectives indicated in graded 11 and 112. (Alderson et-al,1995:173)

Language Components(areas):

Refers to grammar (with all its elements,units and classes) vocabulary(list of words used in the text book concerned) mechanics(conservation of capitalization, punctuation and spelling) pronunciation (the ability to make sound,contrasts see and sea (Heaton ,1988)

Language skills: Writing, reading, listening and speaking(Heaton 1988:8)

1.9 ORGANIZATION OF THE THESIS

Basically, this thesis contains five major chapters. In the first chapter, it describes the background, statement of the problem, objectives of the study, research questions, significance of the study, delimitation of the study, limitation of the study and the organization of the thesis. The second chapter deals with review of related literature while the third chapter is about the research methodology and design used to carry out the research. Following this, chapter four presents results, data analysis and interpretation. Finally, chapter five, discusses summary of major findings, conclusion and recommendations. References and appendices are stated respectively and bear towards the end of the thesis

CHAPTER TWO

2. REVIEW OF RELATED LITERATURE

This chapter presents the review of literature on an overview of testing in general, and communicative testing in particular. It addresses the following points: Meaning of testing, purpose of testing, meaning of communicative testing and its difference from conventional testing approaches, relation between language teaching and testing, contents of language test, language testing methods, and the relationship between the content validity and wash back.

2.1 OVERVIEW OF EDUCATIONAL TESTING

Richards (1990) states that tests are means of obtaining systematic evidence on which to base instructional decision. Educators see tests as motivators that stimulate individuals to do their best. If they are well designed and properly used, tests can effectively enhance the educational process, Educational testing is in fact a world endeavor. In everyday life, there is a necessity to use some devices, determining people's ability. It is difficult to imagine, for example, some organizations hiring interpreters without some knowledge of their proficiency. Therefore, tests will be needed in order to provide information about the achievement of the testees, without which we cannot make decisions.

A test lasts from half an hour to one hour and covers one or more units. An exam is two hours or longer and covers at least half of the content of the course. Despite these distinctions, in this paper, test or testing will be used as an umbrella term to refer to any type of measurement procedures. That is, teacher made and standardized tests.

In any consideration of educational testing, a distinction must be drawn between teacher-made tests of the classroom and those formal standardized tests which are usually prepared by professional testing services to assist students' admission to universities. According to Weir (1993), Classroom tests are generally prepared and scored by one teacher. Test objectives can be

based directly on course content; the students know what is expected of them and what is likely to be covered in the test questions. Standardized tests, on the other hand, are designed to be used with hundreds of thousands of subjects throughout the world.

They are prepared by a team of testing specialists without personal knowledge of the examinees. Such tests often take years to construct as opposed to a few days for a teacher-made test. Perhaps the most common use of educational tests is to pinpoint strengths and weaknesses in the learned abilities of the students. Linden et al (1974) consider evaluation of students' progress to be a major aspect of the teacher's job. It gives a sense of where the students are, relative to the curriculum and to other students, as well as how students are progressing toward the attainment of specified objectives. A test is also a tool which teachers need in their evaluative repertoires.

2.2 COMMUNICATIVE LANGUAGE TESTING

Communicative language testing is intended to provide the tester with information about the testee's ability to perform in the target language in certain context-specific tasks. It has to be recognized that given the constraints of time and practicality, only a small sample of the testee's language can be collected, and that however realistic the tasks may be intended to be, the testee's performance will inevitably reflect the fact that s/he was performing under test conditions.

2.3 TEACHING AND TESTING

Testing and teaching are not separate entities. In no way should testing be separated from teaching. Teaching has always been process of helping other to discover "new" ideas and "new" ways of organizing that which they learned, while testing is an integral part of teaching due to the fact that testing has paramount importance in the course of teaching and learning process for different stakeholders.

Testing can be linked to teaching in a number of ways. Supporting this view, Oller (1987:44) described about language tests as follows :<<They serve a number of purposes including instructional, managerial (Providing feedback), motivational (as rewarding Instrument)

diagnostic, and curriculum (helping to define the curriculum (helping to define the curriculum as whole).>>

This means that teachers need to give test for their students in order to know how much student participate, how much effective the teaching and learning approach is, and to what extent the instructional objectives are achieved successfully. Hence, teaching and testing are interrelated activities.

Heaton (1991:5) also explained the firm relationship of these activities. He states, “Both testing and teaching are so closely interrelated that is virtually impossible to work in either field without being constantly concerned with the other.>>

Bachman (1990:3) similarly shows the mutual advantage and interdependence i.e., contribution of teaching to learning and vice versa ,which results in improved learning habits He says,<< Language tests are frequently used as source of information in evaluating the effectiveness of different approaches to language tests can thus provide a useful input in to the process of teaching.>>

Testing is a useful tool at the beginning of the school year. It can help a teacher gain an overview of what student bring to new instruction. Test results early in the school year can help the teacher plan review material and identify potential issues to be faced. Examining past test results can help a teacher who is new to a specific school asses the school setting that he or she will work in. That is, testing is used to assess and evaluate the curriculum in general and the teaching and learning process in particular so that weakness and strengths, and needs will be identified. As result, the teacher designs a plan based on the data or information obtained from the test. (ibid).

Testing can aid in decisions about grouping students in the class. As many scholars agree, a teacher should consider the three type of learners-high, medium and poor, during the teaching and learning process so that every student can access fair learning. If the teacher has specific instructional objectives, testing can yield information that will aid the teacher in assigning

specific students to instructional groups. The teacher can change the groups later after more teaching and testing has taken place.

Testing can be used to diagnose what individual pupils know and what they don't know, to diagnose their strength and weakness. No one source of data can be sufficient to assess what a pupil knows about school-related content. What is called for is triangulation of several kinds of data drawn from various types of tests: standardized tests of achievement and aptitude, teacher-made quizzes, observations of behavior, and the like. Diagnosis does not necessarily mean prescription unless the data collected have demonstrated high reliability and validity.

Testing can help the teacher determine the pace of classroom instruction. Teachers tend to use tests that they prepared themselves much more often than any other type of test to monitor what has been previously learned. These tests may take the form of oral questioning of the class or individual students, or paper-and-pencil tests. Systematic observation of a student applying a skill can be thought of as a form of performance testing. Testing used in these ways are Prerequisite for determining how quickly new material can be presented. These testes help the teacher gain a perspective of the range of attained learning as well as individual competence.

In no way should testing be separated from teaching. Testing should be a reflection of the way we teach. Testing could be mechanism to recognize to what extent our method of teaching is appropriate to the level of the class, to know students weakness and strength in the teaching process and to indicate the general direction of the program.

The purpose of testes and the merit of its connection of language teaching is that language testes can be important source of information about the success of learning and teaching. They can also be used as a source of feedback on learning and teaching, (Bachman, 1990)

Therefore, the appropriateness and constructiveness of the exam is very important, otherwise delude of teachers, students and other administrative may occur, and also invalid testes may lead

students to wrong study habits, and such relationship of teaching and testing is said to be wash back effect, harmful or beneficial .

There is likely to be harmful back wash if the test content and testing techniques are not matching with the objectives of the course. Students' habit is imposed to study only the facts if a test measures only facts and simple level of thinking, which means the test is not truthful measurement of the intended objective and contents of the syllabus, which makes the decision maker to put falls judgments on the students' assessment. And this is negative wash back.

If the test measures, what is supposed to measure, the response shows a good feedback, and which shows that wash back is also beneficial. Accurate decisions about the achievement of the objectives and coverage of contents in the syllabus can be made by the test experts. It may also show that the teaching methods employed are effective and appropriate. This is a positive wash back effect. Alderson and wall (1993) pose 15 possible hypothesis that they hope will lead to the eventual refinement of the wash back construct in empirical investigations (120-121):

- ❖ A test will influence teaching
- ❖ A test will influence learning
- ❖ A test will influence what teachers teach
- ❖ A test will influence how teachers teach
- ❖ A test will influence what learners learn
- ❖ A test will influence how learners learn
- ❖ A test will influence the rate and sequence of teaching
- ❖ A test will influence the rate and sequence of learning
- ❖ A test will influence the degree and depth of teaching

- ❖ A test will influence the degree and depth of learning
- ❖ A test will influence attitudes of content, methods, etc of teaching and learning
- ❖ Tests that have important consequences will have wash back; and conversely
- ❖ Tests that do not have important consequences will have no wash back
- ❖ Test will have wash back on all learners and teachers
- ❖ Tests will have wash back effects for some learners and some teachers

To this end, it is vital to know that a test should be the reflection of the actual teaching and the course being followed. In other words, if a more traditional, structural approach to language learning has been adopted, the test specifications should closely reflect such a structural approach. If, on the other hand, a communicative approach to language learning has been adopted, the test specification should be based on the types of language tasks included in the learning program.

Heaton (1988:13) summarized the discussion on the relationship of the test to language teaching approaches, as “It is clearly unfair to administer a test devised entirely along communicative line to those students who have followed a course concentrating on the learning of structures and grammar.”

To sum up, the purpose of language tests, they may be constructed primarily as an instrument to reinforce learning and to motivate the student or primarily as a means of assessing the student's performance in the language. To measure language proficiency regardless of any language course that candidates may have achieved the objectives of a course study, to diagnose students strength and weakness, to identify what they know and what they don't know, and to assist placement of students by identifying the stage or part of a teaching program most appropriate to their ability.

Testing can be used to help teachers and administrators in a way that they can be used to make promotion and retention decisions; test results are important device to share information with boards of education, parents and the general public through the media. Test results are also useful tools for measuring the effectiveness of instruction and learning.

In conclusion, better tests mean better teaching; better teaching means better learning. A well designed testing system can spearhead educational improvements ,but a poorly designed system can sabotage the most dedicated efforts to improve instruction quality. Thus, if a test is to promote better teaching and learning, it must be consistent with the principles of good teaching and learning.

2.4. Contents of Language Test

The language areas and the instructional objectives (learning outcomes) are the two main categories tested in language. As Heaton (1988) confirms, due to the complexity and close relation to the total context in which they are used as well as too many non-linguistic skills (gestures eye-movement, etc), it may often seem impossible to separate language skills for the purpose of any kind of assessment. The four major skills in communicating through language are often broadly defined as listening, speaking reading and writing. In many intuitions where English is taught for general purposes, these skills should be carefully integrated and used to perform as many genuinely communicative tasks as possible. Where this is the case, it is important for test writer to concentrate on those types of items, which appear directly relevant to the ability to use language for real life communication especially, in oral interaction. Thus, questions which test the ability to understand and respond appropriately to polite requests, advice instructions, etc would be preferred to tests of reading aloud or telling stories.

Regarding the ways of assessing performance in the four major skills, Heaton (1988:8) lists out as follows:

Listening (auditory) comprehension in which short utterances dialogues talks and lectures are given to the tests, speaking ability in the form of an interview, a picture description, role play and a problem solving task involving pair work or group work, reading comprehension in which questions are set to test the students ability to understand the gist of a text and to extract key information on letters reports memos messages -instructions and accounts of past events etc.

As can be deduced from this, it is test designer's task to assess the relative importance of these skills. Nevertheless, Heaton (1988) as cited in Nigussie (2002) remarks it is often awfully difficult to separate one skill from another, for the very division of the four skills is an artificial one and the concept itself constitutes a vast oversimplification of the issues involved in communication. Test of a foreign language should seek for useful information about functional language ability, general language proficiency and some areas of linguistic knowledge, so problems of knowing what to include in test (i.e. sampling) may sometimes arise in designing an achievement test because there is often so much ground to cover. We may thus find it difficult to know what to leave out of the test and what to put.

In line with what to include in achievement test, Hughes (1989) suggests focusing on objectives rather than on detailed teaching and text book content because objectives provide the real image of what has actually been attained, so teaching and learning will tend to be evaluated against those objectives. As a result, there will be constant pressure to achieve them.

Regarding the assessment of language contents, skills and linguistic areas are cognitive domains. These are intended curriculum objectives expected to be achieved at a certain stage (end of educational plan(phase)(Ibid).

According to different educators, the learning outcomes are classified in to different categories for instructional purpose. For example, according to Bloom and his associates, they are classified as the psychomotor cognitive and affective domains. Perhaps the best known taxonomic classification of learning outcomes in education is those developed by Benjamin Bloom (1956),

David Krathwoh (1964) and Anita Harrow (1972). Bloom and his colleagues developed the cognitive domain, Krathwoh and others the affective domain and Harrow was responsible for the psychomotor domain, (Mekonnen, 1992).

The cognitive part of the intended curriculum, however, is usually given emphasis in the assessment of national exams such as the achievement tests of the EGSEC and EUE English examinations in spite of the three domains in the syllabus. In line with this, Meherens and Lehman (1992) in Mekonnen (1992) support the view explaining as follows.

Achievement tests-be they teacher –made or standardized- are not designed to measure, even remotely, the affective components of our education enterprise. They do not tell us anything about whether Allen or Ilene is interested in school, enjoys reading, or can interact effectively with peers. What achievement tests do, however, measuring the cognitive aim of instruction in an objective manner-is to accomplish more validity and reliability than is possible with teachers' subjective judgments.

From this, it is quite clear to conclude that the cognitive domain categories are the main aspect to be assessed than the affective and psychomotor domain in achievement tests. Therefore, due to the fact that the Bloom's and his associate's taxonomy is the most popular and widely used one, and it minimizes the degree of overlap occurring between the objectives as well as the contents, and also it is advocated for analysis, the study undergoing tends to analyse the exams based on cognitive educational domain.

Generally speaking, what has to be tested is determined based the syllabus, texts and other instructional materials designed in the curriculum. Concerning this, Canale and Swain (1980:34) have proposed how and what to test as follows:

Communicative testing must be devoted not only to what the learner knows about the second language and about how to use it (competence) but also to what extent the learner is able to actually demonstrate this knowledge in a meaningful communicative situation.

2.5. The Content Validity of Tests

As discussed above, and Gronlund(1990) supports as well, the extent to which the interpretation of scores will be appropriate, meaningful and useful for the intended application of results is the most significant point to be considered during the construction or selection of achievement tests and other instruments of evaluation. This is because tests and other instruments of evaluation give different uses, for example, achievement test can be used for selection and placement, diagnosis or certification of mastery, while aptitude tests might be used for predicting success in future learning activities or job.

According to Henning (1987) and Finecchiare and Sake (1938), every measurement should have such characteristics as validity, reliability and comprehensiveness. Besides, Harrison (1983) adds practicality and usability as another feature of all measurements.

Regarding validity in relation to testing and evaluation, Gronlund (1990:48-50) noted down the following points:

Validity is a matter of degree, it does not exist on an all- or none basis. Consequently, we should avoid thinking of evaluation results as valid or invalid. Validity is best considered in terms of categories that specify degree, such as high validity, moderate validity and low validity. Validity is always specific to some particular use or interpretation. No test is valid for all purposes. This is because evaluation results have a different degree of validity for each interpretation to be made.

Content validity, which is the major concern of the study, is one of many ways of gathering evidence to support or challenge the validity of an interpretation of test scores. Adding on this, Hughes (1989:22) Affirm the importance of content validity as “the greater a test’s content validity, the more likely it is to be an accurate measure of what is supposed to measure.”

The word validity is defined by different intellectuals. To mention some, scholars like Weir (1990:119) and Alderson (1981:15) defined content validity as “The extent to which the Selection of tasks are representative of the larger set ” Henning (1987:89) also explained validity as “ validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure.”In other words, Alderson (1995) briefs that the test should have adequate sampling or representation of the content. This makes the test content valid. Supporting this,(Weir 1990) said that a test will have content validity if it simulates the dimension of observable performance. Therefore, in achievement tests, greater attention is given to content validity than such validities as predictive validity, construct validity, face validity and concurrent validity. Supporting this view, Aggarwal (1998:267) states the following:

Content validity is important primarily for measures of achievement. The test maker first determines the widely accepted goals of instruction in the subject and then prepares a blue print for the test. Test content is drawn from the course content and weighted according to the weight age of the objectives of the course and the course content.

From this, this, therefore, it can be concluded that the content validity of a test is not designed haphazardly; it rather needs a thorough and in depth examination of the actual test tasks. In line with the way and the care to be taken on content validity, especially for achievement tests, different scholars suggested guidelines. They have been discussed here under.

2.6 Language Testing Methods

Influential information about the content and method of teaching is conveyed to such stakeholders as teachers, parents and students by examination and national assessments.

Nevertheless, according to Baker (1989), the idea of testing was not given attention before the Second World War. There has not been clean understanding about the difference between teaching and testing until the last decades. Such means as composition, translation, dictation,

etc were used for both assessing and teaching the learners. Briefly speaking, the teachers used to use the same mechanism both to teach and assess learners.

However, since the time that applied linguists, curriculum designers and language teachers were interested in the role of testing, testing has got an increasing credit for one thing since education has become an engine for economic growth and for another thing it has paramount importance as the gateway mechanism for learners, a means for monitoring educational quality and efficiency of a nation's schools. Hence the system, knowledge and approach develop with it. Therefore, when the teaching methods change, the testing mechanisms also change with it.

Thence, four major methods of language testing such as the essay translation method, the structural method, the integrative and the communicative method are categorized by Heaton (1988).

Among the four methods, the essay translation method is known as the pre-scientific stage of language testing. During this time, students were evaluated or judged subjectively. There was no trend of testing students like these days. Hence, writing, translation and grammatical analyses were commonly used. Consequently, memorizing lists of language patterns and labeling parts of a sentence were expected from the learners. What is more, the tests used to lack validity. During this time, as can be deduced from this (the above discussion), the use of testing for language teaching and teachers as well was ignored (Madson, 1985).

Besides, in line with the development stages of the language testing, Spolsky (1978) forwards the pre scientific, the psychometric structuralism, psychometric and socio linguistic. Due to the high demand of using integrative tests and mastering elements and skills of language, the methods and analysis of contents of the psychometric structural approach were criticized severely. During this time, the kind of description of language use was better than structural analysis (Baker, 1989).

Similarly, the great deal of attention on the analysis and measure of the target language, which is phonology, vocabulary and grammar, was criticized by Heaton (1988). However, the advocates

testing language contextually. As a result, meaning and total communicative effect of discourse are the main areas of interest, which characterize the integrative approach.

Therefore, the learners are assessed on the bases of their using more than two skills at the same time rather than in isolation. In accordance with this, the very important thing in integrative testing is not assessing learners based on the purpose of language proficiency.

In line with this, Heaton (1988) explains that integrative test is characterized by the use of functional language, closed testing and dictation, oral interview, translation, and essay writing.

After the integrative approach, the current approach, called the communication approach, which is concerned with the evaluation of language use instead of mastering language form come to existence. Consequently, communicative language teaching got opportunity to be practiced (Weir, 1990).

According to such intellectuals as Heaton (1988), Hughes (1989), Bachman (1990) and Weir (1990) is more preferable to the formal approaches due to the facts that primary concern is using language in real situation. That is way the communicative approach language testing is intended to assess the students in an authentic way, or in a way they encounter is real life rather than just testing the ability of analyzing language forms as was the case with the psychometric structuralism evaluation. Therefore, the nature of the EGSEC and EUE English examinations, nowadays, needs to be communicative so that it coincides with the intended objectives in the syllabus.

2.7 GUIDELINES FOR CONTENT VALIDITY

Among the fundamental features of any test, as discussed above, content validity is the first. Tests of a good content validity, as many scholars commonly agree, facilitate learning language and other subjects. Hence, Walelign (2006) and Anastasia (1975 & 1990) forward some very significant guidelines for test writers to give a great deal of care for writing tests. They suggest such guidelines, which read as: the behavior domain to be tested must be systematically analyzed

to make ascertain that all major aspects are covered by the test items in the correct proportion. The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared. Finally, content validity depends on the relevance of individual's test response to the behavior area under consideration, rather than on the apparent relevant of item content. Thus, language test writers are advice to go through these guidelines in advance in order to produce test which involves adequate content validity since content validity is one of the key promoters learning language. If the test designers do not follow the suggested guidelines, it is very likely that they would forget to include relevant language elements in to a test. Even the included ones would lack proportionality, and some language areas will be dominant in the test, whereas others will get little attention (Ibid). Therefore, Walelign (2006) suggests that construction of content valid test should take in to consideration the major topics or contents and types of behavioral changes in terms of their relative importance, build TOS (Table of specification) from the weighted lists of topics and behavioral changes, select items in line with TOS's

The closer the test corresponds to the specification indicated, the greater the degree of content validity of an examination should be assessed by relating eat to the course objective, contents of syllabuses represented in the examination and whether these are equally distributed or not.

Content validity is a professional judgment ,which the teacher or tester has .They rely on their knowledge of the language to judge to what extent the test provides satisfactory sample of syllabus, whether real (for achievement testing) imagined (for proficiency of testing)or of the theory or model (for aptitude testing)(Davies 1990, as cited in Nigussie ,2002:22)

The content validity of a test is judged on the specification and the test content, which, according to Alderson et-al (1995:173), involves “Gathering the judgment experts, people whose judgment is prepared to trust, even if it agrees with one's own, and the common way to analyze the content of a test and to compare it with a statement of what the content ought to be.

Such a statement may be test's specification, it may be a formal teaching syllabus or curriculum, or it may be a specification.

Weir (1990:24) also said “ To ensure content validity during test construction particular attention must be paid to content validity in an attempt to ensure the sample of activities to be include in a test is as a representative of the target domain as possible . “ Thus, a test is said to be content valid if it effectively samples the language areas and skills from the already stated instructional objectives, Heaton (1988:160) rephrase the concept of content validity as follows:

This kind of validity depends on a careful analysis of the language being tested and of the particular course objectives. The test should be so constructed as to contain particular course objectives. The test should be so constructed as to contain a representative sample of the course, the relationship between the test items and the course objectives always being appropriate.

To sum up “A test is said to have content validity if its content constitute representatives sample of the language skill, structures etc with it is meant to be concerned. Heaton (1988:49) and Hughes (1989:22). The point here is that sampling now assumes paramount importance i.e. test must cover an adequate and representative section of those areas and skills it is desirable to test.

If all the students who take the test have followed the same learning program, they can simply choose areas from program seeking to maintain a careful balance between tense forms, prepositions, articles, lexical items, etc.

Above all, the kind of language to be tested would be the language use in the class room and in the students immediate surroundings, or the language required for the school or the work for which the students is being assessed. On the other hand, content validity assessment depends on how well it samples the learning targets. Making out the important learning objectives clearly and being certain that the sampling is well-done by the assessment procedure are the two

principles for constructing valid tests. Therefore, giving surety whether what we have thought is emphasized by our assessment procedure, the outcomes specified in state's curriculum framework are accurately represented by our assessment tasks, alignment of the assessment tasks and the current thinking about what to teach and how to assess it and the importance of the content in the assessment procedure for learning are the main issues to be considered when evaluating assessment in a particular relation to the representations and relevance of exams,(Nitko,1996).

It is advised to feature the area which has been spent plenty of time significantly in the assessment because the learners anticipate exams to come out from what they were taught with emphasis. What is more, the learning objectives which are appropriately identified and were taught in the classroom should be mirrored by the assessment method. Therefore, if the contents and objectives in the syllabus are represented by their sample, then a test is said to have content validity.

It is, therefore, concluded in one sentence as, “The test would have content validity only if it included a proper sample of the relevant structures,” Hughes (1989:23).

2.8 LACK OF CONTENT VALIDITY AND BACKWASH

Backwash is an important concept that refers to the impact and influence the test can have on the teaching and learning process. This influence can be positive or negative. The backwash concept is very much connected with the test content validity. If the test lacks this quality, it consequently yields negative wash back on both the teacher and the learner, and vice versa (Siddiek, 2010). The key concept of content validity sometimes known as logical or rational validity can be briefly defined as the extent or adequacy with which the test items adequately and representatively sample the content areas in the syllabi or students book, then we can say that the test is content valid, and if the interaction is less ,we can say that there is low content validity .

In the then teaching and learning process, testing might have a positive or negative impact based on the representativeness of contents of the textbooks in the exam. If a standardized achievement test is not rich in terms of content validity, it often has a harmful wash back effect on teaching and learning and fails to measure accurately whatever it is intended to measure (Hughes, 1988). In other words, teaching which is based on student centered approaches mainly stresses the use of the four skills to maximum that leads to the preparation of test items that requires better test items that require high level of mental effort. Hughes (1989) and Weir (1990) suggest to offer much support to the teaching and learning process. A good test has to satisfy three major criteria: a test should be content valid-test what it is intended to measure, reliable-scores should be consistent when repeated, and practical- a test should be economical, easy to administer and score. This process is known as wash back effect. Hughes (1988:1) defined the word as, <<The effect of testing in teaching and learning.>>

CHAPTER THREE

3. RESEARCH METHODOLOGY AND DESIGN

3.1. RESEARCH METHODOLOGY

This research was carried out to assess the content validity of EUE English Examination of 2011/12-2013/14. In other words, it was a contrastive study over the contents and learning outcomes of grades 11 and 12 syllabus and that of the sample examinations. Therefore, the sample documents (EUE English Examinations of 2011/12-2013/14) were analyzed based on the syllabus of grades 11 and 12 and guideline of blooms taxonomy of cognitive domain. The data were interpreted using inferential statistics testing both the null and alternative hypotheses. The results were triangulated with that of the interview obtained from test experts and subject teachers selected using availability and comprehensive sampling techniques respectively. What will follow is a detailed description of how the study was conducted.

3. 2 RESEARCH DESIGN

In the current study, mixed research design was used. The study employed two data collection techniques, namely document analysis and interview. The use of the two methods was meant to triangulate data. That is, to collect information using a variety of methods. Maxwell (1996) points out that triangulation reduces the risk that your conclusions will reflect only the systematic biases or limitations of a specific method, and it allows you to gain a better assessment of the validity and generality of the explanations that you develop. Hence, Qualitative data analysis techniques were used to interpret, clarify, and justify the data collected through interview and document analysis. Quantitative approach was also used to show the outcome of the inferential statistics such as the Kendall's W Coefficient of Concordance, which is used to check the inter rater agreement so that the reliability of the research tools was tested, chi square test of independence and Cramer's V coefficient of association to test the goodness-of-fit and strength of

association between the contents and objectives of the syllabus of Eleventh and Twelfth grade in comparison with contents of the exams, respectively.

3.2.1 DATA SOURCE

The analysis was between contents in the syllabus with the test items of the EUEE English examinations of 2011/12-2013/14 .Therefore, the basic data sources were grades 11 and 12 syllabus along with Transmission Time table of English Satellite Television programs ,which show the time allotment of all the plasma and non-plasma lessons of the courses of the two grade levels , and the test items of the EUE English Examinations. What is more, the experts of National exam and Language teachers of Butajira Preparatory School were also sources of data.Since it is believed that textbooks of grades eleven and twelve are designed based on the syllabus,the researcher used them along with the syllabus.The reason the national exam experts and the language teachers of Butajira Preparatory School were chosen was due to the fact that the exam experts were responsible bodies and the language teachers of Butajira Preparatory School were easily reached by the researcher because of his being the staff member.

3.2.2 SAMPLING TECHNIQUES

In order to get the data that was used to triangulate ,or make sure that the data from document analysis was valid, interview was employed on two English language test experts who were selected by using availability sampling technique,and five English language teachers of Butajira Preparatory School were chosen comprehensively. The reason these sampling techniques were employed was because, in the case of language experts (though three were purposely intended to be interviewed), two experts were available and were able to give adequate information , and in the case of Language teachers, they were few .Due to the inconvenience of time and business of subjects,only one curriculum developer and two experienced teachers were selected purposively as content validity analysts.

3.2.3 DATA GATHERING INSTRUMENTS

Analysis of the contents and objectives in the syllabus were based upon the classification of contents in the EUE English Examinations and the cognitive domain of Bloom's Taxonomy respectively. The contents in the syllabus consist of such major areas as reading comprehension, Vocabulary, grammar, listening, speaking (which is referred as communicative activities in the examination booklets and in the current study as well) and writing. The cognitive domain consists of six levels.

The lower order thinking skills include the cognitive levels of knowledge, comprehension, and application while the higher-order thinking skills include the three cognitive levels of analysis, synthesis, and evaluation. Therefore, the following data gathering instruments or research tools were needed.

3.2.3.1 DOCUMENT ANALYSIS

In order to code the documents such as the sample examinations and the syllabus selected for the current study in to major content areas, the three taxonomies of educational domain based on learning outcomes and cognitive domains, three forms or coding sheets along with the guide line of Bloom's Taxonomy of Cognitive Domain (prepared following research on the internet) were needed and decided by content analysts. (Appendix A)

In order to establish the validity of the research tool, particularly the Guideline of Bloom's Taxonomy of Cognitive Domain, it was presented to a committee of experts headed by three judges. The committee was asked to examine the definitions of the various levels in accordance with the skills and behaviors demonstrated by each level. The judges assessed that the research tool was valid for use in this study.

Regarding its reliability, Inter-Rater Reliability was established in such a way that the assistant analysts randomly selected 140 syllabus learning outcomes of cognitive domain and categorized them according to the cognitive domain of Bloom's Taxonomy. The agreement coefficient

between the findings of the three analyses was computed and determined using Kendall's coefficient of concordance and the agreement coefficient was established as 0.8, which is acceptable and the reliability of categorization of the learning outcomes according to the cognitive domain of Bloom's Taxonomy was high. Therefore, the research tool used by the analysts was found to be reliable (Appendix-C)

3.2.3.2 INTERVIEW

In addition to the documents analysis, interviews were conducted with test experts in NOE office and teachers. The purpose of the interview was to get information about the procedure the exam

experts followed during the preparation of EUEE English exams and the factors affecting the content validity of these exams and to get feedback from teachers' opinion on last years test.

Therefore, semi structured interview was used as tool of data gathering. Each particular interview question had several sub questions (probes) aimed at answering the basic questions. (Appendices N & O)

3.2.4 DATA GATHERING PROCEDURES

To assess the content validity of the 2011/12-2013/14 EUE English Examinations, the following data gathering steps were followed. The data was collected in three phases. The first and the third phases were carried out by the researcher whereas the second mainly by analysts as follows:

In the first phase, the contents of the syllabus were classified into such major headings or sections as reading comprehension, vocabulary, grammar, Communicative activities, listening and writing, then each of the EUEE English examination items and syllabus contents were categorized Under the major headings.

In the second phase, the syllabus objectives were classified in to the taxonomy of Educational domain, and then the cognitive objectives were listed from the syllabus and so were from the test.

Following this, the lists of identified syllabus cognitive objectives were categorized in to such cognitive sub categories as knowledge, comprehension, application, analysis, synthesis and evaluation.

Finally, Interview was undertaken to the subject teachers and test experts in the NOE office in English. That means, interview schedules were prepared for this purpose in advance. In order to give maximum attention to the interviewee rather than concentrating on taking notes, the interviews were recorded. However, the researcher also took notes while the interviews were underway and immediately after the interview sessions, every effort was made to minimize the

effect of using audio recording upon the interviewees' responses, and it was explained to the interviewees that the recorded interviews were to remain in the researcher's possession.

3.2.5 METHODS OF DATA ANALYSIS AND PROCEDURES

Phase I: The Document Analysis Phase

1. A summary table was prepared based on the major topics (contents) and period allotment of Grade eleven and twelve, and their sum was taken as total periods.
2. After allocating the amount of periods spent for each main heading, the expected number of questions from these headings was computed.
3. A summary table was prepared based on the syllabus objectives which indicate frequencies (in figures and percentages) of the taxonomy of educational domains,
4. A summary table was prepared following the analysis of the syllabus indicating frequencies (in figures and percentages) of the cognitive learning outcomes.
5. A summary table was prepared following the analysis of the examination paper indicating frequencies (in figures and percentages) of the major contents.

6. A summary table was prepared following the analysis of the examination paper indicating frequencies (in figures and percentages) of the learning outcomes
- 7 . The degree of the relationship between the frequencies from the analysis of the content and the learning outcomes of the syllabus and from the analysis of the examination questions was tested by the chi-square statistical method .
8. To determine a significant relationship between the observed (empirical, or frequencies that occurred in each language content areas of the SAT) and expected (theoretical, or frequencies that were expected to occur in the SATs) number of questions, an inferential statistics, known as Pearson's chi-square (χ^2) test of independence was used. This value was interpreted based on the null (there is no difference between the expected and observed variables) and alternative hypothesis (there is difference between the expected and observed number of questions). 0.05 level of significance was taken for decision- making.
9. Then, to determine the strength of association between Contents of the syllabus and sample SAT of this study and to enhance the reliability of the study, Cramer's V coefficient of Contingency was preferred among the χ^2 based measures of association.

Phase II: The Interview Analysis Phase

After the semi structured interviews were carried out with subject teachers and test experts respectively, attempts were made to identify and sort out similar patterns and themes of responses of subject teachers and test experts, then meanings of the responses were condensed and structured using narrative analysis. Finally, the results of the semi structured interviews were made to relate with the statistical obtained results.

CHAPTER FOUR

4. RESULTS, INTERPRETATIONS AND DISCUSSIONS

This chapter of the study deals with the results, interpretations, and discussions of data. The results, interpretation and discussion were done using the chi-square statistical methods and qualitative description to analyze the degree of content validity of the items of EUE English examination of 2011/12-2013/14.

4.1 RESULTS OF DOCUMENT ANALYSIS

To realize the objective of the study , the collected data such as the number of period allotment, the learning objectives from analysis of documents (syllabus of grades 11 and 12, time table of the Satellite Television Program and EUE English examination papers of 2011/2-2013/4) were observed, analyzed and calculated with relevant statistical methods .In other words ,the documents were the compulsory materials so that the goodness of fit of the contents and the cognitive domains of the syllabus were tested using the chi-square distribution, Cramer's V and Kendall's W .

The syllabus of eleventh and twelfth grades together consist of 24 units (12 units for each), with each unit divided in to two parts ,plus four revision units and an end of year examination .The contents are intended to be covered in 301 periods with and without plasma television (real teacher) according to the satellite television program. The intended syllabus objectives of both grade levels are 436 in number.

The syllabus address in the introduction part that the course is designed for third year secondary school in order to develop students' understanding and use of English both as a subject and as a medium of instruction .Grade 11 and 12,therefore, focuses on all four language skills equally, developing students' survival skills as well as building confidence and learning strategies, and also it aims to stimulate students' interest and curiosity.

Such topics as reading, vocabulary, grammar, speaking, listening and writing are the major ones though there are sections for study skills and self-assessment targeting on the language elements and skills mentioned above. Study skills encourage students to apply these skills not only to their English lessons, but to their lessons in other subjects as well. Based on the topic (content) of discussion, these sections were grouped to the major content areas

.However, it is unusual to test students while they are really listening, speaking and writing in Standardized Achievement tests. In addition, the term <speaking> is used as <Communicative Activities> in EUE English Examination booklets whereas <writing> remains the same. Therefore, the ongoing study used the same language, and the term <communicative activities> in the syllabus refers to the use of a set of generally agreed-upon principles or/and different languages, for example, social expressions, registers, etc either in real or simulated or both situations, while only in simulated (creating a situation in written form, for instance dialogue, that students encounter in the real world) in the examinations, and the term <writing> in the EUE English Examinations used as writing aspects such as conventions regarding genres of writing, organization of writings, mechanics (punctuation, capitalization, spelling) etc (Richards, 2006)

With regard to content analysis, such scholars as Underhill (1991), Alderson Clapan and Wall (1995) note that to determine the content validity of a certain test, it is good to categorize the syllabus objectives into major content areas first, and then determining the weight given to each content area in the syllabus and in the sample examinations based on the period allotment. Hence, the major content areas of the syllabus and their weight in terms of the period allotment have been tabulated first.

Whereas the syllabus contents were classified into the major sections by the researcher of the ongoing study, the syllabus objectives were classified into the Taxonomy of Educational Domain and the items of EUE English Examinations into cognitive outcomes by three judges, who were lecturers at different Universities. The Profile of the analysts is presented below. The categorizations were established adapting that of Nigussie (2002:45).

Table 1: Profile of Judges

Judges.	Qualification	Experience in year	Sex
Judge 1	B.A in English	20 years in teaching	Male
2	B.A in English	10 years in teaching	Male
	MA in TEFL	6 years in teaching	
3	BA in English	18 in teaching	Male
	M.A in Educational Psychology	3 years as curriculum developer	

As can be seen from Table 1, the total number of the judges is three. Two of them have qualification of M.A in TEFL and Educational Psychology while one has B.A in English. All of them are qualified experienced in teaching for at least 16 years. Besides, one of the judges has worked as curriculum developer for three years.

Table 2: Major topics and periods allotments of grades 11 and 12

Major topics	Period allotments		Total	%
	Grade			
	11	12		
Reading comprehension	25	26	51	16.9
Vocabulary	22	20	42	14
Grammar	26	14	40	13.2
Communicative activity	36	37	73	24.3
Listening	24	23	47	15.6
Writing	25	23	48	16
Total	158	143	301	100

As can be seen from Table 2 above, communicative activity has been given a considerable number of periods(24%),which is followed by reading comprehension(16.9%),listening(15.6%), vocabulary (14%)and grammar(13.2%). The weight given according to the time table of the Satellite Television Program to the selected syllabus contents is different from the contents' weight in the syllabus. This indicates that there is disparity between the newly introduced Time Table and the syllabus where all skills are given equal emphasis. It can be inferred from the distribution of periods that communication has been given due emphasis so that the nature of the exam should be communicative.

Table 3: Categorization of Test Items of 2011/12-2013/14 in to syllabus contents

Contents	Total questions	%
Reading Comprehension	42	11.7
Vocabulary	88	24.4
Grammar	86	23.9
Communicative Activity	60	16.7
Listening	-----	-----
Writing	84	23.3
Total	360	100.0

As it can be seen from Table 3 above , out of 360 items ,24.4% is covered by vocabulary section ,while 23.9% and 23.3% by grammar and writing respectively. They have been given almost the same attention .However, such contents as reading comprehension and Communicative activities have been deemphasized accounting only for 11.7% and 16.7% respectively. Above all, listening has been neglected at all. Grammar, next to communicative activity, has been emphasized much more than the remaining contents of the syllabus. This is contrary to many scholars' suggestion due to the fact that it is not recommended to give more emphasis to structure and recall of fact on grammar in achievement tests at the expense of others.

Table 4: Chi -square Test Analysis of the Expected and Observed Test Items

Major topics	Period allotted (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Observed Test Items (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Total
Reading Comprehension	51 (42.3)	8.7	1.8	42 (50.7)	-8.7	1.8	93
Vocabulary	42 (59.2)	-17.2	4.99	88 (70.8)	17.2	4.99	130
Grammar	40 (57.4)	-17.4	5.27	86 (68.8)	17.2	5.27	126
Communicative activity	73 (60.6)	12.4	2.5	60 (72.4)	-12.4	2.5	133
Listening	47 (21.5)	25.5	30.2	----- (25.5)	-25.5	30.2	47
Writing	48 (60.2)	-12.2	2.47	84 (71.8)	12.2	2.47	132
Total	301	0	47.23	360	0	47.23	661

$$X^2 = \sum \frac{(O-E)^2}{E} = 47.23 + 47.23 = 94.5$$

$$p < 0.05 \quad V = 0.378 \text{ (See Appendix E)}$$

E

Table 4 reveals the chi-square test of independence of expected and observed test items of the EUE English Examinations of 2011/12-2013/14 in terms of time allotments to syllabus contents.

It indicates that the calculated chi-square value is 94.5 and the Crammer's coefficient of concordance-V is 0.378. This illustrates that the EUE English Examinations of 2011/2-2013/4 were significant at 0.05 significant levels with a given 5 degree of freedom. The chi-square value has to be at least ≥ 11.070 .

This leads to the conclusion that there is a significant difference between the observed and the expected contents of the test items, and the value of Cramer V, being < 0.05 indicates a weaker association between the contents of the syllabus and sample examinations.

Table 5: The Number of Expected and Observed Test Items from syllabus contents

Content Areas	Observed	%	Expected	%
Reading comprehension	42	11.7	50.7	14.1
Vocabulary	88	24.4	70.8	19.7
Grammar	86	23.9	68.8	19.1
Communicative activity	60	16.7	72.4	20.1
Listening	---	----	25.5	7.1
Writing	84	23.3	71.8	19.9
Total	360	100	360	100

As can be seen from table 5 above, 24.4% of the test items are represented by vocabulary, which is followed by 23.9% grammar and 23.3% writing. The remaining contents such as communicative activity and reading comprehension represent 16.7% and 11.7% respectively. This shows that both grammar and vocabulary have been given more attention than the skills.

This implies that grammar and vocabulary were considered as a mechanism effectively evaluating students' language proficiency. Which (i.e., the magnitude given to grammar) violates the recommendation forwarded by many scholars that the emphasis to structure and simple recall of hard facts on grammar should be avoided in such standardized achievement tests. In favor of

this,(Jessica Wu (2009) ,as cited in Mekonnen. et.al (2014) noted <<Over emphasizing on the grammatical accuracy in an exam brings general lack of ability to communicate in English due to <old fashioned > approaches in English education.>>

Reading comprehension and which were Communicative activity given 14.1% and 20.1% weight respectively in the syllabus are reduced in number and percentage to 11.7 % and 16.7% respectively.

Listening, which represents 7.1% in the syllabus hasn't been represented in the exams at all. This biased testing practice can lead to the biased or inappropriate way of studying and learning the ignored skill or area of language in the testing system. This can potentially affect their global language progress. According to Abraham (2009), testing practice of testers by itself forces students to decide on what to study or not.

As confirmed by the chi-square value and Cramer's V in Table3 and the frequency of test items in Table 5 above, there is a great disproportion between the fit values and the observed number of questions and this leads to a conclusion that the contents of the sample examinations are not representative sample of the contents of syllabus, and this may results in negative wash back effect on the students' methods of learning the language.

Content validity is not measured only in terms of the extent to which the content areas of the test items represent the syllabus contents; it is also measured in terms of the representation of the syllabus learning outcomes in the test items. Therefore, the extent to which the learning domains of the test items match with that of the syllabus must be determined.

In order to realize this, the learning domains of the syllabus and the magnitude of their representation in the test has been displayed in the table below.

Table 6: Division of Syllabus Objectives in to Learning Domains

Judges	Learning Domains				Rank			Total
	Affective	Psycho Motor	Cognitive	Unclear	Affective	Psycho Motor	Cognitive	
Judge 1	38	33	365	—	2	3	1	6
Judge 2	36	34	360	6	2	3	1	6
Judge 3	32	30	370	4	2	3	1	6
Total	106	97	365	10	6	9	3	18

*S=18 *W=1

In the table above , the raters ranks to the three learning domains are shown .The Kendall's W is associated with critical values of S =18 as a measure of concordance for K=3 and n=3 ,and was found to be significant at 0.05 level, which shows that the raters have a perfect agreement.

Table 7: Division of Grades 11 and 12 Syllabus Objectives in to Learning Domain Averaged by Judges

Division of Learning Domain	Average Total	%
Affective	35.3	8.2
Psychomotor	32.3	7.4
Cognitive	365	84.4
Total	432	100

Table 7 illustrates the distribution of the grades 11 and 12 syllabus objectives in to learning domain .Hence, cognitive domain has 84.4% opportunity .However, affective and psychomotor domains are given 8.2% and 7.4% respectively. From this ,it is concluded that the syllabus of the

two grades ,11 and 12 give much greater attention to the two learning domains, and this proves that the syllabus don't give equal emphasis to the domains. The affective domain which typically targets the awareness and growth in attitudes, feelings, emotions and values and the psychomotor domain which incorporates doing and motor skills are highly undermined. Hence, such extent of considerations to Psychomotor and affective domains tend to be inadequate to increase the students' language skill, emotions, sensitiveness and ways of expressing themselves ,This awaken the concerned bodies such as the syllabus designers to revise and check the weight given to each taxonomy of educational domain.

Nevertheless, due to the fact that, as different scholars agree, it is ideal ,i.e. not mostly practical to incorporate items from such domains as affective and psychomotor in Standard Achievement Tests. This study, therefore, analyze the content validity in line with cognitive domain. Here under, table 5 shows the distribution of the sub categories of cognitive domain in eleventh and twelfth grades syllabus.

Table 8: Division of syllabus Cognitive Domain in to sub categories and their Rank orders by judges

Sub-categories of cognitive domain	Judge 1		Judge 2		Judge 3		Rank sum (SR)	(SR) ²
	Learning Outcomes	Rank	Learning Outcomes	Rank	Learning Outcomes	Rank		
Knowledge	55	4	60	4	68	3	11	121
Comprehension	142	1	118	1	151	1	3	9
Application	69	2	70	2	74	2	6	36
Analysis	20	6	18	6	13	6	18	324
Synthesis	58	3	74	3	36	4	10	100
Evaluation	21	5	25	5	23	5	15	225
Total	365	21	365	21	365	21	63	815

W= 0.9746, S=153.5 X²=14.6

Table 8 above illustrates the rank of the sub categories of cognitive learning outcomes by three raters (judges). The table shows no tied ranks down each columns of ranks .The Kendall's coefficient of concordance (W) =0.9746 was tested using the table giving critical value of Kendall's S .

The calculated value of S=153.5 was compared with the critical value of S and the significance of coefficient of concordance was tested. Accordingly, when the critical value of Kendall's S associated with W at 0.05 level, it was found to be significant and the judges have about perfect agreement. Therefore, the classification of the syllabus cognitive domain in average has been shown in a table as follows.

Table 9 : Division of syllabus Cognitive Domain categories averaged by judges

Sub categories of Cognitive Domain	Average total	(%)
Knowledge	61	16.7
Comprehension	137	37.5
Application	71	19.5
Analysis	17	4.7
Synthesis	56	15.3
Evaluation	23	6.3
Total	365	100

Table 9 reveals 37.5% of the cognitive domain constitutes comprehension, and this is followed by application which comprises 19.5% .From the lower order thinking skills, Knowledge which weighs 16.7 % is given more emphasis than the higher order thinking skills such as evaluation and analysis which comprise 6.3% and 4.7% .

It is inferred that the attention given to the lower order thinking skill is better than that of the higher order thinking, which verifies the weakness of the syllabus, Furthermore , the emphasis given to the higher order thinking is neither proportional(equal) nor hierarchical. That is, in their order of thinking. This violates the recommendation forwarded by Zohar and Dori (2003) that the emphasis to HOTS (Higher Order Thinking Skills) should exceed that of the LOTS(Lower Order Thinking Skills) for high achieving students and tenth and above grade levels.

Table 10: Categorization of Test Items in to Cognitive Outcomes and their Rank orders

Sub-categories of cognitive domain	Judge 1		Judge 2		Judge3		Rank Sum(SR)	(SR) ²
	Number of items	Rank	Number of items	Rank	Number of items	Rank		
Knowledge	15	5	11	3	7	5	13	169
Comprehension	115	2	158	2	138	1	5	25
Application	140	1	178	1	126	2	4	16
Analysis	40	4	7	4	28	4	12	144
Synthesis	47	3	5	5	59	3	11	121
Evaluation	3	6	1	6	2	6	18	324
Total	360	21		21		21	63	799

S=137.5

W=0.87

$X^2=13.09$

p<0.05

Table 10 shows the categorization of test items in to cognitive sub categories by three judges. As can be seen from the table, the rank sum of the outcomes counts 63 while their square sum is 799.

Kendall's coefficient of concordance was computed so as to check the agreement of the judges. To do a test of the null hypothesis of disagreement, the calculated chi - squared value($X^2=13.09$) was tested comparing the critical value of X^2 at 5 degree of freedom and p=0.05.

The significance of the coefficient was tested using the critical value of Kendall's $S_{0.05}$ for $k=3$ and $n=6$, and the result was found to be statistically significant agreement.

The categorization reveals that the majority of the test items are grouped under application and comprehension skills of cognitive domain. They cover 79% of the total number of items (360). The remaining skills such as Synthesis (10.3%), Analysis (6.9%) and knowledge (3%) were given almost fair distribution among the remaining subcategories of cognitive domains, whereas Evaluation (0.3%) tends to be neglected.

Table 11: The Categorization of Test Items in to Cognitive Sub Domains Averaged Across Judges

Cognitive sub-domain	Average Total	%
Knowledge	11	3
Comprehension	137	38.1
Application	148	41.1
Analysis	25	6.9
Synthesis	37	10.3
Evaluation	2	0.6
Total	360	100

Table 11 illustrates that Application skill is given 41.1% of chance to be tested by the tests items. This is followed by comprehension, which is given 38.1% weight. The maximum emphasis, about (82%) of the test items measures the lower order thinking. As compared to data in table 9, Application is represented more than twice, whereas comprehension exceeds 1% and generally the lower order thinking activities represent 10 % greater than their representation in Table 9, which is illogical in the construction of Standardized Achievement Tests.

On the contrary, evaluation weighs only 0.6%, which shows insignificant attention given in the construction of test items. The distribution of synthesis (10.3) and analysis (6.9), however, seems fair among the remaining sub categories of cognitive domain. Therefore, it is possible to say that

the exams were highly dominated by the lower order thinking skills, especially by comprehension and application with the expense of others.

Below is a table which shows the number of expected test items in EUE English Examinations of 2011/12-2013/4 based on the observed number of learning outcomes in the syllabus of grades eleven and twelve so that if the items of the exam match with the learning outcomes will be checked .As a result the basic research question will be answered.

Table 12: The Number of Expected Test Items from the syllabus Cognitive Domain and the Chi – square values

Sub categories of cognitive domain	Syllabus Outcomes Averaged Across Judges (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Test Items Averaged Across Judges (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Total
Knowledge	61 (36.2)	24.8	17	11 (35.8)	-24.8	17	72
Comprehension	137 (137.9)	-0.9	0.006	137 (136.1)	0.9	0.006	274
Application	71 (110.3)	- 39.3	14	148 (108.7)	39.3	14	219
Analysis	17 (21.1)	-4.1	0.8	25 (20.9)	4.1	0.8	42
Synthesis	56 (46.8)	9.2	1.8	37 (46.2)	-9.2	1.8	93
Evaluation	23 (12.6)	10.4	8.6	2 (12.4)	-10.4	8.7	25
Total	365		42.1	360		42.1	725

$$X^2 = \sum \frac{(O-E)^2}{E} = 2(42.1) = 84.2$$

$$V = 0.116$$

$$p = 0.05$$

E

Table 12 indicates that the calculated chi-square value is 84.2. When it is compared with the critical X^2 value, $X^2_C = 11.07$ at $\alpha = 0.05$ significance level with $(6-1)(2-1) = 5$ degree of freedom from the contingency table, it exceeds about seven times.

The data supports the belief that a significant difference exists between the syllabus outcomes and observed test items. Hence the null hypothesis of this study is rejected due to the fact that the variables have significant relationship to each other.

The statistical or calculated result of Pearson's Chi squared test of independence confirms the bad- fit between the dependent (observed) and independent (expected) variables. It is seven times higher than the table or critical value. The result of Cramer's contingency coefficient value-V is 0.116. This indicates the presence of weak association and much stronger statistical relationship or dependence between the variables.

Table 13: The classification of syllabus cognitive domain Categories averaged by Judges

Sub categories	Averaged Total	(%)
Knowledge	61	16.71
Comprehension	137	37.53
Application	71	19.5
Analysis	17	4.7
Synthesis	56	15.34
Evaluation	23	6.3
Total	365	100

Table 13 depicts the classification of cognitive domain categories by judges. As can be seen, from 365 averaged total of objectives of Cognitive Domain, 37.53%(137) were classified under comprehension, while 19.5 %(71) ,16.7%(61) and 15.34%(56) were grouped under application ,knowledge, and synthesis respectively.

The remaining 6.3 % (23) and 4.7% (17) of the cognitive domain objectives were classified under evaluation and analysis respectively. It can be concluded the emphasis given to the lower thinking skill (knowledge) is more than twice greater than that of the higher order thinking skills such as analysis and evaluation.

As a whole, however, the distribution shows that the higher order thinking activities are emphasized, which proves the unbalanced attention given across the sub categories of cognitive domain.

Table 14 : Classification of the Expected and observed test items in to Syllabus Cognitive Domain

Learning outcomes	Expected	%	Observed	%
Knowledge	(35.8)	9.94	11	3.1
Comprehension	(136.1)	37.8	137	38.1
Application	(108.7)	30.19	148	41.1
Analysis	(20.9)	5.8	25	6.9
Synthesis	(46.2)	12.83	37	10.3
Evaluation	(12.4)	3.4	2	0.5
Total	360	100	360	100

Table 14 above reveals 41% of the tests items is represented by application of cognitive domain, which is followed by comprehension which weighs 38.1% . The representation of these two learning outcomes seems fair as compared with that of the syllabus (expected number). Knowledge, synthesis and evaluation which encompassed 9.94%, 12.8% and 3.4% in the syllabus receive only 3.1%, 10.3% and 0.5% respectively. The magnitude of the representation of these three cognitive subcategories is reduced in the sample test items. Evaluation is insignificantly represented despite its coverage in the syllabus.

Generally, such skills as knowledge, comprehension and application, i.e., the lower order thinking activities according to Bloom's taxonomy, represent the greater share (82.3%) in comparison with the HOTS (Higher Order Thinking) activities such as analysis, synthesis and evaluation.

This much unfair representation, overuse of the lower order thinking with the expense of the higher order thinking activities which involve the learning of judgmental skills such as critical thinking and problem solving impact on the students' ability to do something with facts ,understand them ,infer from them ,connect them to other facts and concepts ,categorize them, manipulate them ,put them in new or novel ways, and apply them as we seek new solution to new problems.

To conclude, the evidence ,as shown in table 14 ,the values of Pearson chi-square and Cramer's V ,and the representation of the expected and observed test items in table 12 reveals that there is disparity in treating the learning outcomes of cognitive domain in the sample test items in comparison with their treatment in the syllabus . The items do not represent proportionally from all cognitive domain of syllabus. The interpretation to be made from this analysis is that the test items of the EUE English Examinations are not valid in terms of the cognitive domains representativeness ,which violates the claim of Hughes(1989)<< To say a test has content validity, it should constitute a representative sample .>> (26-27)

4.2 Interview Results

In order to enrich the findings obtained from document analysis and find out the extent to which the items of EUE English Examinations are subjectively perceived to reflect the stated purpose (i.e. to determine the face validity), different stakeholders of EUE English Examinations were interviewed. Among the respondents, there were two teachers of AAU as exam experts and five English language teachers of Butajira Preparatory school, which is found in SNNPRG, Gurage Zone, Meskanworeda.

Hence the interviewees were interviewed similar questions, except few accordingly (on the basis of their responsibilities and experiences). They were forwarded questions in line with such points as approach of the EUE English Examinations, content, proportionality, time allotment, appropriateness (familiarity of item format (stem) adequacy of context, difficulty level), appearance of the test papers, validity and the wash back effect (the influence of EUE English Examinations on the teaching and learning process)

Accordingly, it is revealed by the responses of all interviewees that the EUE English examinations seem to be presented almost in communicative approach. It is different from the former examinations in a way that it attempts to present the language skills except listening, and grammar more in a communicative approach which shows the improvement of SATs of English. They said that the current EUE English Examinations encompass more items than the former so that the

exams have got more opportunity to include sample representatives than the former ones.

Regarding the content of the exams, all of the respondents confirmed that the listening aspect is neglected. Besides the vast majority of items of such skills as writing and speaking are from cognitive domain with the expense of affective and psychomotor domain. As a result, they said, the learners are provided little opportunity to be tested their motor skills.

Similarly, it is not common to see variety of literatures like poems in the history of EUE English Examinations despite their presence in the syllabus for reading comprehension items.

As to the proportionality of the test contents, the teacher respondents explained that it was not fair as compared with the variety of content areas and the magnitude of their representation in the syllabus. Due to the fact that the size and content of the syllabus or text books of both grades :11 and 12 from where the exam items are drawn are too large as a result the contents are too many even to be covered during the teaching and learning period in the academic years. However, the number of items in the sample booklets is only 120, which is hardly fair number to represent. The items in the vocabulary section, especially in the substitution part, are emphasized more than they are in the syllabus .On the contrary, grammar is given much less attention than vocabulary unlike it is in the syllabus, and the experts put their experience on how they determine number of items and said that they decide based on the degree of recurrence of macro skill, minimum and maximum learning competencies and mainly on difficulty, practicality and time to be given to administer the exams.

Regarding the allotment of time to the Examinations, Educators forwarded mostly negative comments. They explained that the exam takers usually complain about the time allotment. Texts for Reading comprehension, for example, were very lengthy and it was quite a big section. Most of the students either got bored after the second page or do part of it otherwise skip it as a whole running after time. The reason behind the problem is that there is no time given to read the passages, but only 120 minutes for 120 questions, which seems one minute to one question, which is different from their practice during classroom language examinations in which extra time is given for reading. However, the experts explained that all items are not of equal difficulty level as complexity level is one of the criteria to select items. From this it is inferred that the exams are not doable in the allotted time due to lack of students' experience (training) on how to use time.

As to the appropriateness of items, the educators provided comments mentioning that the context given to guess the meaning of words or vocabularies is not adequate enough despite the instructions. One repeated comment was that, writing subsections were unable to hold the students' attention because they were repetitive or had bland topics, for instance regarding types of writing. Some representative comments included: Some of the parts seemed to be a lot harder than the students' abilities—even for students who normally get really good grades due to the fact that the content areas are given little attention in somewhat different approach or format in the syllabus (textbook) which resulted in lack of opportunity for students to practice or exercise the items format. The activities that students are given on writing in the class room are usually more to write than to choose or analyze extracts .The students are given more chance to exercise writing so that they develop their psychomotor domain. Regardless of this, the items in the sample exams are from cognitive domain. The experts, on the other hand ,said that there was hardly any strong evaluation has been made with regard to the effectiveness of items particularly on EUE English Examinations and added that due to time, financial and material constraints ,it is unlikely to construct items which measure the students affective and psychomotor domains. This leads a conclusion that little attention is given to quality of testing, and the response of educators (teachers) implies that there is disparity between the learning outcomes of the syllabus and the sample representatives in the test items. If, on the other hand, a communicative approach to language learning has been adopted, the test specification should be based on the types of language tasks included in the learning program so that appropriate items are constructed.

With regard to the way the test experts ensure the content Validity, they said that TOS (Table of specification) is used so that the syllabus contents will be specified and weight of the selected contents will be determined on the bases of their recurrence in the syllabus. However, literatures suggest deciding the weight of contents based on the number of periods given during the teaching and learning, and the Table of Specification should be used not only to contents but also to learning outcomes.

Otherwise, there will be disparity between the syllabus contents and learning outcomes. What is more, the newly introduced time table, which is called English Satellite Television Programmers Transmission Timetable, which is being used currently in the Preparatory schools, should have been used to determine the actual periods and selected syllabus contents. It is advised that test developers should construct items based on commonly used materials.

The teachers highly criticized the general (physical) appearance of the test papers for the items are not sequenced uniformly in the booklets. On the other hand, the test experts confirmed that the items are written in four different booklets in order to minimize the students cheating rate. The order of items in one booklet is reshuffled in the other. However, this results in inconsistent layout, and this violates the principle of item sequence-simple to complex. As a result, the face validity is affected.

The response of the teachers on the wash back effect of the EUE English Examinations shows that the students are negatively influenced in a way that they don't give attention to the syllabus contents which do not frequently appear on the examinations and they are less interested in subjective type of assessments and higher order thinking activities. They prefer cognitive activities to psychomotor and affective

Regarding the challenges affecting Content validity, the test experts forwarded that material, financial, time and qualified manpower in area are the main ones.

To sum up, all the interviewees agreed up on the majority of the points. The teachers' response generally reveals that the EUE English Examinations have got limitation of measuring the students' higher order thinking skill, layout and fair representation of syllabus contents, especially syllabus skills and learning outcomes by the test items, and so do the response of test experts. For instance, the following examples were illustrative comments:

Example 1: *It is not an easy task to construct a standardized test You know...there are, of course, many things to be considered during item construction, but in our case, since the exam is Standardized, priority is given to the major content areas...I mean, skills and*

language elements ...mainly at macro level and their frequency .I mean their repetition in the language syllabus. At the same time, consideration of the learning competencies-the minimum and maximum is the other important thing The standardized test to be constructed is specific for high school final examination; hence it must measure the pre-stated goals in the curriculum. Thus, the construction of the test must follow the pre-determined goals of competence. Another thing, which is not less important than these, is the issue of practicality. Practicality in a sense... that covers financial limitations, time constraints, ease of administration and scoring and interpretation.

Example 2:*It is ideal to encompass all language skills and language elements since the limitations are there .Briefly speaking ,students should be either observed or recorded and evaluated while they are performing ,but to do this high technology and a great deal of money are needed. Otherwise, experts tend to be limited to specific learning domain.*

Example 3:*Subjectivity of scoring and personal interpretations are another challenges. Therefore, by hook or by crook you will be forced to prepare objective questions .I know there is still another headache_ convenience for cheating, so to overcome this problem, some mechanisms or techniques should be designed-but..at this time ...nowadays the tentative solution is to shuffle items in booklets.*

As can be inferred from the implications of the above examples, the findings are congruent to that of document analysis. In other words, Tables 4, 5,7,9,11,12 and 14 depict what the responses of the interview confirm.

There have been few studies assessing the content validity of language tests of different grade levels. Their findings too are wholly or partly compatible with the present findings. They indicate that the exams lack content validity .For instance, Alemu (1983:3), (Tesfaye, 1982 and Dejene, 1990, as cited in Nigussie, 2002:35) conducted a study on assessment of grades six and eight English National Examinations and they found out that the exams had defects, particularly in contents and formats, and the items were dominated by grammar and vocabulary questions. Furthermore, despite the different formats employed in the textbooks the examinations are purely objective type. As compared with the findings of the present study, they are compatible.

Besides, Nigussie (2002) carried out similar study on grade ten EGSECE English Examination of 1993 .His findings are consistent with that of this study in line with all the above issues. Furthermore, he found that HOTS (Higher Order Thinking Skills) were ignored, and so were in the EUE English Examinations of 2011/12-2013/14.However the researcher found that the test items of EGSECE of 1993 were drawn proportionally from all cognitive domain categories.

Dawit (2000) analyzed the content validity of EGSECE of 1997 and 1998 in terms of cognitive taxonomy. The study sought to answer: How are the levels of Bloom's Taxonomy represented in the two examinations? The result showed that lower order cognitive skills were most frequent in the two examinations, and the results revealed that in the two examinations analyzed, memorization-the lowest cell in Bloom's Revised Taxonomy, was the most frequent code followed by application. Understanding meta cognitive knowledge were absent in the examinations. These are partly congruent findings with that of the current study.

CHAPTER FIVE

5. SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 SUMMARY OF MAJOR FINDINGS

The objective of the current study was to assess the content validity of EUE English Examinations of 2011/12-2013/14. To attain this, mixed method approach was used. Document analysis and semi structured interview were conducted. Inferential statistics such as Chi-square test of independence along with Cramer's coefficient of contingency and Kendall's W were computed and interpreted during the document analysis and the goodness- of- fit was tested. Then to check out the consistency of the findings generated by document analysis ,results of interview were triangulated.

Accordingly, the findings revealed that the EUE English Examinations of 2011/12-2013/24 were not content valid; neither the syllabus contents nor the learning outcomes of the cognitive domain were fairly represented, and exams tended to have poor face validity due to the fact that they were perceived by the stakeholders as they fail to measure the intended objectives effectively.

Briefly speaking, there was a significant difference between the observed and the expected contents of the test items .Hence, the items of EUE English Examinations of 2011/2-2013/4 did not match with the syllabus content.

There was weak association and much stronger statistical relationship or independence between the variables. This means that the items of the examinations did not adequately represent the syllabus outcomes. There was a significant difference between the syllabus outcomes and observed test items.

Regarding the appropriateness of the Examinations, the items looked in appropriate in a way that the formats of the tests items were not consistent with that of the syllabus, the context given

to guess the meaning of words or vocabularies was not adequate. Some of the parts seemed to be a lot harder than the students' abilities.

The EUE English Examinations were perceived by the stake holders as not good.

The vast majority of the test items were drawn from such content areas as vocabulary, grammar and writing respectively.

The items representing cognitive learning outcomes were drawn from such skills(sub categories) as knowledge, comprehension and application.

The lower order thinking activities according to Bloom's taxonomy represented the greater share in comparison with the HOTs (Higher Order Thinking) activities such as analysis, synthesis and evaluation.

5.2 CONCLUSIONS

As Popham (1981) has described, “The content validity of achievement tests refers to representativeness and adequacy with which a test taps the topics it should.” Therefore, it is not that difficult to teach a unit or course to your student, but it is important to know how successful that teaching is. The success of that course or unit can be seen through the right assessment of the efforts exerted to teach the course. Good teaching consequently yields good testing. The good testing is supposed to yield a positive backwash on both the learner and the teacher simultaneously.

In examining the EUE English Examinations (EUEEE) the researcher came to find that those examinations lacked the characteristics of a good English Language test (Standardized Achievement Test). They lacked comprehensiveness, content validity. They were also perceived by the stakeholders as the items did not measure what they were purported to, and did not possess any positive backwash on the learners and teachers. This incomprehensiveness and lack of content validity yielded very poor educational practices in terms of foreign language teaching in our country. From this, conclusions are deduced as follows:

- Teachers focused on teaching the examination techniques to help students collect as much marks as possible rather than doing real teaching to change the students' behavior towards language learning (i.e. there was the wrong perception of teachers or test constructors towards the relationship between testing and teaching.)
- Students did not refer to their textbooks but focused their efforts in learning how to do the exam rather than learning the language to use in real life situations.
- They didn't practice and develop their productive skill. As a result, the majority of students had lack of confidence in speaking and writing both in the classroom and outside classroom, which is against the intended learning outcome of the course.

- The students lack confidence due to the tendency of appearance of difficult item first in the booklet .Consequently, their achievement was not so good as expected.
- The educational aims were not secured by the teaching of the syllabus imbedded in the textbooks. The result of this practices yielded poor attainment of national and personal pedagogical and educational goals

5.3 RECOMMENDATIONS

Based on the conclusions derived above, the following recommendations have been forwarded:

- Test constructors should note that the test items must function as intended. Improper functioning of test items could sometimes occur if they contain inappropriate vocabulary, clues to answer, unclear directions or some other defects.
- Test constructors should use not only the syllabus but also the time table of the satellite television program so as to identify the syllabus contents intended to be taught in the academic years and period allotted for them so that the syllabus contents will be fairly represented by the test items.
- The test constructors should use the Students Text Books along with the syllabus so that how activities on the intended syllabus contents are presented will be recognized which help the test constructors prepare items in a way that the students have practiced, which ensures the validity of test items since it is when the learners are tested what they were taught in a way they were taught that valid data is obtained.
- Either the test constructors should increase the number of items thereby increasing the representation of contents and learning outcomes from the different parts of the material or the curriculum developers reduce the number of syllabus contents so that the contents will get sufficient time to be taught during the teaching and learning period and adequate number of sample representative during test construction. In relation with this point, Heaton(1988:12) said, <<The longer the test, the more reliable a measuring instrument it'll be.>>
- The test constructors should use TOS(Table Of Specification) not only to determine the syllabus contents but also to the learning outcomes since the procedure for demonstrating content validity in achievement test is essentially a matter of analyzing and identifying the two compounds such as content and domain of outcomes.

- In order to assess such difficult skills as affective and psychomotor, continuous assessment should be introduced in a formal way and should have value to be added on the result of EUE English Examinations, and to realize this government should allocate enough budget to MOE generally and equip secondary and preparatory schools with language learning materials particularly
- Experts of the EUE English Examinations should present variety of literatures such as poems and novels for reading comprehension and give greater attention for items sequence-simple to complex.
- The current study has got limitations of providing information about other validities and level of item difficulty of the EUE English Examinations of 2011/2-2013/4 thoroughly. Therefore, other researchers are suggested to conduct a study on the aforementioned issues so that it might help the test experts as input if need arises.

References

- Alderson, J.C. (1981) "Report of Discussion on Communicative Language Testing", The British Council (ELT) Documents.
- Alderson, J.C. et al. (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.
- Alemu Tsegay (1983) *Assessment of Grades Six And Eight English National Examinations*, M.A. Thesis. Addis Ababa: Addis Ababa University. (Unpublished)
- Aggarwal, Y.A (1998) *Statistical Methods Concepts, Application and Computation*, Third Revised and Enlarged Ed. Sterling Pub.Priv.Limited New Delhi-India
- Anastasi, A. (1975 & 1990). *Psychological Testing*. New York, Macmillan publishing company.
- Bachman, L.F. (1990). *Fundamental Consideration in Language Testing*, Oxford, Oxford University Press
- Baker, D. (1989) *Language Testing, A Critical Survey and Practical Guide*, Edward Arnold
- Canale, M and Swain, M (1980) Theoretical Cases of Communicative Approach to Second Language Teaching and Testing, In *Applied Linguistics*, Vol. 1. No. 1. USA.
- Finocchiaro, M and Sako, S (1983) *Foreign Language Testing: A practical Approach*, New York: Regents pub. Comp, Inc.

Girma Lemma(1997)Representativeness and coverage adequacy of physics and Mathematics
ESLCE

items,”Institute of Educational Journal”-V.IV1997.Addis Ababa University
.Retrieved
on 1,August2014 from:<http://www.inst,jo.org.ev/>

Gronlund, Norman. E (1990) Measurement and Evaluation in Teaching, 6th Ed Macmillan Published
Company. New York, London

Harrison(1983).A language Testing Handbook,London:Macmilan Published,Ltd.

Heaton, J.B. (1975) Writing English Language Tests, London: Longman .

Heaton J.B (1988). Writing English Language Tests, London: Longman.

Hughes, A (1989). Testing for Language Teachers. Cambridge University Press, UK.

Maxwell(1996).Qualitative Research Design:An international Approach(2nd edition).__:SAGE
publications

Meherens and Lehmann (1991) Measurement and Evaluation in Education and Psychology,
4th Ed, London: Harcourt Brace College Publishers.

MekonnenYibrah et.al (2014)Assessment of Content Validity in Ethiopian General Secondary
*Education Certificate of English Examinations (2009-2012)Global Scholars
Journal of Linguistics and Communication Studies* Vol.1 (1), (p.3).Retrieved
on 28, July 2014 from<http://www.globalscholarsjournals.org/SS/GSJLCS/index.php>

NigussieTamirat (2002)*The Content validity of the Ethiopian General Secondary Education
Certificate English Examination* (MA Thesis) Retrieved on 4, August 2014 from
<http://www.loc.st.org.E>

Nitko, J. Anthony (1996) Educational Assessment of Students Ohio, A Simon and Schuster Company Englewood Cliffs Hall, Inc.

Popham, W.J. (1981) Modern Educational Measurement, Prentice Hall, Inc.

Siddiek, Ahmed Gumaa. (2009). *Assessment of Sudan School Certificate English Examination*. Published (Doctoral thesis). Retrieved on 23, July 2014 from [http:// www.Lulu.org](http://www.Lulu.org) 2009. UK.

Siddiek, Ahmed Gumaa. (2010). *The Impact of Test Content Validity on Language Teaching and Learning* Journal Article Vol. 6, No. 12; December 20 Dawadami Community College, Shaqra University Retrieved on 3, July 2014 from <http://www.ccsenet.org/ass>

Underhill (1991). Testing Spoken Language. Cambridge: Cambridge University press.

Waleign (2006). Educational Measurement (Epsy312). Department of Pedagogical Science, (unpublished Handout), Haramaya University.

Weir C.J. (1990) Communicative Language Testing, London: Prentice Hall International

Weir, G. J. (1993). *Communicative Language Testing*. Prentice Hall, New York. Retrieved on 3, July 2014 from <http://www.tsukuba-g.ac.jp/.../5.LANGHAM.pdf>

UCAS (2013) *International qualifications for entry to university or college in 2014 Journal Article*. vol.1(p.21) Retrieved on 24, July from <http://www.ucas.com>.

(_____) <Item and test analysis by language groups > Retrieved 21, July from.. [http://www.nae.gov.et/download/ Grade10.pdf](http://www.nae.gov.et/download/Grade10.pdf)

Zohar and Dori (2003). *Higher Order Thinking Skills and Lower Achieving Students: Are They Mutually Exclusive ?* The Journal of The Learning Science, 12(2), 145-181. Lawrence Erlbaum Associates, Inc. Retrieved on 5, July from <http://www.tsukuba-g.ac.jp/...>

Appendix

Domain Categories

Put a thick mark (/) against the cognitive domain category below

No	Instructional objectives	Knowl edge	Compreh ension	Applic ation	Analy sis	Synthe sis	Evalu ation
1	Write a report						
2	Make notes in a table						
3	Scan factual information						
4	Listen to an extended lesson						
5	Identify stressed syllable in a word						
6	Categorize nationality words						
7	Read an anthem and answer comprehension question						
8	Change active sentence in to passive form						
9	Discuss advantages and disadvantage using concessive marks						
10	Debate a contentious issue						
11	Run meeting using language of meeting						
12	Read story						
13	Listen to tests to take notes in note summary form						
14	Evaluate themselves what kind of language learners they are						
15	Write an informal letter using an informal letter writing conventions						
16	Identify the links of words in spoken English						
17	Report sentences using indefinite reported speech						
18	Give a solution to a problem reading a puzzle type text						
19	Write a magazine article including several main points						
20	Give reason to learn English, rank (value their reason						
21	Write a sentence in first condition						
22	Complete sentence with either <if> or<unless>						
23	Build words using suffixes <-tion><-or>,<-er>						
24	Identify relative pronoun of defining reported speech and state their function						
25	Take notes and ensure questions listening						
26	Define words in listening script						
27	Read a text on medical innovations and write notes						
28	Write medical words						
29	Read a poem and analyze it						
30	Construct a sentence about past using 'would' and 'used to'						

31	Pronounce ' ough' in different contexts accurately						
----	--	--	--	--	--	--	--

32	Read a table summary and answer questions						
33	Read contractions in complete form in a sentence						
34	Read a health leaf let						
35	Write sentence in third condition						
36	Discuss pronunciation problems						
37	Give solution for pronunciation problem of others						
38	Define the meaning of make and 'let'						
39	Express performance and reservations						
40	Answer a quiz before listening to a script						
41	Write a sentence using < incase>						
42	Give advice using a large of structures						
43	Role play as a counselor and advisee						
44	Match phrases (sentence) with their rhythm						
45	Perform rhythm chain story						
46	Perform rhythm chain story						
47	Write a letter of reply						
48	Define acronyms and initials						
49	Describe changes or trends of data						
50	Talk about parts of speech						
51	Give evidence to support opinion						
52	Interpret simple statistics and write a report						
53	Get interested in using dictionary						
54	Brain storm words to add to mind map						
55	Listen to dialogues to identify gist						
56	Group nouns and verbs which have related meaning						
57	Building bonus from base verbs adding 'ion'						
58	Pronounce nouns ending in 'ion' correctly						
59	Write a guided paragraph						
60	Speak using appropriate register and language in a situation						
61	Develop convincing argument and persuasive language						
62	Give meanings of new words using contextual clues						
63	Give meaning of new words using knowledge of word formation						
64	Scan a text to obtain information						
65	Read a text to indentify main point						
66	Read a text and explain the feature of the text						
67	Fill gaps in a text						
68	Distinguish factual adjectives from the subjective opinion adjective opinion adjectives in a text						
69	Describe places and activities persuasively						

70	Write a short formal letter using current conventions						
71	Drive a rule for ‘-ed’ and ‘-ing’ ending adjectives						
72	Examine their own writing critically						
73	Set a realistic writing goal						
74	Read an excerpt from a novel and answer multiple choice questions						
75	Listen to a story to sequence events						
76	Define narrative tense						
77	Analyze story plot						
78	Create a story orally						
79	Match extreme adjectives with their base form						
80	Change one form of question to another form						
81	Match phrasal verbs with their meanings						
82	Differentiate genres of texts						
83	Sequence events from a novel						
84	Write a short narrative using						
85	Complete sentence with correct narrative verbs						
86	Group sequence indicators based on their function						
87	Write a summary of story joining lists of events						
88	Listen to text and make notes in table						
89	Complete definition of terms						
90	Take about climate change						
91	Group discourses markers according to their function						
92	Write a summary						
93	Write an information leaflet						
94	Link pair of sentence with appropriate markers						
95	Complete sentence with suitable cause result verb						
96	Express wishes and hops using variety of language structure						
97	Survey a text to a usually questions						
98	Complete mind map						
99	Brainstorm and exchange information						
100	Discuss an issue						
101	Make a presentation						
102	Identify rules of using ‘as ‘ and ‘like ‘						
103	Complete sentences with either adjective and adverbs accordingly						
104	Change adjective to adverb form and vice versa						
105	Assess themselves						
106	Express their wishes using ‘wish’						
107	Take part in role play to inform						
108	Writ two to three paragraph to inform						
109	Assess own ability in terms of speaking and writing						

110	Skim a text to find out general information						
111	Complete sentence with a suitable verb form						
112	Use 'can' and be able' to express ability						
113	Develop conversation about unit topic						
114	Match wh- words with the questions						
115	Comment on report						
116	Match meaning to prefixes						
117	Rewrite sentence replacing verbs						
118	Write a survey report						
119	Match meaning to words in a poem						
120	Set a listening goal						
121	Highlight silent consonants in a word.						
122	Assess own formulated strategies on leading						
123	Write sentence using synonyms and antonyms of 'rich' 'poor'						
124	Listen to an interview						
125	Complete rules of past simple & present perfect						
126	Express agreement and disagreement						
127	Read job advertisement						
128	Make plan and write a brochure						
129	Differentiate the grammar of mother tongue and that of English						
130	Assess own progress based on pre set goals						
131	Decide and rank inventions						
132	Do pre listening activities						
133	Predict content of a text						
134	Set a grammar goal for a week						
135	Build a word from its family						
136	Write an opinion essay						
137	Make notes on personal views						
138	Write speculation about the future with range of expressions						
139	Listen to a text to do multiple choice questions						
140	Develop a family discussion about a problem						
141	Compare (facts aspects of life) in a text with own life						
142	Put words in the correct order to make sentence						
143	Define rules for using prefix						
144	I identify main point in speech and write it note form						
145	Discuss the advantage and disadvantages of methods of communication						
146	Read a poem to identify the subject						
147	Recite a poem						
148	Rate how well language skill can be done						
149	Match film scenes with their definitions						
150	Compare notes						

151	Deal with problems using range of language to deal with & solve problems						
152	Match jobs, with their definition						
153	Write an essay based on thought bubbles						
154	Write an essay based on thought bubbles						
155	Match picture with its description						
156	Punctuate a short passage						
157	Complete cross words						
158	Compare and contrast poem and novel						
159	Explain idioms						
160	Play a speaking game using language of changing the subject and coming back						
161	Listen to film scripts and write brief notes (out line)						
162	Listen to a review of film and make note reviewers opinion						
163	Write comments on film reviews						
164	Connect a verb with the particle 'off' to get phrase						
165	Ask and answer questions						
166	Match the beginnings and endings of sentence using the correct form of 'will'						
167	Express purpose using variety of words expansions						
168	Write sentences with suitable word ending-ship						
169	Write a learn let following instruction						
170	Listen to description and write brief note						
171	Describe qualities and skills needed for a range of jobs						
172	Group words based on their suffixes						
173	Criticize activities						
174	Make suggestion						
175	Read to gain information						
176	Write CV						
177	Listen to do comprehension questions						
178	Construct reduced relative clause correctly						
179	Assemble sentence from different parts of a sentence						
180	Write a profile of company or an organization						
181	Complete sentence with the form of past tense						
182	Write an essay following the seven stages						
183	Write letter of application along with CV						
184	Construct a job interview in to present , past and present perfect						
185	Read questions with various intonation						
186	Take note from lecture						
187	Discuss issues using a range of structure to highly information						
188	Re write sentence using modals						

189	Critically evaluate self judgment						
190	Pronounce weak vowels in un stressed syllables						
191	Group adjectives based on common proposition						
192	Make notes on a range of topics						
193	Role- play on a theme in which there is a conflict						
194	Write a short passage punctuated correctly						
195	Write words in word search puzzle						
196	Write an argumentative essay						
197	Find words in word search puzzle						
198	Identify highlighted information in sentences						
199	Transform wh-questions to highlighted form						
200	Present the press and com of						
201	Create a role – play						
202	Listen to broad casts and match stories to pictures						
203	Enjoy Ethiopian films						
204	Write a descriptive essay						
205	Identify different past verb forms written material						
206	Changes direct speech to indirect and vice versa						
207	Hold a debate						
208	Use language for classification appropriate						
209	Convince others using persuasive language appropriately						
210	Listen to others with respect						
211	Accept responsibility for one's learning						
212	Recite the anthem of Ethiopia						
213	Practice exam strategies						
214	Retell stories						
215	Organize a talk show						
216	Ask for suggestion						
217	Cooperate in group activities						
218	Make a plan and produce class magazine						
219	Make self evaluation						
220	Search for best strategies to improve writing						
221	Contribute to and develop conversation						
222	Express facts using simple present tense						
223	Propose a plan to one's language improvement						
224	Ask for someone's opinion						
225	Act out conversation and tell stories						
226	Read a conversation aloud						
227	Read part of story and predict the next						
228	Make comparison sentence correctly						
229	Write a paragraph using phrasal verbs						
230	Narrate stories from pictures						
231	Express contrast necessity and permission						

232	Play language games						
233	Detect non verbs communication was						
234	Take dictation						
235	Listen to others with respect						
236	Give replies to requests						
237	Express reason and result						
238	Evaluate the effectiveness of word association strategy to oneself						
239	Ask and offer help						
240	Arrange paragraphs to make a meaningful passage						
241	Express intentions						
242	Write a letter of complaint						
243	Use verb followed by infinitive and gerund in a sentence correctly						
244	Extend vocabulary						
245	Make active and passive sentences correctly						
246	Read silently and answers comprehension questions						
247	Make sentence using probable condition						
248	Participate in guessing games						
249	Write notes						
250	Take and make notes						
251	Analyze one's own approach to language leaning						
252	Take part in a role- play about a family situation						
253	Make sentence about leg lets						
254	Review and discuss different ways of recording vocabulary						
255	Distinguish kinds communication						
256	Make questions						
257	Extend knowledge of prefixes						
258	Carry out self assessment exercise and then write report on strength and weakness						
259	use fillers appropriately						
260	Assess one's own speaking and leading						
261	Change the subject in conversation						
262	Extend ability to use adjectives						
263	Create and carryout a class survey participation in the arts						
264	Take part in an extensive leading						
265	Write a film review						
266	Asserts one's own listening and reading						
267	Describe leaders using appropriate adjectives						
268	Talk about fid use events with range of strictures						
269	Express purpose using different language expression						
270	Identify words commonly used n development						
271	Use expression it is time appropriately						
272	Practice research skills						

273	Describe figures in percentage and express proportions						
274	Identify, commonly used words about manufacturing and trade						
275	Describe a process						
276	Take about globalization						
277	Write a business letter						
278	Talk about future career						
279	Use defining relative clause appropriately						
280	Describe personal qualities and skills						
281	Discuss the process of applying for a job						
282	Read texts used when applying for a job						
283	Express ability using modals						
284	Read news papers						
285	Express past intentions accurately						
286	Write a dialogue						
287	Change direct question to embedded form						
288	Role – play a TV talk show						
289	Role- play a review of a TV show						
290	Give short responses						
291	Read review of film						
292	Define phrasal verbs with ‘away’						
293	Read extracts from magazine articles						
294	Appreciates a magazine cover						
295	Ask and answer questions to find out about a person						
296	States names of Africa countries and nationalities						
297	Discuss solution to a riddle						
298	Make contrastive sentence stress accurately						
299	Decide to be an effective independent learner						
300	Give a one minute talk						
301	Identify names for different medical stage of and medical equipment						
302	Identity names of common illnesses						
303	Talk about inners						
304	Write description of graphs						
305	Identify words about tourism						
306	Write a narrative						
307	Identify homophones						
308	Talk part in three role – play						
309	Express hope and fears using appropriate structure						
310	Use ‘will’ and ‘going’ to appropriately to talk about the future						
311	Build words by adding affixes						
312	Relate what is lead with own experience						
313	Talk about a hypothetical present						
314	Listen and identify speaker’s feelings through use of intonation						

315	Use a range of structures to demand explanation						
316	Elaborate ideas / Justify opinions in essay by giving examples						
317	Summaries information and suggestions						
318	Agree, disagree and express simple counter arguments						
319	Summaries a text using bullet points						
320	Critically evaluate texts to inform own comments judgments						
321	Research , deliver initiate disc mission on a short probation a chosen topic						
322	Identify the gist of a short monologue						
323	Identify main points reading an autobiographical account						
324	Write and present a 500 word report on a given topic for a specific audience including recommendations						

N.B. The remaining objectives are redundancy!

Appendix-C

Reliability Test of the Research Tool

Sub-categories of cognitive domain	Judge 1		Judge 2		Judge 3		Rank sum (SR)	(SR) ²
	Learning Outcomes	Rank	Learning Outcomes	Rank	Learning Outcomes	Rank		
Knowledge	15	5	9	5	13	4	16	196
Comprehension	40	1	49	1	38	2	3	16
Application	38	2	40	2	44	1	6	25
Analysis	17	4	20	3	22	3	10	100
Synthesis	20	3	17	4	11	6	11	169
Evaluation	10	6	5	6	12	5	17	289
Total	140	21	140	21	140	21	63	795

*W=0.84

$$W = \frac{12S}{M^2(N^3-N)-M \odot T} \quad \text{OR} \quad \frac{12S}{M^2(N^3-N)}$$

$$M^2(N^3-N)-M \odot T$$

$$M^2(N^3-N)$$

Where

M=number of judges

N=number of objects(topics rank)

T=a correction factor for tied ranks

S=Sum of squared deviation of total ranks about the mean (Appendix-G)

XI

Appendix- D

Major topics and periods allotments of grades 11 and 12

Major topics	Period allotments		Total	%
	Grade			
	11	12		
Reading comprehension				
Vocabulary				
Grammar				
Communicative ativities				
Listening				
Writing				
Total				

Appendix-E

Chi -square Test Analysis of the Expected and Observed Test Items

Major topics	Period allotted (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Observed Test Items (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Total
Reading Comprehension							
Vocabulary							
Grammar							
Communicative activity							
Listening							
Writing							
Total							

$$X^2 = \sum \frac{(O-E)^2}{E} = 47.23 + 47.23 = 94.5$$

$$p=0.05 \quad V=0.378$$

O= the observed frequency count in population r for level c of the categorical variable.
(the observed number of items in the exam)

E=the expected frequency count in population r for level c of the categorical variable.
(Number of expected items based on the syllabus)

P-value=the probability of observing a sample statistic as extreme as the test statistic

V=the square root of the chi-squared statistic divided by the sample size and the minimum dimension minus 1

Appendix-F

The Number of Expected and Observed Test Items from syllabus content

Content Areas	Observed	%	Expected	%
Reading comprehension				
Vocabulary				
Grammar				
Communicative activity				
Listening				
Writing				
Total				

Appendix-G

The coefficient of concordance W and the Observed value of S on the categorization of test items in to the educational taxonomies.

Ranking judges	Domain Categories			Sum of Rank Total	Mean Rank Total	Sum of squared deviation of rank total about the means	The correction factor	Coefficient of concordance
	Cognitive	Affective	Psychomotor	$\sum R_j$	$\frac{\sum R_j}{N}$	$S = \sum (R_j - \frac{\sum R_j}{N})^2$	$T = \sum (\frac{t^3}{12} - t)$	W
Judge1								
Judge2								
Judge3								
Rank total(R)							-----	

$$W = \frac{12S}{M^2(N^3 - N) - M \sum T} \quad \text{OR} \quad \frac{12S}{M^2(N^3 - N)}$$

Where

M=number of judges

N=number of objects(topics rank)

T=a correction factor for tied ranks

S=Sum of squared deviation of total ranks about the mean

Appendix-H

Division of Syllabus Objectives in to Learning Domains

Judges	Learning Domains				Rank			Total
	Affective	Psycho Motor	Cognitive	Unclear	Affective	Psycho Motor	Cognitive	
Judge 1								
Judge 2								
Judge 3								
Total								

Appendix-I

Division of Grades 11 and 12 Syllabus Objectives in to Learning Domain Averaged by Judges

Division of Learning Domain	Average Total	%
Affective		
Psychomotor		
Cognitive		
Total		

Appendix-J

The coefficient of concordance W on the categorization of syllabus cognitive outcomes

Ranking judges	Domain Categories						Sum of Rank Total	Mean Rank Total	Sum of squared deviation of rank total about the means	The correction factor	Coefficient of concordance
	1	2	3	4	5	6					
Judge1							$\sum R_j$	$\frac{\sum R_j}{N}$	$S = \sum (R_j - \frac{\sum R_j}{N})^2$	$T = \sum \frac{(t^3 - t)}{12}$	W
Judge2											
Judge3											
Rank total(R)										-----	

Appendix-K

Division of syllabus Cognitive Domain in to sub categories and their Rank orders by judges

Sub-categories of cognitive domain	Judge 1		Judge 2		Judge 3		Rank sum	$(SR)^2$
	Learning Outcomes	Rank	Learning Outcomes	Rank	Learning Outcomes	Rank	(SR)	
Knowledge								
Comprehension								
Application								
Analysis								
Synthesis								
Evaluation								
Total								

Appendix--L

The Number of Expected Test Items from the syllabus Cognitive Domain and the Chi –square values

Sub categories of cognitive domain	Syllabus Outcomes Averaged Across Judges (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Test Items Averaged Across Judges (O) (E)	(O-E)	$\frac{(O-E)^2}{E}$	Total
Knowledge							
Comprehension							
Application							
Analysis							
Synthesis							
Evaluation							
Total							

$$X^2 = \sum \frac{(O-E)^2}{E} = 2(42.1) = 84.2$$

E

$$V = 0.116$$

$$p = 0.05$$

Appendix-M

Classification of the Expected and observed test items in to Syllabus Cognitive Domain

Learning outcomes	Expected	%	Observed	%
Knowledge				
Comprehension				
Application				
Analysis				
Synthesis				
Evaluation				
Total				

Appendix-N

Interview for Exam Experts

I.PERSONAL INFORMATION

1.1 Place of work: _____

1.2 Code of the interviewee: _____

1.3 Qualification: _____

1.4 Field of specialization: _____

1.5 Responsibility (position) at present: _____

II.BODY OF THE INTERVIEW

Topic 1: Approach

1. Would you please tell me the approach of the EUE English Examination?

- What makes it similar to, or different from the then, or former examinations?
- What changes have been made regarding the construction of the current EUE English examinations, if any, why?

Topic 2: Content

2. How do you decide the content and number of test items ?

- Which of the language skills and language areas are encompassed by the exams?
- Is there any language skill or language area not included in the examinations, which one(s)?

Topic 3: proportion

- what do you say about the sufficiency of the test items?
- Are the test items selected based on the taxonomy of educational domain in the syllabus?
- How much proportional are the cognitive domain categories in the exam?

Topic 4: validity

4. How do you ensure the content validity?

What kind of materials do you use so as to construct the examination?

➤ How do you use the student Textbook and the syllabus during the construction of the examinations?

Which one do you use more, why?

➤ To what extent do you use the TOS (Table Of Specification)?

➤ When do you use TOS-before or after, or before and after the test construction?

Topic 5: Layout (general appearance)

5. To what extent are the exams perceived to reflect the stated purpose? (=Do they appear to measure the knowledge or ability it claims to measure?)

➤ Do you think the exams look right or fair or valid when assessed by the stake holders?

➤ How are the items sequenced in the exam paper: easy to difficult? Here and there?

➤ To what extent do you think the items and directions are clear and un complicated?

➤ Tell me about how much doable the test within the allotted time limit.

➤ Tell me the extent to which the difficulty level of the exam presents a reasonable challenge, which is appropriate for students?

➤ Tell me how well the exam is constructed and has expected format with familiar tasks.

Topic 6: Challenges

6. What are the challenges affecting the content validity of EUE English Examination?

Appendix-O

Interview for Subject Teachers

I.PERSONAL INFORMATION

- 1.1 Place of work: _____
- 1.2 Code of the interviewee: _____
- 1.3 Qualification: _____
- 1.4 Field of specialization: _____
- 1.5 Responsibility (position) at present: _____

II.BODY OF THE INTERVIEW

Topic 1: Approach

1. Would you please tell me the approach of the EUE English Examination?
- What makes it similar to, or different from the then, or former examinations?
 - What changes have been made regarding the construction of the current EUE English examinations, if any, why?

Topic 2: Content

2. How do you decide the content and number of test items ?
- Which of the language skills and language areas are encompassed by the exams?
 - Is there any language skill or language area not included in the examinations, which one(s)?

Topic 3: proportion

- what do you say about the sufficiency of the test items?
- Are the test items selected based on the taxonomy of educational domain in the syllabus?
- How much proportional are the cognitive domain categories in the exam?

Topic 4: validity

4. How do you ensure the content validity?

What kind of materials do you use so as to construct the examination?

➤ How do you use the student Textbook and the syllabus during the construction of the examinations?

Which one do you use more, why?

➤ To what extent do you use the TOS (Table Of Specification)?

➤ When do you use TOS-before or after, or before and after the test construction?

Topic 5: Layout (general appearance)

5. To what extent are the exams perceived to reflect the stated purpose? (=Do they appear to measure the knowledge or ability it claims to measure?)

➤ Do you think the exams look right or fair or valid when assessed by the stake holders?

➤ How are the items sequenced in the exam paper: easy to difficult? Here and there?

➤ To what extent do you think the items and directions are clear and un complicated?

➤ Tell me about how much doable the test within the allotted time limit.

➤ Tell me the extent to which the difficulty level of the exam presents a reasonable challenge, which is appropriate for students?

➤ Tell me how well the exam is constructed and has expected format with familiar tasks.

Topic 6 :Wahback Effect

6 .How do you explain the influence of the EUE English Examinations on the teaching and learning process?

➤ What are the positive influences of the EUE English Examinations on language teaching and learning?

➤ What are the negative influences of the EUE English Examinations on language teaching and learning

Topic 7: Challenges

7. What are the challenges affecting the content validity of EUE English Examination?