International Journal of Innovations in TESOL and Applied Linguistics
Vol. 3, No. 1; 2017
ISSN 2454-6887
Published by ASLA, Amity University, Gurgaon, India
© 2017

Assessing Content Validity of the EGSEC English Examinations

Mekonnen Yibrah

Haramaya University, Ethiopia

Received: Apr. 13, 2017 Accepted: Apr. 23, 2017 Online Published: May. 27, 2017

Abstract

The purpose of this study was to assess whether the Ethiopian EGSEC English exams administered by the NEAEA fairly represent the content coverage and adequate sampling of the objectives stated in the syllabi. To attain this objective, quantitative and qualitative research design were employed. The required data for the study were collected using textbooks/syllabi and SATs content analysis, questionnaire and unstructured interview. To cross check the data obtained from the document analysis, questionnaire and unstructured interview were for 12 teachers and 3 testing experts. The subjects of this study were selected using purposive sampling technique. The content validity of the contents of the textbooks of grade 9 and 10; and EGSEC SATs were analysed using the standards of measuring content validity. The collected data were analysed 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 textbooks and sample SATs respectively. The findings of the study revealed that the degree of relationship and strength of association between the contents of the sample SATs and textbooks were found to be divergent and have weak strength of association. The result of the study generally shows that the exam papers had problems in adequately sampling all sections of the language items. Especially, in most cases, grammar and reading dominated the exams whereas listening skill was totally ignored in the testing system. This indicates a mismatch in allocating the test items with respect to the prescribed weight of contents in the textbooks. Besides, the less emphasis goes on praising listening and speaking as the former is not tested at all and the latter is not tested beyond paper and pencil dialogue format. Thus, the findings account for the students' low level of global language competence.

1. Introduction

In education and educational institutions, teaching, learning, and testing are the most important interlaced activities. Scholars in the area of testing such as Nigussie (2002), Alderson (1991), Weir (1993), Teshome (1995), Heaton (1990), Hughes (1989), Ebel and Frisible (1991) and Bachman (1990) have given due attention in their works on content validity.

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. These content areas are all the points that students have been practising in the class. If the interaction between the contents of the textbooks and sample EGSEC SATs is high, we can say that the test is content valid and if the interaction is very less, we can say that there is low content validity. In the teaching and learning process, teaching-learning and testing might have a positive or negative relationships 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 washback effect on teaching and learning and fails to measure accurately whatever it is intended to measure (Hughes, 1996:1). Not using enough teaching techniques that need the memorization or recitation of surface meanings or hard facts can have a negative effect on testing in which the teacher may prepare test items that need the memorization of these surface meanings.

In other words, teaching which is based on the recent student-centred communicative approaches mainly stresses on the use of high amount of cognitive efforts that leads to the preparation of better test items that requires high level of mental effort. The quality of the test items might have an impact on the students' preparation for exams. The effect of the relationship of teaching and learning may have a positive or negative impact on the behaviour of the students as the researcher has been observing in his six years working time. This process is known as back wash effect. Hughes (1990:1) defined the word as "the effect of testing in teaching and learning is known as backwash".

Hughes (1989) and Weir (1990) suggest that to offer much support to the teaching and learning process, a test has to satisfy three major criteria of a good test: a test should be content valid - test what is intended to measure, reliable - scores should be consistent when repeated, and practical- a test should be economical, easy to administer and score.

Although validity, reliability and practicality have a mutual contribution to the good quality of a test, this study focuses on the content validity of the grade ten's EGSEC English examinations, which are prepared by the National Educational Assessment and Examination Agency (NEAEA). NEAEA is a branch of the Ministry of Education of Ethiopia that is authorized to prepare, evaluate, administer, score and announce students' results, since EGSEC examinations qualify candidates for pre-university admission in country. The researcher feels that a study on content validity, especially in relation to standardized achievement tests (SATs) is worthy and has paramount importance in the quality of education in the country as it contributes to the improvement of the curriculum in general and language education (teaching-learning and testing) programme in particular.

A number of local researchers like Nugussie (2002); Asmare (2008); Abraham (2009); Teshome (1995); Kifle (1995); Nuru (1992); Tibebe (1992) and Alemu (1983) have made their studies on testing areas. Some of them were conducted on validity aspects. For example, Nugussie (2002) has conducted a study on the 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. The then researcher found the content validity of the EGSEC English examination items in 1993 E.C. were very weak and divergent from the contents of the syllabus. And the statistical result showed that the test items of the EGSEC English examinations were not relatively valid and the test items did not match with the syllabus contents. The gap the present researcher has noticed in that research is it was confined to one year test items only and failed to see other years' test items to have guarantee on the invalidity of the EGSEC English examinations.

Another research by Asmare (2008) discuses content validity of three years' teacher made achievement tests (TMTs) of English language at Hawassa College of Health Sciences with reference to the observation of the textbook or syllabi and sample test papers' contents with the purpose of assessing if the coverage of English language tests administered in the college fairly represent the coverage of the textbooks. The result of the study shows that the contents of the sample test papers do not adequately represent the coverage of the textbooks.

Kifle (1995) has also conducted a study on content validity of grade ten English language tests with reference to the former textbook-*English for Ethiopia*- which was structural approach based. Alemu (1983) has also conducted a research on National Examinations for grades six and eight. One of the specific objectives of the study was to see the content validity of the examinations. The results of these study show that the examinations lack content validity.

The present study focuses on assessing content validity of English language tests at GEQAEA of EGSEC level. It aims to assess the extent to which EGSEC English language tests of NEAEA are valid content wise. As stated earlier, the previous studies by Alemu, Kifle, Abraham and Tamirat on content validity (old textbooks) of each sample tests were found to be weak. The present study focuses on assessing the content validity of four EGSEC English examinations (2009-2012) in comparison with the 1996 (old) and 2010 (new) textbooks editions (syllabi). Are these standardized achievement tests adequate representatives of the contents of the textbooks? The current study has tried to answer this question. The previous research papers had measured only the relationship between the contents of the textbooks and contents of the test papers. In this research, apart from the relationship, the strength of association between the contents of textbooks and sample SATs are also included to make the research more valid.

The present study is distinct from the previous studies in terms of procedural analysis too. The present work is more so of assessment than judgemental in nature. To arrive at the right conclusion, sample test papers, official students' textbooks and teacher's guide or syllabi were collected. Then the amount of periods allocated(old textbooks) and frequencies of practice items (new textbooks) for each major headings or sections in both grades' syllabi were observed and determined how many questions are expected from each major heading or section by using chi-squared test of independence. And finally, the result of the chi-square test tells us the content validity of the selected sample test papers, that is, the relationship between the contents of the textbooks and sample SATs.

1.2. Statement of the Problem

There are several reasons why the researcher intends to emphasize on studying the content validity of EGSEC English examinations. The main reasons for studying the content validity of the EGSEC English language SATs were the observation of test papers while the researcher was in Axum Secondary School as a secondary school teacher. The treatment of all contents of the textbooks did not seem well represented in the SATs of EGSEC English language exams in accordance with the syllabi. Consequently, more than half of the school students were not able to do well in the exams and in turn were unable to get admission for preparatory school.

Hughes (1989) notes that if a test lacks content validity, it results in a harmful washback effect. It means, if a test is biased to certain content areas in a syllabus/textbook at the expense of other content areas, those content areas which are less considered or not considered, will receive little attention or no attention from learners while practising and studying. This means students give less attention to less considered or not considered skills and language areas when they study, that is, language skills development and use will receive partial treatment (psychomotor domain).

On the other hand, if all content areas in the domain treated in the classroom are given equal or nearly equal chance to appear in the test, there will be beneficial washback effect on teaching. That is learners will give balanced attention to all content areas of the syllabus during their practice and study. The attitude and motivation (affective domain) of students towards studying these contents will be high. This is well strengthened by Siddiek (2010), as tests are supposed to focus on core syllabus constituents to find out how much the learner achieved. Generally, the more representative samples a test contains, the more reliable will be the assessment of student's knowledge and ability.

Testing experts and their co-workers in the GEQAEA are currently working and contributing their efforts for the implementation of the new curriculum devised in 1994 and attained a good progress in the quality of testing nationwide. However, some researchers in the area of testing are expressing their threats on the content validity of the EGSEC English exams. There are many complaints not only from students' parents about the results of their children but also from employers, who are unhappy with the performance of some graduates in their implementation of language skills in real life situations

Thus, the researcher intends to assess the content validity of the EGSEC English language exams in response to the aforementioned divergent views. The question remains open whether the problem is on the part of EGSEC English language test designers; or the tests themselves lack content validity.

1.3. Objectives of the Study

1.3.1. General objective

The general objective of this study is to assess whether the test papers of EGSEC English examinations prepared and administered by NEAEA from 2009-2012 possess content validity or not.

1.3.2. Specific objectives

The specific objectives of the study were:

1. To assess the extent of conformity between all language areas (phonological, morphological, lexical, and syntactic) stated in the syllabi or textbooks and the contents of EGSEC English test papers.

2. To explore the degree of covering textbooks contents (language skills) in the EGSEC English examinations.

3. To investigate the views and awareness of English language teachers and testing experts about content validity of EGSEC English examinations.

1.4. Research Questions of the Study

1. To what extent all the language areas (phonological, morphological, lexical, and syntactic structure) stated in the syllabi and textbooks are clearly addressed in EGSEC exams?

2. To what extent all the language skills (speaking, listening, writing, and reading), vocabulary and grammar stated in the syllabi and textbooks are clearly addressed in the EGSEC exam?

3. What are the school teachers' and testing experts' views on the content validity of the EGSEC Exams; awareness and practice of testing valid tests, and the impact of content valid English language tests on language teaching and learning.

1.5. Scope of the Study

The study covered only the four recent consecutive standardized English language achievement tests prepared and administered by the National Educational Assessment and Examination Agency from 2009-2012. Moreover, views of English language teachers of Harar Secondary School on content validity of EGSEC English examinations and testing experts of the regional bureau of education were included.

1.6. Significance of the Study

As a matter of fact, the result of any study should have importance to the users. The result of this study is hoped to contribute towards preparing high content valid EGSEC English examination and providing feedback to testing experts in the National Educational Assessment and Examination Agency because results of the study can explain the experts' strategies, quality and efficiency. Furthermore, it will serve as springboard for potential researchers who would like to conduct a research on content validity and related issues.

1.7. Delimitations of the Study

The purpose of this paper is to examine whether or not the EGSEC English examinations administered from 2009-2012 adequately represent both the contents of the textbooks and learning outcomes stated in the syllabi. The research could be more comprehensive if it were to analyse whether the item analysis (item difficulty and item discrimination) to see the test qualities on which effectiveness could be assessed. However, the time and financial constraint could not allow this comprehensive assessment to be done. Therefore, the researcher has delimited the scope to the content validity of EGSEC English examinations only.

1.8. Limitations of the Study

This study focused on assessing of grade 9 and 10 English textbooks and four recent standardized achievement tests administered by the National Educational Assessment and Examinations Agency. However, the researcher encountered several kinds of problems in effectively conducting his research activities as it was scheduled. The first problem was finding the test papers, 1996 editions (old textbooks) and their syllabi. The second problem was the 2010 editions syllabi do not have clearly stated periods allocation for each major language section. To compensate this problem, the frequencies of practice items or tasks in the textbooks were used. And finally, this paper was confined only to the study of content validity of standardized achievement tests.

1.9. Definitions of the Key Terms

EGSEC - an exam designed by the National Educational Assessment and Examination Agency, administered and used to select those who could only join preparatory or move onto the technical vocational and educational training (TVET) in the years 2009-2012.

Content validity- refers to the EGSEC English examination contents which cover or are representative sample of all the language areas (phonological, morphological, syntactic structure) and skills (speaking, reading, listening, writing) plus grammar and vocabulary specified in grades 9 and10 English language (1) syllabi, (2) teachers guide and (3) student's textbooks.

Backwash - (1) The effect of test on teaching and learning, including effects on all aspects of the curriculum, including materials and teaching approaches, as well as on what the students do to learn and prepare for tests. (2) The effect of test on teaching and learning.

Major Language sections - refer to the six main sections of 'English for Ethiopia' (writing, reading, speaking, listening, grammar and vocabulary) used in the textbooks and EGSEC English exams of the grades under study.

Standardised Achievement Tests - tests used to identify how well students have met textbooks' objectives or mastered textbooks' contents.

Teacher Made Tests - refer to tests made by classroom teachers

Item Analysis - process of statistically identifying how well the questions did their jobs, and which ones might need improvement. It includes looking at item difficulty (level of difficulty).

Item Discrimination - estimate of how well an item separates high- and low-ability test takers.

Behaviour/behavioural Domain: a type of knowledge, skill, or performance that the EGSEC English examination is claimed to assess.

2. REVIEW OF RELATED LITERATURE

Many relevant studies have been made on content validity related areas. Therefore, the researcher makes a brief survey of previous studies related to the problem. For this reason, issues like the meaning of content validity, pros and cons of content validity, evaluating the content validity of achievement tests, the importance of content valid achievement tests to language learning, guidelines to establish content validity, the importance of preparing a table of specifications in content validation of tests, steps to be followed in constructing test of content validity, standards of measuring content validity and finally, lack of content validity and backwash are discussed below.

2.1. The Meaning of Content Validity

Many testing scholars have given their definitions to the concept of content validity. All definitions are similar in their central focus except the way they are worded. The following are some of the definitions given by different scholars.

Harrison (1983) and Alderson, Claphan, and wall (1995) defined content validity as a test's ability to include or represent all of the contents of a particular domain (syllabus) in a proportional manner. Similarly, Bachman (1990) defines "Content Validity is the extent to which the tasks required in the test adequately represent the behavioural domain in question." Moreover, Weir (1990) describes content validity as "The extent that tests sample as widely as possible relevant, critical and communicative items from the syllabus in order to have a beneficial washback effect on teaching."

In line with this Hughes (1989:26-27) stated the meaning of content validity as follows:

> *A test is said to have content validity if its content constitutes a representative samples of the language skills, structures and etc. with which it is meant to be concerned. The test would have content validity only if it included a proper sample of the relevant structures. Just what are the relevant structures will depend, of course, upon the purpose of the test. We wouldn't expect an achievement test for intermediate learners to contain just the same structures as one for advanced learners. In order to judge whether or not a test has content validity, we need specification of skills and structures that is meant to cover. This should be made in an early stage of test construction. A comparison of test specification and test content is the basis for judgments as to content validity.*

From the above idea it is possible to understand that to conduct a standardised achievement test preliminary requirements such as identifying the purpose of the test, setting specifications and analysing the contents of the textbooks (language skills, vocabulary, grammar, etc.) are very important aspects.

According to Underhill (1991) content validity refers to the concept if the test produces a reasonable sample of the contents of the textbooks and/or syllabus. Content validity is more important than other validity measures such as face validity, construct validity and concurrent validity. This idea is more strengthened by Hughes (1989) in her book entitled 'Test for Language Teachers' as content validity is quite relevant because it is a means to assure how deeply language tests sample the instructional objectives or the universe of certain behaviour. The test items help to check the performance of the students or their level of progress or level of language competency in each content areas of the syllabus. In addition to this, the test assesses every objectives of the course attended by the students in the classroom. Other test experts like Heaton (1988 and 1990) regard content validity as an important feature of tests and more specifically Weir (1990) focuses on contents of standardized achievement tests must reflect what has been in the syllabus and treated nationwide to arrive at the relevant decisions about students' performances.

Moreover, many authors have oriented the definition of content validity and its role in measurement and evaluation process. In this regard Ogunniyi (1991:84) stated that the content validity of a test deals with whether or not the test can measure the ability in question. Thus, a test whose items do not reflect what has been thought cannot be regarded as content valid. From this point, we can realize that to say a test has content validity; it should not deviate from classroom instruction. In related issue, Gronlund (1982:79) stated that for a test to have high content validity, it should contain a representative sample from the topics and course covered. Robert M. Thorndike (1997) elaborated the content validity under the title 'Content –Related Validity' that "when the test has maintained an appropriate balance in emphasis for both content and mental process by allocating a different number of items to each main heading and process on the test, then it can be said contently valid." Content validity of a test is the representativeness or sampling adequacy of the content, the substance, the matter, the topics of a measuring instrument (Kerlinger, 1973). In related issue Mehrens and Lehman (1991) stated content validity as it is related to how adequately the content of the test samples the domain about which inferences are to be made. In related development Lennon (1980) further stressed that for achievement tests, content validity evidence is by far the most important type of validity evidence.

Content validity, sometimes called logical or rational validity, is the estimate of how much a measure represents every single element of a construct. For example, an educational test with strong content validity will represent the subjects actually taught to students, rather than asking unrelated questions (Carmines & Zeller, 1991, P.20). Content Validity is based on the extent to which a measurement reflects the specific intended domain of content (Ibid).

According to the Wikipedia, the free Encyclopaedia, a test has content validity if it measures knowledge of the content domain of which it was designed to measure knowledge. Another way of saying this is that content validity concerns, primarily, the adequacy with which the test items adequately and representatively sample the content area to be measured. For example, a comprehensive maths achievement test would lack content validity if good scores depended primarily on knowledge of English, or if it only had questions about one aspect of maths (e.g., algebra). Content validity is primarily an issue for educational tests, certain industrial tests, and other tests of content knowledge like the Psychology Licensing.

2.2. The Pros and Cons of Content Validity

To check the degree of students' performance achieved through textbooks, a representative test of the instructional objectives is crucial. Determining whether students’ are making progress or not is possible only if the exam is based on the content coverage of their textbook they have practiced in the classroom.

Furthermore, ensuring content validity while constructing tests, have a paramount importance in quality of education in general and that of language proficiency of students in particular. In line with this Hughes (1998:27) also stated about the importance of content validity as follows:

What is the importance of content validity? First, the greater a test’s content validity, the more likely it is to be an accurate measure of what it is supposed to measure. A test in which major areas identified in the specification are under-represented or not represented is unlikely be an accurate. Secondly, such a test is likely to have harmful backwash effect. Areas which are not tested are likely to become areas ignored in teaching and learning. Too often the content of tests is determined by what is easy to test rather than what is important to test. The test safeguard against this is to write full test specifications and to ensure that the test content is a fair reflection of these. For this reason, content validation should be carried out while a test is being developed; it should not wait until the test is already being used.

From the above extract, it is possible to infer that the content validity of tests should be carefully scrutinized and considered especially preparing content validity chart guarantees the content validity of tests. If not, the consequence of the test would have a harmful washback effect for the teacher(s) and students. Teachers should not evaluate their students' achievement power by constructing content poor and invalid tests.

When preparing a content valid test, problems of knowing what to include in test, that is sampling, may sometimes arise in designing an achievement test because there is often so much ground to cover. It may find it difficult to know what to leave out of the test and what to put in.

Hughes (1989) has definitely put where to stress our attention, particularly, in achievement testing in the following lines: “If achievement tests are based on the detailed teaching and textbooks content/objectives, they will provide a truer picture of what has actually been achieved in teaching and learning and it will tend to be evaluated against the objectives of those contents. As a result there will be constant pressure to achieve them.” Supporting this idea Messick (1975:961) confirmed that the primary limitation of content validity, then, is that it focuses on tests, rather than test scores. This limitation has been characterized by Messick in an extended simile:

Content validity is like the barker outside a circus tent touting two bowing aerialists, a waving clown, and a poster of a lady riding a unicorn as a sample of the show you will see inside. It is not just that the sample is not representative of the variety of individuals and animals inside or even an accurate portrayal of them; it is that you do not see any performances.

It has frequently been claimed, for example, that authentic tests are valid simply because they are authentic, and look good. At the extreme, the notions of content and face validity become touchstones of test validity (Messick, 1989).

2.2.1. Consequences of content validity

When writing achievement test items, writers must begin with a list of content standards written by content specialists which specify exactly what students are expected to learn in a given school year. The goal of item writers is to create test items that measure the most important skills and knowledge attained in a given grade level. The number and type of test items written is determined by the grade level content standards. Content validity is determined by the representativeness of the items included on the final test (http://en.wikipedia.org/wiki/Achievement_test).

Content validity –representing of sample of the language skills, structures, etc- is a very important factor in the success of tests. When all the language elements and skills are included in the test it will put pressure on the learner to cover the whole syllabus to read all the materials in the syllabus. It also makes teachers focus on teaching the specific components within the specific course in the specific time. Content of test puts the learner and the teacher on the right track by committing themselves to the syllabus which is especially designed by experts to secure the educational objectives of the individual and the community (Siddiek 2010).

On the other hand, if the sampled test papers are not representative of the stated syllabus, the chance of test takers to demonstrate their English language competence will be limited. Hughes (1989) notes that if a test lacks content validity, it results in a harmful washback effect. It means, if a standardized achievement test is subject to certain content areas in a syllabus or textbooks at the expense of other content areas, those content areas which are less considered or not considered, will receive little attention or no attention from learners and teachers while studying and teaching, respectively.

To arrive at the right conclusion whether an exam is content valid or invalid, evaluating the content validity of test papers has a paramount importance to ensure quality of tests. Many testing experts tried to mention the way on how to evaluate the content validity of tests in their own perspectives at various times. The process of content validation involves having experts' judges who compare systematically the test items to the content domain. If the table of specifications were available, it would make easier the process of content validation. For this reason, a summary of the statements of these experts is compiled in the following way.

According to Mehrens and Lehman (1991) there is no single commonly used numerical expression for content validity, content validity is typically determined by a thorough inspection of the items. Each item is judged on whether or not it represents the total domain or the specified sub domain. They also noted that "some individuals report a content validity index as the proportion of items rated as matching the domain or sub-domain which it was originally intended to sample."

The Standards for Educational and Psychological Testing (AERA, 1999) is a joint publication of the American Educational Research Association, American Psychological Association and the National Council on Measurement and it suggests that the most common method of providing evidence of content validity for any test or assessment is to have content area experts rate the degree to which each test item represents the objective or domain. The items are like the critical tasks and the domain is represented by the state objectives. This statement can be rephrased as content validity is typically estimated by gathering a group of subject matter experts (SMEs) together to review the test items. Specifically, these SMEs are given the list of content areas specified in the test blueprint, along with the test items and objectives intended to be based on each content area. The SMEs are then asked to indicate whether or not they agree that each item is appropriately matched to the content area indicated. Any items that the SMEs identify as being inadequately matched to the test blueprint or flawed in any other way, are either revised or dropped from the test.

On the other hand, Thorndike (1997:131) explicitly stated the way how to evaluate the competence in English language in content valid way as follows:

First, we must reach some agreement as to the skills and knowledge that comprise correct and effective use of English. If a test is to be used to appraise the effect of classroom instruction, we must specify the subset of skills and knowledge that have been the objectives of that instruction. Then we must examine the test to see what skills, knowledge, and understanding it calls for. Finally, we must match the analysis of test content with the course content and instructional objectives to see how the former represents the latter.

According to Alderson *et al* (1996:193), to evaluate the content validity of tests we need to compare the test contents with the content of the textbooks or syllabus. To obtain further information so as to strengthen the information about content validity, we need to discuss with subject teachers or specialists in this regard. In addition to these, comparing the test contents with content of the textbook with precise list of criteria is one of the ways to evaluate content validity.

Harris (1979:19) further added that:

If a test is designed to measure mastery of a specific skill or the content of a particular course of study, we should expect the test to be based upon a careful analysis of the skill or an outline, not just those aspects which lend themselves most readily to a particular kind of test question.

From the above definitions and procedures, which were presented by different testing experts, to assess the content validity of tests, we must see the match between the test content with the content of the course material or classroom instruction. In other words, it is to say that to identify whether a given test has content validity or not, it is very important to compare the content of the test with the content of curriculum materials so as to judge whether a test proportionally represents, under represents or not represents the course content.

Generally, content validity refers to what goes into the test in relation to what has been taught or covered in the classroom. So, in such achievement tests more emphasis is given to content validity than the other validity types such as predictive validity, construct validity, face validity and concurrent validity types. Supporting this view Aggrawal (1998:267) describes as:

Content validity is a professional judgement which has the teacher or tester. They rely on their knowledge of the language to judge that to what extent the test provides a satisfactory samples of the syllabus, whether real (for achievement testing) imagined (for proficiency of testing) or of the theory or model (for aptitude Testing).

From this view, therefore, an appraisal of content validity involves careful and detailed examination of the actual test tasks. In the same view, Davies (1990:23) has also stated as follows:

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 of the objectives of the course and the course content/syllabus.

Another aspect in conceptualizing content validity is as to ensure the validity of the test. The basis of judgment of the test’s content validity is a comparison of the specification and the test content. 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) paraphrased the concept of content validity as “this kind of validity depends on a careful analysis of the language being tested and of the particular course objectives or course contents".

In preparation for exams, Nitko (1996:39) stated that students expect exams to appear from what has been emphasized in class. If teachers have spent lots of time in one of the content, that area should be featured prominently in the assessment. Besides, the assessment method should reflect learning targets, which are appropriately identified and was taught in the classroom. Therefore, a test is said to have content validity if it is a representative sample of the contents and objectives in the syllabus. To conclude, Hughes (1989:23) supports the idea as, "the test would have content validity only if it included a proper sample of the relevant contents of the textbooks".

2.3. The Importance of Content Valid Achievement Tests to Language Learning

According to Hughes (1989) claims that, though all aspects of validity have significance in teaching and learning, content validity, for example, is the most relevant one because it is a means to check the attainment of objectives of each contents of the syllabus or the universe of a certain domain. If a certain test is valid content coverage wise, it provides relevant information to the concerned bodies that how much students have progressed in each contents or the syllabus.

In addition to this, tests which involve a proper content validity initiate learners to study with hard commitment to each content of the course. Otherwise, students will avoid studying and practising those language areas and skills which do not appear or appear less in tests. They give a considerable attention to those language areas which they are going to be tested. In favour of this, kohonom (1999) notes as:

It is known that anticipation of testing procedures has a washback effect on learning; learners prepare for examinations and organize knowledge in memory in the light of how they are going to be tested. ... evaluation thus affects both quality and quantity of learning. Therefore, it needs to be examined in terms of both the learning process and the outcomes of the learning.

Weir (1993) also says, “Students could carry out 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 at a particular grade level.

To sum up, if tests contain satisfactory samples from each content areas of a given syllabus, they can have a positive influence on teaching-learning. It is important, therefore, that tests should sample as widely as possible test items from each contents of the domain. As a result, decisions that one makes about students’ performance level and the program or syllabus, will be more acceptable. Decisions, based on scores gained from tests which are poor at content validity, are likely to be imprudent (less acceptable). Tests with appropriate content validity are helpful to increase quality and quantity of learning providing valid and reliable information about students’ performance.

2.4. Guidelines to Establish Content Validity

As in the aforementioned discussion, content validity is a vital feature of any test. Multitude of scholars in the area of testing share the idea that tests of a good content validity facilitates learning a language and other subjects. Hence, test writers should give a considerable care while writing tests. Anastasi (1976) and Weir (1990) suggest some very important guidelines that help test constructors to establish content validity of tests. These guidelines read as: the behaviour 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. And finally, content validity depends on the relevance of individual’s test response to the behaviour area under consideration, rather than on the apparent relevance of item content.

Thus, language test writers are advised to undergo through these guidelines in advance in order to produce tests which involve adequate content validity since content validity is one of the key promoters of learning language. If the test designers do not follow the suggested guidelines, it is very likely that they forget to include relevant language elements into a test. Even the included ones will lack proportionality and some language areas will be dominant in the test whereas others will get little attention (Ibid).

2.5. The Importance of Preparing Table of Specifications in Content Validation of Tests

Walelign (2006:62) noted that table of specification (TOS) is a device that has been widely used for the purpose of making a test valid measure of the instructional objectives and course content. It is sometimes called test grid, or content validity chart. According to Siddiek's (2010) idea, analysis and planning of the content of an achievement test is very crucial. To represent the achievement test for any educational material, we must analyze the content of this educational material to see how much of truth the specific test shows by calculating and then determining the relative weight on the basis of time spent in the teaching of each subject. The identification of the relative weights of the educational objectives of the subject depends on the experience of the teacher. It is often the active and knowable expert teacher we need to consult in determining these weights depending on the objectives to be measured. Achievement tests should be representative sample of educational outcomes and the content of educational material. Practically, we cannot put all the questions that we want to ask in the test at the end of the semester, where there are hundreds of facts and terminology we expect learners to know within the time allowed; but we can insert an appropriate number of paragraphs in the test to measure parts of the facts or applications.

A test without plan will be a harmful washback effect. Here, it must be emphasized that before starting to write any test items, the test constructor should draw up a detailed TOS showing aspects of the skills and language areas being tested and giving a comprehensive coverage of the specific language elements to be included. A classroom test should be closely related to the ground and cover the classroom teaching (Gronlund, 1990).

2.5.1. The purpose of TOS on content validation

The most important purpose of TOS is to achieve a balance in the test, and make sure that the measurement of the samples is highly representative of the objectives of our teaching, and the subject content that we want to measure in the achievement test, to secure the ultimate goal of education. Educational goals embody the mental, physical and spiritual growth of the human individual. It also constitutes of the aim to change the individual person's attitudes positively by engineering behaviour to secure pedagogical objectives (Siddiek, 2010).

2.5.2. Benefits of TOS

Table of specification helps build a balanced test with the size of efforts in teaching each topic and gives real weight of each part of the subject, and therefore, every subject deserves questions according to its relative importance. It also helps in the selection of a representative sample of the objectives of teaching, in an orderly manner. And finally, it gives the student a lot of confidence of test fairness, which will assist the candidate in the organization of his/her time.

2.5.3. Component of table of specification

The TOS will consist of the following components. These are the relative weight of the subject and vocabulary that will be measured by a student achievement performance and the relative weight of the objectives to be measured. It also consists of the relative weights that will determine the number of questions contained in the test of each topic and each level of the cognitive objectives as well as the relative weights to determine the questions for each level of objectives to be covered by the test.

2.5.4. Principles be taken into account when building TOS

Any test constructor must consider the following principals when writing a TOS: the nature of the subject, the educational set of behavioural objectives; the length of time it took to teach each topic of the course, and the nature of students with regard to the level of study.

2.6. Steps to be Followed in Constructing Test of Content Validity

To construct a content valid test (Walelign, 2006:246-247), one should do the following steps:
1. List major topics or contents and types of behavioural changes to be measured separately.
2. Give weight for various subject matter/topics and behavioural changes in terms of their relative importance.
3. Build TOS from the weighted lists of topics and behavioural changes. The table should show the relative emphasis given to each topic and behavioural change.
4. Select/construct test items in line with the TOSs. The closer the test corresponds to the specification indicated, the greater the degree of content validity.

To assess the content validity of an examination, one should relate it with the course objective i.e. asking questions such as: what were the aims of the courses? Do the questions in the examination afford a serviceable means in determining the extent to which the aims have been realized? What were the contents of the syllabus? Does the examination adequately sample these contents? And do the syllabus and the examination correspond in terms of properties and distribution of emphasis?

2.7. Standards of Measuring Content Validity

The purpose of this study is to assess the content validity of the EGSEC English examinations based on the prevailing standards (notions) of Anne R. Fitzpatrick about content validity and the process of content validation are described as they pertain to the following concepts (standards):

(a) Domain Sampling
(b) Domain Relevance
(c) Domain Clarity
(d) Technical Quality

2.7.1. The concept of domain sampling

Central to most test specialists' views of content validity is the general concept that this validity refers to the adequacy with which a test samples the domains that the test is claimed to cover. However, when the specific terms used to explain this concept are examined, it becomes clear that these specialists have interpreted and operationalized the concept in different ways, but the discussion here is limited to the adequacy with which the EGSEC tests sample the behavioural domains stated in the textbooks or syllabi.

2.7.1.1. The test as a content sample.

According to Cronbach (1971), Lin (1974,1980), Loveinger (1957) and Messick (1975) view content validity as relevant to test content issues and have indicated that this validity is dependent upon the adequacy with which the items of a measure constitute an adequate sample of the content domains that a test is claimed to cover. Using the term "universe" when referring to the content domain, Cronbach designed this view when he stated "an achievement test is said to represent a body of content outlined in the test manual. To ask "Are the tasks used in collecting data truly representative of the specified universe?" is to examine content validity.

To assess the adequacy of a test as a content sample, it has typically been suggested that content experts be asked to judge: (a) how well each item of a test corresponds to the defined content domain that the item was written to reflect, and (b) how well sets of items represent the content domains to which they are judged to correspond (Brown, 1976; Thorndike and Hagen, 1977). In the standards, it is said that the items of a test be examined to determine whether they appear to call for a representative sample of the intended behaviours (Roseboom, 1966; Anatasi, 1976; Mehrens and Lehman, 1978)

2.7.2 The concept of domain relevance

The relevance of test content: measurement textbooks and test development manuals have commonly indicated that a content valid test must cover important aspects of the content universe that the test user wishes to assess (APA *et al.*, 1974; Thorndike and Hagen, 1977). They imply, therefore, that a measure's content validity depends, in part, upon whether the content domains that define a measure are relevant to the important part of some universe of, say, academic, or job content that interests a test user. The assumption of this study is: *Are the contents of the SATs being used in the EGSEC English examinations relevant (adequate representative sample of the actual universe) to identify students who could continue to grade 11 and 12, leave school after grade 10 or move into the technical vocational and educational* training (TVET) in the years 2009-2012?

Cureton (1951) had the following meaning in mind when he discuss test relevance, which he viewed as " the degree to which the test operations as performed upon the test materials in the test situations agree with the actual material in the situation normal to the task.

Generally, the relevance of a test content is when a person wishes to use a test appraise performance on a particular content universe of interest, the relevance of a test's content to this universe will be important to determine. Many test users who wish to do this; teachers commonly use tests to appraise students learning of the material presented in the course. By 99establishing that the content of test represents important aspects of the content universe that is of interest, a user can gain logical support for the claim that examinee's performance on this universe (Brown, 1976).

2.7.3. The concept of domain clarity

Also included in most test specialists' views of content validity is the idea that this validity is determined, in part, by the clarity with which the content domains of a measure are defined (Anatasi, 1976; APA *et al.*, 1974; Lennon 1980; Linn, 1980; Rozeboom, 1966). Measurement texts have indicated that achievement measures are well specified when the subject matter and cognitive process to be measured are indicated, and the number and format of the items that are used to measure each behaviour of interest are noted (Brown, 1976; Rozeboom, 1966; Thorndike and Hagen 1977).

Specifically, it has been suggested that this definition comprises a detailed description of the content, structure and scoring of the items that are used to measure each behaviour a test is intended to assess (APA *et al.*, 1974; Cronbach, 1971; Popham, 1978*).*

2.7.4. The concept of technical quality in test items.

Study of the technical quality of test items, using appropriate empirical and logical procedures is the final kind of investigation that has been mentioned by test specialists, albeit infrequently, as requisite for establishing the content validity of a measure. Hamilton and Eignor (1979) and Benson (1981) viewed content validity as resting upon how adequately the items of a test represent a test's content domains. These researchers suggested that test items that are flawed cannot be considered adequate representative of any content domain associated with a test, so that such items, when present will diminish the content validity of the tests.

The presence of numerous discussions in measurement textbooks of the principles and methods of devising effective items suggests that test specialists have ascribed considerable import to the notion that test items should have good technical properties. Commonly, it is said that such properties are necessary in order for a measure to have the level of difficulty, reliability and validity that is desired (Brown, 1976; Mehrens and Lehman 1978). This research emphasises on validity, particularly, content validity rather than level of difficulty and reliability.

As Hamilton and Eignor (1979) implied, it can be assumed good technical quality is a requirement implicit in the specifications of a test's content domain. Consequently, poor items that appear in a test will jeopardize the fit between the test and its definitions. Thus, it appears important to establish that the items of a test are technically sound. To do this, analysis of item content (hereafter contents of the test and textbooks) was carried out to assure the adequate representativeness of the textbooks contents in the leaving SATs of EGSEC.

2.8. Lack of Content Validity and Backwash

Backwash is an important concept in the field of applied linguistics. It refers to the impact and influence that the test can have on the teaching and the 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 backwash on both the teacher and the learner and vice versa (Siddiek 2010).

3. METHODOLOGY

This chapter elaborates an overview of the research methodology. It incorporates an account of the research design, research setting, description of the participants and sampling procedures, data gathering tools, procedures of data gathering and methods of data analysis.

3.1. Design of the Study

As described earlier, the main purpose of this study is to assess the content validity of the EGSEC English examinations by assessing the representativeness of the contents of the sample SATs with respect to the contents of the textbooks or behavioural domains. Therefore, to address the intended research questions, qualitative and quantitative research methods have been used. Qualitative method has been used to interpret, clarify, and justify the data collected through interview and questionnaire, because, interview and open-ended 3questions are mostly the feelings, beliefs, and attitude of respondents which require justification of the researcher. On the other hand, quantitative approach has been used to show the outcome of the chi-square test of independence and Cramer's V contingency coefficient to test the goodness-of-fit and strength of association between the contents of the textbooks and contents of the SATs, respectively. Qualitative approach mostly deals with close-ended questions that do not require too much justification and clarification, in contrast to the qualitative approach.

Goodness of fit refers to how close the observed data are to those predicted from the content domain (universe). Here the key idea of the χ^2 (pronounced 'kigh square') test of independence is to compare the observed numbers of questions in the sample SATs and expected numbers of questions from the behavioural domain (contents of the textbooks) or to determine if the two variables (expected and observed) of interest are independent or dependent. The reason of using chi-square test is to examine differences with nominal categorical variables, because it is a "test of goodness". In this research, it is used for the purpose of two similar but distinct circumstances: (a) for estimating how closely an observed distribution matches an expected distribution- it is referred to this as the "goodness of fit test"; and (b) for estimating whether the expected and observed variables are independent.

The chi-square distribution is determined by the number of degrees of freedom, as in degrees of freedom = *n*-1, where 'n' is the number of population

3.2. The Research Setting

This study has been conducted on four years of EGSEC English Examinations which were prepared by the NEAEA from 2009-2012. The study was conducted in Harari Regional State where the researcher lives which facilitated the researcher to have easy access to the interviewees for the sake of data collection and distribution of questionnaires.

3.3. Participants and Sampling Procedure

The standardized achievement English examinations which were constructed and administered in Ethiopia from 2009-2012 of grade ten were taken for assessment with respect to their syllabi/textbooks. In addition to this, twelve English language teachers of Harar Secondary School and three testing experts of Harari Bureau of Education were selected as respondents in the study. For the basis of the sampling of test papers, the sample standardized achievement test papers, sample interviewees and sample individuals to fill the questionnaire were selected purposively. The criteria for the selection of the subjects were: (a) master or bachelor of degree in the field; and (b) a minimum of six years of direct work experience in the field.

3.4. Instruments of Data Collection

In order to conduct the study regarding the content validity of the exams, the researcher has used three instruments of data collection. These were document (content) analysis, interview and questionnaires with open and closed-ended questions.

3.4.1. Document analysis

As it was stated in the review of literature, the usual way of analysing the content validity of test papers is to compare the test papers with the content coverage of the textbooks and objectives of the syllabi.

Therefore, the researcher has investigated the representativeness between the contents of the textbooks/syllabi practiced in the classrooms against the content coverage of the sample test papers administered from 2009-2012. This is checking the correspondence between the contents of the textbooks and time allocated for each major heading in the syllabi with respect to the administered exam papers. To do so, first, the contents of the textbooks were classified into six major headings or sections as reading comprehension, listening, speaking, writing, vocabulary, and grammar items. In order to increase the reliability of categorization of the main headings of the textbooks, teachers of English language in the study area were interviewed. Then by taking the total number of periods allocated or frequency of tasks for each major heading in both grade levels, their average was taken as total periods.

After allocating the amount of periods spent for each main heading, the amounts of expected number of questions from these headings were computed. In other words, if it is requires to prepare 100 questions for a final year model exam for grade10, the time allocated for the whole course is 200 periods and the amount of time shared for 'reading comprehension' is 20 periods. So, how many 'reading comprehension questions' are expected in this exam? To obtain the expected amount of questions (frequencies) for any cell in any cross-tabulation in which the two variables are assumed independent, multiply the row and column totals for that cell and divide the product by the total number of cases in the table. This can be represented in a simple formula as:

$$fe = \frac{(Row\ Totals)(Column\ Totals)}{Grand\ Total}$$

The above question can be easily computed by multiplying the total number of questions times the total number of periods allocated for reading comprehension and dividing the product by the total period's allotment of the course. The quotient will be 10, which is the number of reading comprehension questions that are expected to be included in the exam paper.

$$\frac{(100\ questions)(20\ periods)}{200\ periods} = 10\ questions$$

Ten reading comprehension questions are expected and this figure be compared with observed reading comprehension questions in the actual test paper. This process is applied to all major headings of the textbooks until 100 questions are prepared.

To determine a significant relationship between the observed and expected number of questions, an inferential statistics, known as Pearson's chi-square (χ^2) test of independence was used. Once we have been calculating the expected amount of questions, the steps of calculating the chi-square were as follows.

(a) Subtract expected values from observed values (O-E)
(b) Square the differences $(O-E)^2$
(c) Divide the square differences by the expected values $((O-E)^2/E)$
(d) Add all quotients

Therefore, the sum of the quotients gives the χ^2 result. This value was interpreted based on the null and alternative hypothesis. In order to triangulate the information obtained from the chi-squared test, questionnaire and unstructured interview were also used. These were employed for twelve teachers of Harar Secondary School and three testing experts in Harari Bureau of Education. The questionnaire and interviews were analyzed qualitatively.

3.4.2. Questionnaire

The questionnaire was used to gather qualitative data by integrating both closed and open ended questions. This means, the researcher gathered information from the school teachers by asking if they test all the sub-skills and areas in the classroom, their views on content validity of the EGSEC English exams prepared by the NEAEA are based on the textbooks and syllabi of the stated grades. They were also asked for their experience as school teachers in the construction of contently valid tests and their overall know how on content validity to compare with the NEAEA's standardized achievement tests. Moreover, testing experts' reaction in Harari Bureau of Education ideas was also appraised.

In doing so, the researcher developed around 20 closed and open ended questions. Some of these items included 1-4/5 Likert scale, containing rating scales. Therefore, the researcher had the chance to triangulate the data with document analysis and unstructured interviews. Most of the results are coded and inserted in tabular form.

3.4.3. Interview

Apart from the textbooks content or syllabi or test papers analysis and administration of the questionnaire, unstructured interview with English language teachers was conducted in the schools to make sure if the contents of the sample tests were sampled from the whole major headings of the textbooks they have taught or not , to know their attitude of teaching towards the forgotten skills in testing (often speaking and listening) and phonological structures at classroom and to request them to deduce the factors that reduce content validity of test items, and so on. The researcher has preferred using unstructured interview because teachers are likely to be reluctant in supplying reliable information. Moreover, unstructured interview also helped in validating the findings drawn from document analysis and questionnaire. To increase the reliability of the interview, oral responses were recorded and documented.

3.5. Procedures of Data Collection

To assess the content validity of the 2009 - 2012 EGSEC English examinations, the following steps were followed. A list of syllabus objectives, teacher guide and student's English textbook for grade nine and ten together with the test items were collected by the content validity researcher so as to code the contents of the sample SATs into themes. The task was done as follows:

1. List of syllabus objectives of grade nine and ten together with the teacher guide and student's textbooks were collected so as to code them with respect to the objectives.

2. Referring to the textbooks and the syllabi of grade nine and ten lists of content areas were drawn up, which were followed by the period allotment in 1996 editions and frequencies of practice tasks or items in 2010 editions.

3. The amount of periods and frequencies of practice tasks allotment for each section of both grades were worked out and put in a tabular form.
4. Both grades' periods' allotment (old textbooks) and frequencies of tasks (new textbooks) were added and their average was taken as total periods and frequencies allotment as the exams were prepared from both of the textbooks.

5. From the compiled periods and frequencies of tasks allotment table, the expected number of questions was determined.

6. Then the number of test items from each section was observed in the test papers.

7. Unstructured interview was conducted for English language teachers in Harar Secondary School, and test experts of Harari Bureau of Education.

8. Questionnaire were distributed among the sample size, collated and compiled (coded).

3.6. Data Analysis

The major headings of the textbooks were divided into six categories reading comprehension, listening comprehension, speaking, writing, grammar, and vocabulary. The total amount of periods allotted for each of these major headings of both syllabi was taken and the average of the figure with its percentages was inserted in a table. Using these prompts as functional exponents, the amount of expected questions from all major headings were computed and those amount of questions were compared with the observed questions in the actual exam papers to see if the exam papers are representative samples of the classroom instruction, partially representative or not at all. To assess the contentvalidity of the EGSEC English examinations, it is believed that the following categorical variables would validate the study.

3.6.1. Independent variable

The independent variable is a variable being used to predict the dependent variable. It is an assessment that involves making a change in the value of the dependent variable. The independent variables used in this research are the period's allocation of the contents of the textbooks and observed values of the test papers.

3.6.2. Dependent variable

The dependent variable is a variable which is affected by the independent variables. This is the chi-square result. The relationship and strength of association between the dependent and independent variables was determined by the value of the chi-square (χ^2) and Cramer's contingency coefficient.

3.6.3. Null hypothesis (H_o)

The null hypothesis (H_o) of this research is there is no association between the row-column variables. In other words, the variables are independent, that is, no relationship or dependence exists between them. When the calculated value of the chi-square (χ^2) test is less than the critical (table) value ($\chi^2 < 11.070$) at $\alpha = 0.05$ significance level with N-1 degree of freedom (DF=5), the null hypothesis is retained or the null hypothesis is not rejected. This means the difference between the observed and expected questions is attributable- that means the observed questions from the six categories (sections) do not significantly differ from the expected questions. The observed ones have a "good-fit" with what was expected.

On the other hand, when the computed value of the chi-square exceeds table value ($\chi^2 > 11.070$) at $\alpha = 0.05$ significance level with N-1 degree of freedom (DF=5), the χ^2 would be declared significant. This means the observed questions would have a "bad fit" with what would be expected. This indicates that there is an association between the variables. Hence, the researcher would reject the null hypothesis.

For example, assume that critical or table value ($\chi^2{}_c$) at α=0.05, with 2 degree of freedom is 5.991. If the computed value of the χ^2 is 42.92, which is seven times greater than the critical value of 5.991, we declare the chi square value significant. This indicates the observed values do not fit with the expected values. Hence, the null hypothesis is rejected, whereas if the calculated value of the χ^2 would be less than the critical value, it indicates the observed questions have a "good fit" with what was expected. For this reason, the null hypothesis is retained or attributable.

3.6.4. Alternative hypothesis (H_1)

The research hypothesis of this study states that there is a relationship between the row and column variables. This hypothesis depends on the conclusion and interpretation of the null hypothesis.

The degree of relationship between the contents of the textbooks and sample test papers was computed by using Pearson's Chi-Squared (χ^2) Test statistical analysis formula as:

$$\chi 2 = \sum \frac{[(O - E)]}{E}$$

Where χ^2 = the value of chi-squared or relationship
Σ=summation sign, read "the sum of..."
O = observed amount of questions
E = expected amount of questions

Summary of steps in a chi-squared test:
Where expected values are not known,
1. Draw up a contingency table of the observed values.
2. (a) Compute the column totals (b) Compute the row totals
3. Assume the null hypothesis of no association between the two variables and work out the expected values for each cell under this hypothesis from the row and column totals. This is done by applying the formula
:

$$Expected = \frac{\text{(row total)(column total)}}{\text{grand total}}$$

4. Compute the value of the chi-square from the formula:

$$\chi 2 = \sum \frac{[(O - E)]}{E}$$

5. Workout the degree of freedom:

$$df = (No. of\ rows - 1)(No. of\ columns - 1)$$

6. Look at the relevant P-value in the table of the χ^2 distribution.

The chi-square test of independence tells us whether the two nominal categorical variables are related or not. It does not tell us how strong that relationship is. When we produce a significant chi-square (the two variables are related), it is natural to wonder how strong the relation is. Chi-square test is not complete without other measures of association such as Pearson's Contingency Coefficient, Phi Coefficient, Siegel's Contingency coefficient, Cramer's Contingency Coefficient for the strength of association for nominal categorical variables. To know the strength of association between contents of the textbooks and sample SATs of this study, Cramer's V coefficient of contingency is preferred among the χ^2 based measures of association. This measure is defined as*:*

$$V = \sqrt{\frac{\chi 2}{n(q-1)}}$$

Where q is the smaller of the number of rows or the number of columns. The smaller the two numbers is used to represent the variable q. If r is number of rows, and c is the number of columns, then:

$$q = (r - 1, c - 1)$$

Cramer's V always takes value in the interval [0, 1]. If Cramer's V is close to 1, then the value indicates a strong relationship/association between the expected and observed variables whereas if the value of Cramer's is close to 0, then the value indicates a weak association between these variables. On the other hand if Cramer's V is equals to 0, there is no association or strength between the two variables, and has a maximum value of 1 when there is a very strong relationship between the two variables (here preferred to say perfect relationship or strength).

The data that was collected using the unstructured interview and questionnaire was analysed in qualitative way of analysis as it did not involve numerical values. The collected data were evaluated, coded and tabulated in percentages accordingly. Finally, the result of the tools was discussed based on the demand of test content validity and quality of language teaching and learning.

4. RESULTS AND DISCUSSIONS

This chapter presents the results, interpretations and discussions on the assessment of the relationship of four EGSEC SATs of English examinations and strength of association between these examinations and their textbooks. It also contains the analysis of the SATs by sub-skills, teachers' and testing experts' awareness of content validity and their real practice on the ground and the impacts of content validity upon language learning in general and the specialists' views on the content validity of EGSEC English language tests in particular.

The results were attained through non-parametric statistics known as Pearson's chi-squared test of independence which is an inferential statistics technique designed to test for significant relationship between two nominal categorical variables. The chi-squared test of independence was used to analyse the content validity of the SATs of the EGSEC English language tests of the stated years. The chi-square test was preferred to test the goodness-of-fit between contents of the sample standardized achievement test papers and contents of the textbooks.

Thus, the χ^2_c value 11.070 with 5 degree of freedom was used for all chi-square based tables to measure the relationship between the content validity of the SATs with respect to the contents of the textbooks. The reason of using this nonparametric statistics is to test the "fit" between the contents of the sample test papers and objectives of the syllabi of the stated grades (expected and observed number of questions). Cramer's V was also used to determine the strength of association between the two variables.

The old editions of English for Ethiopia of grade nine and ten had fourteen units apiece based around a topic of 170 and 140 episodes, respectively, each of which were classified into six major sections. The sections in each unit are reading comprehension, listening comprehension, speaking, writing, vocabulary and grammar. In both grades of the new editions of the textbooks the number of units and episodes were reduced to 12 and 136, respectively. The division of major content areas is similar to the old, but here frequencies of practice items were taken instead of allotment of periods. Hence, the researcher categorized the contents of the textbooks into six main sections (components) based on the classification of the textbooks (syllabus) of the grade levels.

4.1. Analysis and Results of Content Validity of EGSEC English Language SATs of 2009 - 2012

In this sub-section, four of the SATs of EGSEC English examinations constructed, administered, scored and announced by the NEAEA from 2009-2012 have been thoroughly analyzed to address the second research question in order to identify whether the tests constructed in the agency of the aforesaid years do really contain content validity or not. To know the goodness- of- fit between the observed and expected number of questions, chi-square test is used.

χ^2 = (Kigh square) is the summation of empirical (observed) values minus theoretical (expected) values the whole squared divided by the expected (fit) values. Here, it should be noted that "observed values" refer to frequencies occurred in each language content areas of the SATs, whereas "expected values" refer to frequencies that were expected to occur in the SATs or appear in each category or content area. To find the χ^2 value three fundamental steps are followed.

Step 1:

Determine the content areas of the textbooks and sample test papers. Then, put the items in each content area in frequencies (period allotment). Then after, determine row and column totals. The row total is the sum of item frequency in the textbooks and in test papers across each content area. But the column total for the textbooks is the sum of frequencies (periods) of practice items in all content areas of the textbooks, however, the column total for sample test papers is the sum of frequencies of test items in each content area of test papers. In short:

$$\chi 2 = \sum \frac{[(Observed - Expected)]}{Expected}$$

Step 2:

Calculate the expected frequencies (f_e) for each observed *(f_o)* values using the formula:

$$Expected = \frac{(\text{row totals})(\text{column total})}{\text{grand total}}$$

Step 3:

(a) Subtract expected values from observed values (O-E)

(b) Square the differences $(O\text{-}E)^2$

(c) Divide the square differences by the expected values $(O\text{-}E)^2/E$

(d) Add all quotients.

Therefore, the sum of the quotients gives the χ^2 result.

4.1.1 Analysis and results of content validity of 2009 EGSEC English language test items

Table 1. Results of chi-square test analysis with regard to the content coverage of 2009 EGSEC English Language Test Items

Items	Coverage in numbers and percentages							
	Expected(E)		Observed(O)		O-E		$(O\text{-}E)^2$	$(O\text{-}E)^2/E$
	N	%	N	%		%		
Reading	13	17.3	23	30.70	+10	13.4	100	7.69
Speaking	14	18.7	13	17.33	-1	1.4	1	0.07
Listening	10	13.3	0	0	-10	13.3	100	10
Writing	13	17.3	7	9.30	-6	8	36	2.77
Grammar	13	17.3	23	30.67	+10	13.37	100	7.69
Vocabulary	12	16	9	12	-3	4	9	0.75
Total	75	100	75	100	Σ(O-E) =0			$\sum \frac{O-E}{E}$ = =28.97

*N= number of questions and %= the amount of questions in percent.

In the 2009 EGSEC English examination test items, it is difficult to observe any correspondence between the data frequencies (observed test items) and fit values (expected test items). As we can see from the above table, grammar and reading have been given due attention in the test with equal emphasis. In this exam, both of them dominated the testing practice. The fit values of grammar from the total amount of constructed questions which are 75 in number are 13 (17.3%). However, 23 (30.67%) questions were observed in this test paper, which is almost two times greater than it deserves. This means, the amount of test items prepared for each language items is not proportional to the amount of their coverage in the textbook/ syllabi.

The treatment of the different content areas is not balanced. Much of the weight given in the test is to the grammar and reading. On the other hand, vocabulary, and the other skills such as listening and writing are not well treated. The assumption made in this study is content validity as the proportion of test items allotted to each content area with regard to the instructional emphasis and importance of the language items in the classroom instruction.

In this examination, vocabulary and speaking have received relatively fair representation as compared to the other test items, though speaking was represented with its several limitations. The tester(s) tried to assess students' communicative ability through dialogue paper and pencil task format, however, fails to assess students' oral fluency and accuracy as the best way of testing speaking is by making students speak.

Writing test items were also constructed almost less than half of the fit values whereas listening comprehension tasks which were fairly treated in the syllabi (13.145%) received no attention (0%) in this exam. The maltreatment of this receptive skill in such SATs indicates that it will get little attention by students while they are practising the task in the classroom and consequently, they are unable to use this skill in real life situations or target language use domain. This leads to fade the flavour of the communicative competence of listening from the intention of the students as it does not appear in the exam.

The "bad-fit" of the observed and expected number of questions can be statistically determined using Pearson's chi-squared test. Here it must be noted that the larger chi-square value indicates there is a bad-fit between the two variables than the smaller chi-square value and vice-versa. The value of the chi-square test is sensitive to sample size. The larger chi-square value appears to imply a much stronger statistical relationship (significant relationship) between two variables. The table with the larger chi-square value generally provides stronger evidence there is disparity/disproportions between the two variables.

The calculated chi-squared ($\chi^{2)}$ value of the above table, which is 28.97 compared to the critical χ^2 value, χ^2_c=11.070 at α=0.05 significance level with N-1 degree of freedom, DF=5 (Please see Appendix Table 4). The two tailed p-value is <0.001. The p-value of the χ^2 28.97 is =0.00002345. By conventional criteria, this difference is considered to be extremely statistically significant. The p-value answers this question: if the theory that generated the expected values were correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values? A small p-value is evidence that the data are not adequately sampled from the contents of the textbooks.

Thus, the above statement can be further rephrased as the larger χ^2 value indicates that there is a great disparity/disproportion between the observed (empirical) and expected (theoretical) number of questions. Especially, when we look at the representation of listening skill of testing students' comprehension was totally missed out of the test. As the calculated result of the chi-square test (χ^2=28.97) exceeds the critical chi-squared value (χ^2_c=11.070), the data supports the belief that a significant difference exists between the expected and observed number of questions. In line with this Ebel (1971) indicated that for the test to have high content validity, it should be a representative sample of a given course or unit. For this reason, the null hypothesis of this study is rejected as the variables are highly statistically independent or have significant relationship to each other. Hence the exam lacks content validity.

4.1.2. Analysis and results of content validity of 2010 EGSEC English language test items

Table 2: Results of chi-square test analysis with regard to 2010 EGSEC English

Items	Coverage in numbers and percentages							
	Expected(E)		Observed(O)		O -E		(O-E)^2	(O-E)^2/E
	N	%	N	%		%		
Reading	13	17.3	29	38.67	+16	21.37	256	19.69
Speaking	14	18.7	11	14.67	-3	4.03	9	0.64
Listening	10	13.3	0	0	-10	13.3	100	10
Writing	13	17.3	6	8	-7	9.3	49	3.77
Grammar	13	17.3	24	32	+11	14.7	121	9.31
Vocabulary	12	16	5	6.67	-7	10.7	49	4.08
Total	75	100	75	100	Σ(O-E) =0			$\sum\frac{O-E}{E} = 47.49$

*N= number of questions %= the amount of questions and differences in percent

As it is plainly indicated in table 2, the 2010 EGSEC English examination has a great disproportion or disparity between the numbers of test items constructed by the GEQAEA compared to the expected amount of questions. Had there been a perfect match between the expected and observed number of questions in each section their difference would have been zero and the chi-square too. As the table clearly shows, the exam was more of reading and grammar dominated at the expense of other test items.

The data of the table portray that the coverage of grammar and reading in the exam is almost two times higher than what is rightly expected to appear. The representation of writing and vocabulary have reduced in number almost by half than expected to represent in the exam whereas speaking has got relatively fair representation. This means writing and speaking have received a small amount of representation. Here it should be noted that speaking has got its representation in the form of paper and pencil through dialogue task format. Supporting this concept, Harris (1979) said that we cannot test speaking (oral test) through paper and pencil testing. From Harris's statement we can infer that testers lack awareness of testing speaking skill as it requires extended response prompt task format.

The observed data shows how much testers are in favour of grammar (32%) and the receptive skill-reading (38.7%). Here, it is possible to infer, especially, grammar as a mechanism of effectively evaluating the students' language proficiency. Above all, listening skill does not receive any representation at all. This biased testing practice can lead to the biased or inappropriate way of studying and learning to the ignored skill or area of language in the testing system. This can potentially affect their global language progress.

The statistical result of Pearson's chi-squared test also confirms the bad-fit between the dependent and independent variables. When the calculated χ^2 value of the above table (χ^2=47.49) is compared with the critical or table value (χ^2_c=11.070) at α=0.05 significance level with N-1 degree of freedom (DF=5), it is four times higher than the critical or table value. This means observed questions have a bad-fit with what was expected. This indicates that association dependence exists between the variables. The appearance of larger chi-square value implies there is much stronger statistical/significant relationship between the variables. In other words, the greater the chi-square value, the greater the independence between the tabulated variables in the population and, therefore, the null hypothesis is rejected. Evidence for the rejection of this null hypothesis is reflected in large positive value of the chi-squared statistics which is 47.49.

Supporting the above data, Hughes (1989:26-27) confirms that "to say a test has content validity, it should constitute a representative sample of the language skills, structure, etc". However, the practically observed scene is in contrary to this literature. Hence, the exam greatly lacks content validity. This table entails that, there is a great disproportion between the fit-values and observed number of questions and this may resulted in negative washback effect on the students' methods of learning the language.

4.1.3. Analysis and results of content validity of 2011 EGSEC English language test items

Table 3. Results of chi-square test analysis with regard to 2011 EGSEC English Language Test Items

Items	Coverage in numbers and percentages							
	Expected(E)		Observed(O)		O -E		(O-E)^2	(O-E)^2/E
	N	%	N	%		%		
Reading	14	17.5	27	33.75	+13	16.25	169	12.07
Speaking	15	18.75	14	17.5	-1	1.25	1	0.07
Listening	11	13.75	0	0	-11	13.75	121	11.00
Writing	14	17.5	6	7.5	-8	10	64	4.57
Grammar	14	17.5	2 8	35	14	17.5	196	14
Vocabulary	12	15	5	6.25	-7	8.75	49	4.08
Total	80	100	80	100	Σ(O-E) =0		-	$\sum\frac{O-E}{E} = 45.79$

*N= number of questions %= the amount of questions in percent.

Table 3 reveals 35% of the test items represent the grammar and this is followed by 33.75% of the reading skill. This 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. On the other hand, vocabulary (all reiterations except synonyms) and the skills such as listening (totally not), reading, and writing were not well treated in the test paper.

Vocabulary and writing, which were well treated in the syllabi, each of which constitutes 15% and 17.5%, apiece received less than half representation in the exam, which is 6.25% and 7.5%, respectively. In both grades, on average, listening comprehension as a major section shared 21 (13.4%) periods (Please see Appendix Table 1). Despite the coverage indicated in the syllabi, it was totally ignored in all the EGSEC English language testing system as if it does not have any share in the syllabi. This may affect students' attention towards learning listening activities in the classroom. And finally, speaking has got a reasonable representation in terms of quantity, but with its limitations that was already stated above with a task format problem. Students potential of expressing, narrating, eliciting, directing, reporting, describing, etc, were not assessed. Weir (1993) emphasizes students are usually eager to learn what appears in the exam and Nitko (1996:39) added that students expect exams to appear from what has been emphasized in the class.

It can be argued that there is a disparity in treating the language skills/areas of the test items. The syllabus gives weight to the receptive skills that is reading 17.5% and listening 13.75%, but the test items do not weigh them expectantly. The listening skill which has 13.14% in the syllabus is ignored in the construction of items. Only, speaking 18.75% in the syllabi is fairly treated in the dialogue paper and pencil task format which has a weight of 17.5% in the test. The listening skill has become the forgotten skill. The general implication of this test construction leads to negative impact on students' receptive skills of English competency due to the harmful washback effect.

Apart from the above constellations of observed figures, the result of the chi-squared test of independence more elaborates the disparity between the two variables. As shown in table 3, the calculated χ^2 value (χ^2= 45.79) at α=0.05 significance level with 5 degree of freedom is four times higher than the χ^2_c. This is the boldest statement that shows the mismatch or bad-fit between the expected (fit values) and observed amount of questions in the test paper.

Since the statistical result of the chi-squared test is greater than the value of the decision making significance level (α=0.05) with 5 degree of freedom (χ^2_c =11.070) that is $\chi^2>\chi^2_c$, it indicates a disagreement with the null hypothesis and the null hypothesis is not retained or attributable. A p-value of 0.05 or less is usually regarded as statistically significant, that is, the observed deviation from the null hypothesis is significant. Hence, the null hypothesis is rejected. This means the two variables are statistically dependent or related to each other.

On the whole it can be summarized that, both the learning outcomes and contents of the test items did not reach statistically significant agreement with the observed learning outcomes and contents of the syllabi at 0.05 significance level. This means that the test items were not constructed taking into consideration the magnitude and emphasis given to each skill and language area in the syllabi. The EGESEC English examination of 2011 was not relatively content valid. This will lead to the conclusion that the procedure to the construction of test items of the English examination of the EGSEC is not based on a well designed table of specifications (TOS). Besides, the result of the study seems that the NEAEA do not test the content validity of the tests before its administration. Therefore, the test greatly lacks content validity.

4.1.4. Analysis and results of content validity of 2012 EGSEC English language test items

Table 4 Results of chi-square test analysis with regard to 2012 EGSEC English test items

Items	Coverage in numbers and percentages							
	Expected(E)		Observed(O)		O -E		(O-E)^2	(O-E)^2/E
	N	%	N	%		%		
Reading	14	17.5	33	41.25	19	23.75	361	25.79
Speaking	20	25	6	7.5	-14	17.5	196	9.80
Listening	8	10	0	0	-8	10	64	8.00
Writing	13	16.25	7	8.75	-6	7.5	36	2.77
Grammar	15	18.75	27	33.75	12	15	144	9.60
Vocabulary	10	12.5	7	8.75	-3	3.75	9	0.90
Total	80	100	80	100	Σ(O-E) =0		-	$\sum\frac{O-E}{E} = 56.86$

*N= number of questions %= the amount of questions in percent.

When the above table is inspected in terms of domination, it is similar to the exam administered one year before that. (Please see table 3 above). It indicates the domination of the receptive skill - reading and grammar that constitute two times and almost two times greater than expected, respectively. The imbalance between the expected and observed values goes not only to reading and grammar, but also to the rest of the sections. The observed questions in the exam were, less than half for speaking (7.5%), 0% for listening, half for writing (8.75%) and almost half for vocabulary (7%) representations than expected to be appeared in the examination. The implication of the disproportion of this exam diverts or confines the students' attitude towards studying and learning grammar and reading or to language areas that appear more and frequently in number in the exam.

What does the chi-square test of independence illustrate about the above table? The 2012 EGSEC English examination to be significant at 0.05 significance level with a given 5 degree of freedom, the χ^2 value has to be at least $\geq$11.070. Therefore, we can reject the null hypothesis of no association between the two variables. This is because the data tells us that there is a significant relationship between the contents of the textbooks and the sample standardized achievement test.

As findings of the study in the table revealed, the calculated χ^2 value 56.86 (actually 56.855) is far greater than the critical χ^2 value of 11.070 with 5 degree of freedom at α=0.05. This proves the presence of great disparity or disproportion between the two categorical variables- expected and observed number of questions. This is because the data tells us there is a significant relationship between the contents of the textbooks and the sample standardized achievement test. Both reading and grammar together should contribute 36.25% of the exam, but in reality these sections contributed the highest share, that is, three-fourth (75%) of the observed amount at the expense of the other sections.

Table 4 can be more elaborated as: the computed chi-square result is 56.86. To answer the basic question, the degree of freedom from the contingency table (6-1) (2-1) is found to be 5. The critical χ^2 value with 5 DF at α= 0.05 level of significance is 11.070 or the calculated χ^2 P-value is less than the estimated P-value (P<0.0001). By conventional criteria, this difference is considered to be extremely significant statistically. The p-value answers this question: if the theory that generated the expected values were correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values? A small p-value or a large calculated χ^2 value is evidence that the data are not sampled from the distribution that was expected. This leads us to the conclusion that there is a significant difference between the observed and expected contents of the test items. Therefore, the content of the EGSEC English examination was not adequately representative sample of the contents of the textbooks. All the above reasons lead us to the conclusion that all the from 2009-2012 EGSEC English examinations lack content validity.

Generally speaking, the chi-squared test of independence tested the statistical relationship of the EGSEC English examinations administered from 2009-2012 in accordance with the weight given in the syllabi. The result of the chi-squared shows that four of the SATs of EGSEC English examinations were statistically significant as compare with the contents of the textbooks. This implies all the contents of the exams were not adequately representative sample of the contents of the textbooks. Thus, four of the exam papers lack content validity.

4.2. Strength of Association between the Textbooks and Sample SATs

The chi-square test of independence used in this study tells us whether two nominal categorical variables are statistically significant (independent) or statistically not significant (dependent). It does not tell us how strong that relationship or association is. When we produce a significant chi-square (this time the two variables are related), it is natural to wonder how strong the relation/association is. Chi-square test of independence is not complete without other measures of association such as Pearson's coefficient of contingency (C), Phi coefficient (Φ), Siegel's Contingency Coefficient (C), Cramer's Contingency Coefficient (V), etc. for the strength of association for nominal categorical variables. To know the strength of association between contents of the textbooks and sample SATs of this study, Cramer's V was preferred measure among the χ^2 based measures of association because it is useful for comparing multiple χ^2 test statistics and is generalizable across contingency tables of varying sizes or independent of the size of the variables. This measure is defined as*:*

$$V = \sqrt{\frac{\chi 2}{n(q-1)}}$$

Where *q,* is the smaller of the number of rows or the number of columns. The smaller the two numbers is used to represent the variable *q.* If *r* is number of rows and *c* is the number of columns, then

q=minimum (r-1, c-1)

Where: *V*= Cramer's contingency coefficient
n = grand total (the sum of all row totals or simply grand sample size)

Cramer's V always takes value in the interval [0.00, 1.00]. The coefficient ranges from 0 (no association) to 1(perfect association). In practice the close V is to 0, the weaker the association between the categorical variables. On the other hand, V being close to 1 is an indication of strong association between the variables. If the variables have a good fit, then V=1 (perfect association). Whereas if Cramer's V is equals "0" (V=0) when there is no relationship between the observed and expected variables. Generally, tables which have a larger value for Cramer's V (V>0.5) can be considered to have a stronger relationship between the two variables, with a smaller value for V (V<0.5) indicating a weaker relationship (weak association) between the two observations (contents of the textbooks and sample SATs of EGESC English examinations).

Like Underhill (1991), Alderson, Claphan and Wall (1996) note that to determine whether a certain test is valid in terms of content, the first step is to categorize the syllabus objectives into major content areas. The second step is to determine the number of period allotment or frequency of practice items in each content area of the textbook and the frequency of test items in each content areas of the test paper. Hence, the sample SATs in which this study focuses was classified into content areas.

The horizontal cells of each row show major content areas of tests, whereas the vertical columns contain number of questions of each section in the SATs by test year. The analysis of the distribution of different content areas of the samples test papers is summarized in hard figures and percentages (Please See table 5 below).

4.2.1. Total Frequencies of SAT Items versus Content Area

The analysis of the distribution of different content areas of the sample standardised achievement test papers in relation to their contents of the sections is summarized in figures and percentages in the following table. Its purpose is to obtain the sum of three years' test items, and in turn, this sum is helpful to obtain the grand total.

Table 5. Summary of total frequencies of items in various content areas in sample test papers by content area and test year (2009, 2010 and 2011).

Content area of sample SATs	Frequency by test year and content areas							
	Total test items in 2009		Total test items in 2010		Total test items in 2011		Total test items	
	figure	%	Figure	%	figure	%	figure	%
Reading	23	30.7	29	38.7	27	33.75	79	34.3
Speaking	13	17.3	9	12	14	17.5	36	15.6
Listening	0	0	0	0	0	0	0	0
Writing	7	9.3	6	8	6	7.5	19	8.3
Grammar	23	33.7	26	34.7	28	35	77	33.5
Vocabulary	9	12	5	6.7	5	6.25	19	8.3
Total	75	100	75	100	80	100	230	100

As can be seen from the above table, the sum of test items of each section of the three years along with its percentage is determined. It has been found to be 230. This figure will serve later as total frequencies of items in the test papers

4.2.2. Test content areas versus textbooks' content areas

So far, the two sets of important data have been obtained by analyzing contents of the textbooks (1996 editions) and sample test papers (2009-2011) in terms of frequencies of periods. The extent of relationship between the two observations can be computed using data from the textbooks and syllabi (For detailed information, see Appendix Table 1 of the average column). In the same way, data from sampled test papers (former three exams) analysis are presented in table 5 (See table 5 above). The total frequencies of items for both observations are put in figures and percentages in table 6 below.

Table 6 Total periods of content areas of textbooks (old) and frequencies of SATs (2009-2011)

Content area	Frequencies of periods in textbooks		Frequencies of items in test papers		Total frequencies -
	Figure	%	Figure	%	
Reading	28	18.115	79	34.3	107
Speaking	28.5	18.735	36	15.6	64.5
Listening	21	13.140	0	0	21
Writing	26.5	17.235	19	8.3	45.5
Grammar	27	17.530	77	33.5	104
Vocabulary	24	15.235	19	8.3	43
Total	155	100	230	100	385

Grand total

4.2.3. Degree of relationship between the textbooks versus sample SATs

Table 7. Degree of relationship between the textbooks and sample test papers.

	Content area	Observed(Pds)	Expected values	(O-E)^2/E
	Reading	28	$\frac{(155x107)}{385} = 43.08$	$\frac{(28 - 43.08)^2}{43.08} = 5.29$
	Speaking	28.5	$\frac{(155x64.5)}{385} = 25.97$	$\frac{(28.5 - 25.97)^2}{25.97} = 0.25$
Periods in textbooks	Listening	21	$\frac{(155x21)}{385} = 8.45$	$\frac{(21 - 8.46)^2}{8.46} = 18.59$
	Writing	26.5	$\frac{(155x45.5)}{385} = 18.32$	$\frac{(26 - 18.32)^2}{18.32} = 3.65$
		27	$\frac{(155x104)}{385} = 41.87$	$\frac{(27 - 41.87)^2}{41.87} = 5.28$
	Vocabulary	24	$\frac{(155x43)}{385} = 17.31$	$\frac{(24 - 17.31)^2}{17.31} = 2.59$
Frequencies in test papers	Reading	79	$\frac{(230x107)}{385} = 63.92$	$\frac{(79 - 63.92)^2}{63.92} = 3.56$
	Speaking	36	$\frac{(230x64.5)}{385} = 38.53$	$\frac{(36 - 38.53)^2}{38.53} = 0.17$
	Listening	0	$\frac{(230x21)}{385} = 12.55$	$\frac{(0 - 12.55)^2}{12.55} = 12.55$
	Writing	19	$\frac{(230x45.5)}{385} = 27.18$	$\frac{(19 - 27.18)^2}{27.18} = 2.46$
	Grammar	77	$\frac{(230x104)}{385} = 62.13$	$\frac{(77 - 62.13)^2}{62.13} = 3.56$
	Vocabulary	19	$\frac{(230x43)}{385} = 25.69$	$\frac{(19 - 25.69)^2}{25.69} = 1.74$
	Grand Total	385	385	$\sum \frac{(O - E)}{E} = 59.69$

The contingency table of this study is 2-by-6 (2 columns and 6 rows), q =2, which is the smaller of the matrix. As table 7 shows, the χ^2 result is determined. It has been found to be 59.69. As a result, the strength of relationship between the contents of the textbooks (old versions) and sample SATs (2009-2011) has been computed in the following way.

Applying Cramer's coefficient of contingency $V = \sqrt{\frac{\chi^2}{n(q-1)}}$, where n=385 (grand sample size , see in table 6 above), q=2 and χ^2=59.69 (See table 7 above), substituting we get

$$V = \sqrt{\frac{59.69}{385(2-1)}}$$

$$V = \sqrt{\frac{59.69}{385}}$$

$$V = \sqrt{0.155038961}$$

$$V = 0.393749871$$

$$V \approx 0.39$$

By using Cramer's statistical contingency coefficient formula, the extent of relationship or strength of association between the contents of the official textbooks (1996 versions) and sample SAT papers of (2009, 2010 and 2011) has been determined. In other words, the question "do the tests of EGSEC English examinations of the stated years reflect the required strength or association with the textbooks' coverage?" is answered. The result of Cramer's contingency coefficient value V is 0.39.

According to Cramer's contingency coefficient V, two observations are said to have a perfect relationship if coefficient value is 1 and if V>0.5<1, it indicates the presence of high or strong association or strong relationship whereas if coefficient value reads V<0.5>0, it indicates the presence of weak association. The obtained value of 0.39 in this case indicates the presence of weak association between the contents of the textbooks and sample test papers. As we can see in appendix table 1, speaking has the biggest share (28 periods) in the textbooks, however, in the 2009 sample test paper (See Table 1 above), it was given second position (13 items), and grammar which is third dominant in the textbooks was given first grade equal with reading. Generally, in table 7, it was found that the content coverage of the 1996 versions - *English for Ethiopia*-was not properly sampled in the 2009, 2010 and 2011 EGSEC English examinations. They have weak association and hence, they lack content validity.

4.2.4. Précis of total frequencies of test items of 2012

From September 2011 onwards, all Ethiopian students of grade 9 and 10 have been learning the new 2010 editions- English for Ethiopia. Hence, the 2012 EGSEC English examination was sampled from these new textbooks. So, it was the researcher's responsibility to assess the content coverage of these textbooks and their syllabi too. The main changes to the previous English syllabi are: the content has been reduced in order to cover the content in the allotted time.

The analysis of the distribution of different content areas of the sample standardised achievement test papers in relation to the contents of its sections is summarized in figures and percentages in the following table. The procedures and significance is similar with that of 4.2.1 above

Table 8. Summary of total frequencies of test items in the various content areas in sample test paper of 2012.

Content area	Frequencies	
	Total test items of 2012	
	in figure	in %
Reading	33	41.24
Speaking	6	7.5

Listening	0	0
Writing	7	8.75
Grammar	27	33.75
Vocabulary	7	8.75
Total	80	100

4.2.5. Total frequencies of 2010 textbooks editions versus EGSEC of 2012

So far, the content coverage (frequencies) of the major language items in practice tasks of grades 9 and 10 (editions of 2010, see appendix table 2) and summary of total frequencies of test items in various content areas in the sample test paper by content area (of 2012, see table 8 above) were determined. The extent of relationship or strength of association can be computed. The total frequencies of items for both observations were merged and put in figures and percentages in the table below.

Table 9 Total frequencies of content areas of 2010 editions of English for Ethiopia and sample test papers of EGSEC English examination (2012)

Content area	Frequencies in textbooks		Frequencies of test items in test paper		Total frequencies	
	Figure	%	Figure	%	Figure	%
Reading	54.5	17	33	41.25	87.5	21.9
Speaking	80.5	25	6	7.5	86.5	21.6
Listening	31	10	0	0	31	7.8
Writing	54	17	7	8.75	61	15.3
Grammar	59	18.5	27	33.75	86	21.5
Vocabulary	40.5	12.5	7	8.75	47.5	11.9
Grand Total	319.5	100	80	100	399.5	100

Grand total (grand sample size)

To determine the strength of association between the contents of the 2010 editions of the textbooks and the 2012 EGSEC English examination sample test paper, as in above, Cramer's contingency coefficient, a better alternative to the other measurements of association in a contingency table is used. It ranges from 0.00 to +1.00 and it is advantageous as it is independent of the size of the table. The procedures of obtaining the χ^2 of table 10 are similar to that of table 7 above.

4.2.6. Degree of relationship between the textbook (2010 version) versus sample EGSEC tests (2012)

Table 10 Degree of relationship between the 2010 editions and 2012 sample test paper

	Contents	Observed	Expected values	$(O\text{-}E)^2/E$
Frequencies in textbooks	Reading	54.5	$\frac{(319.5x87.5)}{399.5} = 69.98$	$\frac{(54-69.98)^2}{69.98} = 3.42$
	Speaking	80.5	$\frac{(319.5x86.5)}{399.5} = 69.18$	$\frac{(80.5-69.18)^2}{69.18} = 1.85$
	Listening	31	$\frac{(319.5x31)}{399.5} = 24.79$	$\frac{(31-24.79)^2}{24.79} = 1.56$
	Writing	54	$\frac{(319.5x61)}{399.5} = 48.79$	$\frac{(24-48.79)^2}{48.79} = 0.56$
	Grammar	59	$\frac{(319.5x86)}{399.5} = 68.78$	$\frac{(59-68.78)^2}{68.78} = 1.39$
	Vocabulary	40.5	$\frac{(319.5x47.5)}{399.5} = 37.99$	$\frac{(40-37.99)^2}{37.99} = 0.11$
Frequencies in test paper	Reading	33	$\frac{(80x87.5)}{399.5} = 17.52$	$\frac{(33-17.52)^2}{17.52} = 13.68$
	Speaking	6	$\frac{(886.50x)}{399.5} = 17.32$	$\frac{(6-17.32)^2}{17.32} = 7.40$
	Listening	0	$\frac{(80x31)}{399.5} = 6.21$	$\frac{(0-6.21)^2}{6.21} = 6.21$
	Writing	7	$\frac{(80x61)}{399.5} = 12.22$	$\frac{(7-12.22)^2}{12.22} = 2.23$
	Grammar	27	$\frac{(80x86)}{399.5} = 17.22$	$\frac{(27-17.22)^2}{17.22} = 5.55$
	Vocabulary	7	$\frac{(80x47.5)}{399.5} = 9.51$	$\frac{(7-9.51)^2}{9.51} = 0.66$
	Total	80	80	35.73
	Grand total	399.5	399.5	$\sum \frac{(O\_E)}{E} = 44.62$

Now it is possible to calculate the relationship or strength of association between the textbooks (editions of 2010) and the 2012 test paper.

Applying Cramer's contingency coefficient statistical formula, $V = \sqrt{\frac{\chi^2}{n(q-1)}}$

Where, χ^2 = chi-square test

n = grand sample size

q= minimum (r-1, c-1)

$$\mathrm{V} = \sqrt{\frac{44.62}{399.5(2-1)}}$$

$$\mathrm{V} = \sqrt{\frac{44.62}{399.5}}$$

$$\mathrm{V} = \sqrt{0.1116896120150188}$$

$$\mathrm{V}= 0.3341999581313839$$

$$\mathrm{V}\approx 0.33$$

By using Cramer's contingency coefficient statistical formula, the strength of association between the contents of the 2010 editions [New textbooks of *English for Ethiopia*] and the EGSEC English sample test paper of 2012 has been determined. It has found to be 0.33.

The calculated value 0.33 in this case indicates the presence of low association between the contents of the 2010 editions and the 2012 test paper. What does "weak" association mean"? In this case weak association means this sample test paper did not adequately represent the practice items or task formats in the textbooks. In other words, there was a sampling bias in the test paper, meaning there was a significant deviation between the empirical data (values) and expected (fit) values or frequencies. The "bad-fit" can also be easily observed from table 9 that speaking and grammar, both constitute 25% and 18.5% respectively in the textbooks whereas the test paper was dominated by 33 (41.25%) items of reading comprehension questions followed by 27 (33.75%) of grammar items which is incongruent with what was being expected.

The composition in terms of sub-skills of reading questions of the 2012 EGSEC English examination, it was dominated by macro-skills of reading-scanning a text to locate specific information and skimming text to obtain the gist - both constituted 23 (28.75%) questions. This data indicates less emphasis was given to the micro-skills - identifying reference of pronouns, using context to guess meanings of unfamiliar words- both of them comprise 10 (12.5%) questions at the expense of others. In addition to this, questions that demand students' higher order thinking of 'identifying stages of an argument', 'identifying examples presented in support of an argument', and 'understanding relation between parts of a text by recognizing indicators in discourse, especially for introduction, development, transition and conclusion of ideas' were not constructively treated (For detailed sub-skills analysis, see 4.3 below).

The above unfair treatment of sub-skills of reading indicates the domination of macro over micro-skills. This clearly indicates that there was not fair distribution of questions. Teachers' response of testing macro-skills of reading, of which 12 (80%) of them confirmed that their reading tests are more of scanning and 11 (73.3%) for skimming at the expense of other sub-skills. In response to this Tilahun and Triwork (2006) stated that to say a test has content validity, it should answer the question "Are the items in a given test adequate representative sample of a universe of content in a subject course?" Thorough observation and deep analysis of the SATs were made and are more elaborated in following section.

4.3. Analysis of the Standardised Achievement Test Papers by Sub-Skills

For the detailed analysis of the exam papers, the major content areas of the textbooks and sample test papers were sub-categorized into their sub-skills to make them specific as part of addressing the first and second research questions stated in the first unit of this study. Some of the major skills were categorized as macro and micro-skills whereas the language areas were split into their simpler components. Splitting the skills and language areas into their sub-skills and areas, helps in creating tasks which will elicit a representative sample of each. In the researcher's view, the greater the detail in the specification of content, the more valid it is likely to be. All of the exam papers were assessed, figured and analysed in this way hereafter (Please see Appendix Table 7).

4.3.1. Testing the components of vocabulary

When the treatment of vocabulary test items of 2009, 2010, 2011 and 2012 were observed, only synonym has got its representation in four of the test papers at the expense of word formation, homonyms, hyponyms, derivatives, dictionary use, antonyms, definitions and gap filling. Though the representation of synonyms 9(12%), 5(6.67%), 5(6.25), and 7(8.75), respectively across the test years was fair, the other sub-skills that have significant role in the textbooks entitled "Increase Your Word Power" were ignored in the exam papers. This orients students to give less attention on vocabulary skills resulting into weak competence in vocabulary.

4.3.2. Testing the sub-skills of writing

As far as the concept of content validity of writing items is concerned, writing test items must at least get a fair representation even in the indirect way of testing it. Every year's writing test items were expected to represent from13 (17.3%) to 14 (17.5%) questions in each test paper, but these figures were in contrary with what was being observed. They were rather reduced by half. In terms of sub-skill analysis, testing students' knowledge of mechanics of English [3(3, 75%) to 4(5%) questions every year)] takes the highest class in all the test papers. Narration [1(1.25%) to 2(2.5% or 2.67%)], description [2(2.5% or 2.67%)], letter writing [2(2.67%) only in 2009] have got their representation whereas report writing and note-making, which are the common tasks or practice items in the textbooks were totally neglected/ignored.

4.3.3. Testing grammar components

The third language area is grammar test item. Many, if not most, language testing handbooks encourage the testing of grammar by means of multiple choice items, often to the exclusion of just about any other method. But there are other techniques too to represent grammar in any of the following: paraphrasing, completion, and modified cloze with the imagination these would prove sufficient for most grammar testing purposes. They require the student to supply grammatical structure appropriately and not simply to recognize their correct use. Let the researcher see the test papers in terms of the representativeness of these sub-skills of grammar, if any.

Paraphrasing requires students to write a sentence equivalent in meaning to one that is given. It is helpful to give part of the paraphrase in order to restrict the students to the grammatical structure being tested. Completion testing technique can be used to test a variety of structures. Both of them did not get any representation at all in all the exams. The exams focus on modified cloze (testing prepositions, articles, a varieties of grammatical structures) and sentence linking other than paraphrasing and completion as some of the objectives of paraphrasing and completion are making students to restate simple sentences in their own words without distorting the substance of the original sentence and supply part of the fore- and back-part of the phrase or clause of a sentence respectively to make it complete and meaningful.

Among the represented sub-skills, testing prepositions and articles (almost none) shared the smallest figure. In 2009 only 17 (22.67%) grammatical structures and 6 (8%) of sentence linking, in 2010 similar format with different figures was repeated that is 20 (26.67) grammatical structures, 3 (4%) sentence linking and 1 (1.33) preposition. In 2011 the representation is similar in layout to that of 2009, which is 22 (27.5%) grammatical structure and 6 (7.5%) sentence linking whereas in the 2012 exam paper articles and prepositions have got a marginal representation as compared to the other testing years.

Generally, when we observe the representation of grammar test items in four of the test years, it is almost two times higher than what is rightly expected to appear in the tests. The reason why testers tend to emphasise on grammar is the ease with which large number of items can be administered and scored with a short period of time.

4.3.4. Testing micro and macro-skills of reading

Reading skill was also assessed in terms of its sub-skills. The simple statistical data shows that due attention was given to macro-skills of reading- scanning a text to locate specific information and skimming text to obtain the gist or theme. Other macro-skills of reading such as 'identifying stages of an argument 'and' identifying examples presented in support of an argument, which are among the main objectives of the syllabi, were left untreated in all the test papers. Micro-skills have got very little representation when compared with macro-skills. Among the micro-skills, identifying reference of pronouns and using context to guess meanings of unfamiliar words (deducing) were favourably treated at the expense of understanding relations between parts of a text by recognizing indicators in discourse, especially for introduction, development, transition and conclusion of ideas.

In almost all the test papers, in average, reading skill questions were prepared two times greater than expected. Out of which, in 2009, scanning 7 (9.33%) and skimming 11 (14.67); in 2010, scanning 10 (13.33%) and skimming 11 (14.66%); in 2011, scanning 13 (16.25%) and skimming 7 (8.75%); and in 2012 scanning 12 (15%) and skimming 11 (13.75%) constitutes the highest class. This distribution foretells students to concentrate upon these sub-skill and teachers to tighten their teaching only on scanning and skimming and ignoring the other as if they are not or less important for the exam and/or are not essential in the students' real life situations.

4.3.5. Testing the sub-skills of speaking

Speaking section was represented in a dialogue paper and pencil format only. In this test format, the tester hardly measures students' ability of expressing their own ideas, or whatever they are asked to respond orally, narrate events, or report whatever they know or read, as the best way of testing speaking is by making students speak.

4.3.6. Testing the sub-skills of listening

Listening skill which has a fair representation in the textbooks can be figuratively expressed as "as dead as a dodo". Since EGSEC English examinations became operational, it was totally ignored in the standardized achievement testing system of the General Education Quality Assurance and Examinations Agency (GEQAEA). Not testing listening skill in TMTs and SATs is direct information to students that this skill is either less important or not essential at all. This

Therefore, the progress of students' listening communicative competence may weaken and finally unable to use it the target language use.

4.3.7. Testing the components of morphology

In line with the sub-skill analysis of the contents of the test papers, the following language areas were also assessed. Morphology, the study of how words are put together, developed as a sub-field of linguistics. English speakers tend to think of 'words' as the building blocks of sentences and of sentences as strings of words. In this research, if a word has to be defined, it might first think of it as a unit in the writing system, the so-called orthographic word/s.

As the concept of the linguistic term, morphology is broad enough and complex, its usage in this study is confined only to the concepts of prefixes and suffixes- which are common word formation systems in the student's textbooks are addressed in the EGSEC English examinations. Under suffixes, grammatical categories such as plurals (workers), person (works), tense (worked) or case (John's) and prefixes were assessed.

The findings of the analysis shows that the test items that draw attention of students' knowledge of prefixes were not observed at all. Few items that requests students' background knowledge of suffixes in the form of tense and person were observed. In 2009, 2010, 2011 and 2012 test years, 4%, 2.67%, 0% and 11.25% representations of suffixes were observed respectively. Plurals and case suffixes were not detected. The absence of prefix test items implies a negative impact in the teaching and learning process as students would get confused about their appearance in such exams, perhaps this area may not get a reasonable attention by the students at class or when they study out of the class.

4.3.8. Testing lexical structures

Lexical structures that denote links between words which carry meaning (nouns, verbs, adverbs, adjectives, etc.) in the textbooks were treated in various sections of the test contents. Under this section reiteration particularly, synonyms, antonyms, homonyms and hyponyms and collocation (a way in which certain words occur together) were assessed. Under reiteration part, only synonyms have got their representation under vocabulary test items with 5 (6.7%), 9 (12%), 5 (6.25%) and 7 (8.75%) in 2009, 2010, 1011 and 2012, respectively.

The other parts of reiteration and collocations which are the nuts and bolts of lexical structures were totally forgotten. If such lexical structures are not incorporated in standardised achievement tests, these items will receive little attention by the students when they are practising in the tasks at class and in their daily communicative activities. Students want to give emphasis on what appears in the exam. In view of this idea, Nitko (1996:39) claims that students expect exams to appear from what has been emphasized in the class. For this reason, the test constructors might not able to see the washback effect and hence their exams lack content validity.

4.3.9. Testing syntactic structures

The representation of syntactic structures was also part of the textbooks and, therefore, assessed. Syntax is the study of the way in which phrases, clauses and sentences are structured out of words, that is how lexical items or words in the language form a given sentence out of these set of words. The chosen words are then combined together by a series of syntactic computations in the syntax, thereby, forming a syntactic structure. This concept is expected to answer the first research question: *To what extent the syntactic structures stated in the syllabi are addressed in the EGSEC English examinations?* Four of the exam papers were thoroughly inspected and discovered the following results.

Except the year 2009, test items of 2010, 2011, and 2012 exams focussed on question like "choose *the best arrangement of the words to make a complete sentence"* and the statistical result was appreciable or satisfactory with 3 (4%) questions every year. The washback effect of these test items was found to be positive as students give due attention to concord in their textbooks and class hour.

4.3.10. Testing semantics

The overall objective of the textbooks is to make students have the basic skills of communication and develop their communicative competence so as to solve immediate problems associated with the language. Students need to understand the meaning of words, grammatical semantics and how morphemes' meanings are combined by grammar to form the meaning of utterances. This generally refers to semantics - the study of meanings and its manifestation in language. Except 4 (5%) questions only in 2012 test paper, no observations were found in the tests before. This indicates how much test makers have forgotten it and now it seems to be reviving and is able to construct positive washback effect in the future language teaching and learning.

4.3.11. Testing the components of phonological structures

Regarding the phonological structures of English the summary of grade 9 and 10 syllabi introduces them as "students are able to repeat rhythms that reinforce English sounds, stress and intonation". Accordingly, these items are in the textbooks and students have practiced them at class. More than two syllabic (grouping of sounds for the purpose of articulation) words carry the meanings of that word. Some words have only one syllable like: yes, no, town, etc. whereas others have two or more (poly) syllables (for example, pro-nu-ci-a-tion). Words with more than one syllable always have one strong syllable which is stressed. Many words are stressed on the first syllable, but not all. These and the following examples were taken from the students' book. For example, "SYLL-a-ble" and "ex-AM-ple", etc. Word stress is a common practice task in both of the textbooks, but it did not get any representative sample in the EGSEC English exams of the stated years.

Intonation, which is the rising and falling tone of sentences, is also a common task in the textbooks. For example, the following extract was taken from English for grade 9 task A1.10 page 9 of edition 2010 stated to clarify the intonation.
Rising: May I borrow your pen?
Falling: Yes, of course you can.

The case of length of words, that shows the duration taken to produce sounds (syllables) are also common exercises among regular past endings /t/, /d/, /id/ and pronunciation of words with /ei/ and /ie/ in the textbooks. This information was sourced from English for grade 9 pages 35 and 76 of tasks B2.11 and B4.7 respectively just as it is handy to let one know their representation.

The goal of the pronunciation practice then is to get every student to pronounce the target word or utter sentence as accurately as possible or the reverse process is also possible, the teacher says the word or sentence and the students write it (dictation). Or sometimes drilling or choral repetitions in plenary is also recommended. This will get their ear sensitized to the natural sounds of English and has the advantage that they can use their stronger skill of writing, so the students will not be entirely at sea

To recap, phonological structures are part of the syllabi that the learners are expected to accomplish as part of their classroom tasks. Students have spent several periods to practice those tasks, but there were not any observable items in four of the exam papers. The treatment of these tasks would be possible at least through paper and pencil task formats as used in the classrooms, but in reality phonological structures were ignored in the testing system of the EGSEC English exams as listening do. As it was understood from the school teachers' tests, the majority of them do not include these items in their tests. The implication of neglecting such important components of the language in the testing system of EGSEC English exams makes teachers and students give less attention while teaching and studying, respectively. Therefore, this results in harmful washback effect because SATs should fairly represent to all content areas in a syllabi or textbooks.

4.4 Teachers' and Testing Experts' Reflections on the Testing Practice

This section is being presented as a response to the third research question. Hence, the core concern of this section is to reflect on awareness and practice of testing valid English language tests, views of teachers and testing experts on the content validity of the EGSEC English exams, and the effect of test invalidity on the quality of English language teaching and learning.

4.4.1. Analysis of teachers' views on content validity of EGSEC English examinations.

Every year when the time for national exams comes around, a debate ranges among teachers community whether the national examination has fair distribution across the contents of the textbooks they have taught in the class. Most of the teachers cast doubts over the fairness of these exams when they are used nationwide. Their dissatisfaction can be seen because there is an imbalance between what they really teach and what reflects in the national exams. For example, they radiate their hesitation and ambiguity with their cautious language as "we teach and test several components of the language such as word formation, antonyms, hyponyms, homonyms, etc. but the EGSEC English examinations focus on testing synonyms only. As can be understood from this statement, the test formats of the EGSEC by themselves pressurize school teachers to accustom it.

All the respondents shared the idea that the exams did not encompass all the four skills. From these, as said above, listening skill was totally ignored in the testing system. Even the other skills and language areas were treated with little considerations. This indicates the items of the exams were not representative of the syllabi. The interviewees also reflected that the EGSEC English examinations were not communicative language tests as compared to the syllabi because these exams missed several most important components of the textbooks. For this reason, all the exam papers were haphazardly prepared without taking into account the weight given in the syllabi or allotment of periods and frequency of practice tasks.

All of the respondents strictly criticized the layout of the papers. The exam papers were distributed across the nation with four booklets (or codes) so as to protect cheating. The EGSEC English language test constructors might not aware of the shock on the behaviour of the test takers' when the codes with the difficult test items come first. In this case, students would be victims of coding as some of the codes with difficult items come at the outset and make students frustrated and nervous whereas the students with the easier items at the beginning of the paper would perform without any fretfulness. Reshuffling of the test items in different booklets could reduce the standard of content validity of the exams.

Such a method of testing contradicts with the objective of the syllabi that says, "There is spiral progression through the four skills, grammatical and vocabulary items and other language components are taught at increasing level of difficulty and sophistication within the topic area". This should be in harmony with the idea that test items too should proceed from simple to complex, but it was totally ignored in the testing system of the agency. Such factors minimize the content validity of the SATs.

The respondents were also asked for their opinion if the EGSEC English exams were the representative sample of the content coverage of the textbooks. Most of them were expressing their doubts on the representativeness as those exams are more favourable to certain parts of the contents- grammar, reading, and vocabular3y- and give little attention to others-speaking, writing and no attention at all to listening. As the researcher has been understood from the cautious language, they were also in doubts in generalising the content validity standard of the exams. This way of representing test items indicates the presence of unbalanced or "test bias" (Bachman, 1990). In other words, lack of content validity affects students' attitude negatively toward learning and studying of contents of the textbooks. If the tester becomes impartial to certain parts of his/her favour, and partial to the other portion, the students also become attentive to those parts which frequently appear in the exam and neglect the other parts regardless of their usefulness.

4.4.2. Results and analysis of teachers' awareness and practice of testing content valid tests

In this part, to strengthen the statistical result obtained from the chi-square test of independence and Cramer's coefficient of contingency, the raw data obtained from the questionnaire and interview would also help in validating the research. To do so, each of the tools with 20 questions were distributed and asked among 15 respondents who have a direct relationship with teaching and testing. These questions were not intentionally designed to assess directly the content validity of the EGSEC English examinations but just to triangulate what teachers actually test at classroom level and what the GQAEA tests at higher level. Their response to all the questions in the questionnaire was put in figures and percentiles (Please see Appendix Table 1). These figures were compared with the figures of the agency's test (Please see Appendix Table 7). This was handy to let someone know what teachers actually teach and test in their classroom and contents of EGSEC English exams so as to draw a conclusion on the EGSEC English exam content validity.

The first question for the respondents in the questionnaire was: *Do they test vocabulary test items such as word formation, grammatical categories, pronunciation, dictionary use, antonym and synonyms?* For this question, the respondents responded that 100%, 100%, 53.3%, 33.3%, 100% and 100%, respectively. From their response it is possible to infer that the majority of the respondents test these items at classroom level but in the sample SATs of the EGSEC English exams, synonyms was only treated as a vocabulary test item. This indicates there was a mismatch between what teachers teach and test at class and what the EGSEC English exam tests. In view of the above response the interviewees also confirmed 'the major contents of their test' incorporates grammar, reading and vocabulary to a great extent. Their focus on testing speaking and writing was very little because of large class and time constraint to subjective marking.

To be sure of what the teachers test in reality, a first semester regional final test of 2012 (they called it standardized achievement test) set by a joint relationship between these people was observed. In that exam, no focus was given on dictionary use and pronunciation test items. The others were tolerable as the exam was a mid-course test.

In the second question in the questionnaire on the degree of testing the above test items, varieties of responses were observed. The different responses could be as a result of difference in their teaching experience in different schools. For this reason, they test word formation (66.7%), part of speech (46.6%), antonym, synonyms and gap filling each 40% every semester. Derivatives and dictionary use seem as they were less tested. 60% and 33.3% of the respondents rarely test dictionary use and derivatives respectively. Derivatives, which are used to form varieties of words and dictionary usage, are common practice items, especially in the new editions of the textbooks. But they get little attention in the TMTs and SATs which is in contrary with the meaning of content validity.

With regard to the third question in the questionnaire on the degree of testing the sub-skills of writing, 80% of the respondents occasionally test letter writing and narration, almost 13.3% of them do not test and very few of them (6.7%) test almost in all their regular tests. In the writing sections of the textbooks, letter writing covers a wide range, but school teachers and EGSEC test makers give less attention to this productive skill in the exams. Except in 2009 test paper, no observation was made at any other time. Descriptive and simple report writing were tested almost in every exam by 20% and 26% of the respondents respectively but the majority of the respondents (80% for descriptive and 40% for report writing) test these test items sometimes. Summary writing (73.3%) and note-making (53.3%) also receives similar attention. The reason why they do not test these test items regularly or almost every time is due to subjective marking and time constraint as they have large classes. This may lead students to be dependent on objective type of test items where their spelling, hand writing, capitalization, punctuation, syntax, coherence, etc. are not assessed.

The respondents were asked the priority they give to test 'paraphrasing, completion and modified cloze'. The majority of the respondents (53.3%) give low priority for testing students' ability to restate simple sentences in their own words without losing the substance of the original sentence. The reason is similar to testing writing test items as it demands teachers to mark it subjectively and difficult to mark every students' papers in a limited time. 53.3% of the testers gave high priority for testing one or two word(s) completion. Paraphrasing and sentence completions were totally ignored in the EGSEC English examinations. Regarding testing prepositions 20% of them gave high priority; another 20% gave somewhat priority whereas 53% of the respondents give a moderate priority. These items which are common tasks in the textbooks were also given less attention by the SATs.

Grammatical structures and sentence linking were given high priority by the majority of the respondents (60% and 40%), respectively. Even the locally observed TMTs also confirmed that testing grammatical structures (40%) was given a high priority. This way of test preparation encourages students to focus on grammatical structures that are simply memorizing hard and fast rules giving surface meanings of sentences. Testing several grammatical structures usually restricts students' mastery of the language and its usage in real life.

In response to the fifth question about the frequency of testing the micro and macro-skills of expeditious reading, most of the informants’ response indicated that 80% and 73.3% of them frequently test scanning and skimming respectively whereas the rest of them test these items occasionally. This indicates the dominance of scanning and skimming over identifying stages of an argument and examples presented in support of an argument. There are even few respondents that do not test 'identifying stages of an argument and identifying examples presented in support of an argument' at all with 20% and 6.7% each.

Few of the respondents were able to put the frequency of testing reading items from occasionally to never, but response of the majority fits with the reading test preparation system of the EGSEC English examinations which is a misleading notion with the meaning of content validity given by many testers. Among the expeditious reading, examples presented in support of an argument and understanding relation between parts of a text by recognizing indicators in discourse were prepared by the teachers rarely and so do the SATs. Data from the interviewees also supports majority of them focus more on reading test items as their major section in their classroom tests. In line with this Hughes (2003:138) notes that the washback effect of tests like this is that many students have not been trained to read quickly and efficiently.

It may sound quite strange to test listening separately from speaking, since the two skills are typically exercised together in oral interaction. However, there are circumstances of recording resources, for example, listening to the radio on images of Ethiopia, listening to lectures, listening to a recorded audio CDs about the Ethiopian Airlines advertisements, etc. when no speaking is called for. This receptive skill, which is the ignored skill in the testing system of SATs, 60% of the testers of the school do not prepare listening test items that demand students to scan and skim, following instruction, and recognition of function of structures (such as interrogative as requests, for example, could you pass the salt?), etc.

Moreover, 53.3% of the respondents do not test listening test items that request students to listen and follow directions, take notes (usually lecture) and practice intonation patterns (recognition of sarcasm, etc). Very few respondents (13.7%) test listening test as quiz assessment only to encourage learners. As the syllabi orders in every listening lesson, teachers are expected to read the whole reading text or use pre-recorded cassettes. This time students will comprehend the relationship between sound and meaning, perceive changes in stress and intonation which signal meaning. Of course they do, but my dismay here goes to what extent are students able to understand stress and intonation patterns by listening to speakers of English as a second or foreign language. Having the pre-recorded cassettes/DVDs of native speakers' natural accent (with appropriate pronunciation and intonation or tone of speech) would be recommended, but teachers do not use this as a teaching resource in class at all. Thus, this is an indication of how students are suffering from identifying the intonation and stress pattern of the words or phrase and the message they signal.

The assumption made in this research is that testing oral ability of students helps them develop their ability to interact successfully in English. This involves comprehension as well as production. Tests of speaking should also be set from a representative sample of population of oral tasks that are expected candidates to be able to perform and elicit behaviour which truly represents the candidates' ability, and finally, the samples of behaviour can and will be scored validly and reliably.

The speaking operations that the syllabi introduces for students to achieve are dialogue (conversation between or among people); expressing (likes/dislikes, thanks, requirements, attitude, confirmation, apology, complaints, reasons, justifications, preferences, agreement or disagreement, opinions, etc); directing (instructing, persuading, advising, prioritising); describing (actions, events, objects, people, process); eliciting (information, direction, classification, help);

narration (sequence of events) and reporting (description, comment, decision, choice). Teachers and testing experts were asked the extent of testing these items as part of their classroom tests.

Majority of the respondents replied that they test these operations to a great extent. To mention in figure, 66.7%, 46.7%, 40% and 40% of them test dialogue, directing, eliciting and reporting dialogue respectively. Dialogue was represented as a speaking test in general at the expense of the other speaking test operations as the EGSEC English testers do. Few respondents were observed that they do not test speaking test except dialogue in a paper and pencil format in their tests. To substantiate, 13.3% of them do not test narrating, eliciting and directing and 26.7% of them do not even prepare oral test items that demand students to express and report what has been asked. The problems they have in testing speaking test items are due to large class, time constraint and subjective marking as of writing.

To sum up, the majority of the respondents give emphasis to the above operations, but exam of the agency only assesses dialogue through printed hard copies that restricts students' precision of the language (grammatical and lexical accuracy, appropriateness, range, flexibility and size) are go unheeded.

Language areas were also satisfactorily treated in the practice tasks of both grades of both editions. Stress, length and syllable at word level and intonation at sentence level were thoroughly discussed in the textbooks, especially in grade 9 (For more details see 4.3). As it can be seen from appendix Table 4, the response of the ninth question indicates that some respondents test these areas to some extent. Only 33.3% of them said that they include the above phonological structures of English. This practice adversely influences the students' attitude towards learning these language items appropriately.

Concerning the morphological and lexical structure of English, 73.3% of the respondents have been giving emphasis for testing prefixes and suffixes whereas 53.3% of them focus on testing the lexical structures (reiteration and collocation) always. The rest of them, 26% of the respondents emphasize on testing prefixes and suffixes whereas 46.7% of them lexical structures. This is in clear harmony or accord with the facts in the analysis of their test papers, only suffixes and synonyms. The worst situation goes on to testing collocation, hyponyms, homonyms and antonyms where nobody gave them due attention both in the TMTs and SATs.

The respondents were also asked to rate the degree of testing their students' knowledge of syntactic structures and semantics. 60% of them confirmed that they always test syntax and semantics. 40% of the respondents usually test these items. This data fits with what is in their test papers. Such data encourage learners to stress on these items when learning and the washback is positive as the tests of EGSEC also focus on these.

And finally, the respondents were requested to rate the type of test items they prefer to address the objectives of the textbooks. 100% of them rated to a great extent that their test items are more of multiple choice items. 33.3% of them also greatly emphasise on matching and gap filling. 53.3%, 46.7%, 26.7% and 20% were rated for matching, gap filling, true/false and essay writing, respectively that these test items are tested to somewhat. But in reality, all the items in their test paper were multiple choice items as the EGSEC does.

4.4.3. Views on the causes of content invalidity

In all the earlier findings of this study, both the TMTs and SATs of English examinations were found to be divergent from the contents of the textbooks. Some of the possible reasons collected from the respondents are analysed and discussed hereafter.

The teachers were asked if they are familiarized with the term *'content validity'*. As has been understood from their cautious language, their understanding on this term seems limited which led them not to use table of specifications and in turn led them miss unintentionally very important components of the language in the test. Such awareness has a direct influence on the behavioural domains of the students. The preparation of test items, excluding some part of the syllabi at classroom level would make students wait the EGSEC the same way. In this situation, we can understand that less emphasized language components in the test will obtain little attention by the students when practising and studying.

The school teachers and testing experts of the regional state do not have any relationship with NEAEA. The test makers in the NEAEA do not request regional subject area specialists and testing experts to contribute questions to the agency so as to have an item bank. The teachers' limited awareness and personal attitude affect their practice of designing their content valid classroom level tests and are unable to supply feedback to the concerned bodies in the quality assurance and examination agency about the content validity of the agency's tests. The information regarding what teachers do at school level is presented below.

In view of the respondents' response on the: *The major content of their test*, their test incorporates grammar, reading and vocabulary to a great extent. Their focus on testing speaking and writing was very little and even they do not test listening skill at all. Some of the reasons they cast against testing this skill are: the school is not well equipped with accessible materials, inappropriate testing environment and absence of self-access learning and testing centre. This indicates that phonological structures are not tested. On the other hand, the problems they had in testing writing and speaking test items are due to large class and time constraint to subjective marking.

All the school teachers do not use table of specifications (TOS) before their test construction. They simply prepare their exams by looking at the major contents of the textbooks ignoring the weight given to every content. Such habits of the school teachers indicate their test lacks proportionality. The misuse of the TOS led them to harmful washback effect and thus, low content validity. Supporting this view, Siddiek (2010) summarises the importance of TOS as "it must be emphasized that before starting to write any test item, the test constructor should set up a detailed TOS, showing aspects of all skills and language areas being tested and giving a comprehensive coverage of the specific language elements to be included. Thus, the practice of constructing tests without TOS shows their limited awareness about the concept of content validity and its application.

The absence of periodic long or short term training greatly hampered their professional development in such a way that most of their tests relied on grammar and usage, reading and vocabulary and even these tests have less content validity. From this concept, it is possible to infer that a failure in adequate assessment and appropriate evaluation of students' performance can by and large worsen education quality in general and language teaching and learning in particular. Supporting this view, Abraham (2008) traces the very limited or no opportunity of getting long and short term training which significantly limited the contemporary knowledge of teachers in the areas of quality test development in general and content validity of English tests in particular.

4.4.4. Views on the solutions of content invalidity

Concerning taking a short or long term training, all the respondents except the testing experts replied that they have never taken any short or long-term training except the course 'Educational Measurement and Evaluation' as part of their undergraduate course. The courses 'English Language Testing and Assessment', 'Fundamental considerations in Language Testing', and 'Designing and Analysing Language Tests' could have a great positive impact in preparing quality and content valid language tests for they have special attachment in technical matters of language tests preparation unlike the 'Educational Measurement and Evaluation' course does. The absence of availing the above recommended and other short- term courses perhaps resulted in low levels of awareness concerning content validity

In general, concerning the content validity of the EGSEC English examinations, it can be said that the Assessment and Examination Agency should give due attention and care to the content validity of these SATs exams before and after their construction and administration. Truly speaking, the Ethiopian Ministry of Education is working to a great extent in the decentralization of education under the themes "Education for All" and "No Child Left Behind"; therefore, emphasis should be given to testing too as the main concern in teaching.

Generally speaking, as shown in the analysis, the exams do not possess content validity and have even weak association/ strength with their respective textbooks as well. These resulted in the negative washback effect in students' English language learning. Moreover, the study showed that teachers' limited awareness concerning content validity heavily affects the quality of language teaching and learning.
To finalise the result and discussion, the researcher believes that the following Carr's (2011:19) statement in his recent book entitled 'Designing and Analyzing Language Tests' would summarize this research. He claims that without adequate sampling of the topics, situations, genres, rhetorical models, function, notations, structures and tasks that were covered in the class, it is not possible to claim, in fairness, that the test provides a clear picture of what students have or have not achieved!!

5. SUMMARY, CONCLUSION, AND RECOMMENDATION

This section presents the summary of the study, conclusions and the recommendations made based on the findings of the study.

5.1. Summary

The main aim of this study is to assess if the EGSEC English examinations administered from 2009-2012 possess content validity or not. To achieve this aim, three research questions were formulated. So the data required to answer these questions were collected using document analysis, questionnaire and unstructured interview from the contents of the textbooks, syllabi and SATs, and teachers of Harar Secondary School and testing experts of Harari Bureau of Education. The data collected using these instruments were analysed and discussed quantitatively and qualitatively based on the research objectives and literature review. Taking in to account the content coverage of the major headings of the textbooks and SATs, they were categorized in to six sections: reading, speaking, listening, writing, grammar and vocabulary in separate cells. For all these sections their respective period's allotment and frequency of practice tasks were inserted in each cell. From these cells, the expected numbers of questions were computed by using the coverage size of each item in the textbooks and total number of questions in the paper. Then the number of items in the test was observed, determined and put in a separate column. All the expected and observed data were put in the chi-square table and the $\chi 2$ value for each contingency table was determined using 5 degree of freedom and α=0.05 significance level. The findings of the χ^2 test of independence indicated that the relation between the contents of the textbooks and SATs were found to be deviating. Some content areas such as reading and grammar were given a high attention in the exams whereas the other major sections are either minimally treated or not treated in the exams. The strength of association between the contents of the textbooks and SATs was also determined using Cramer's V coefficient of contingency. Thus, the strength of association between the contents of the 1996 editions and 2009-2011 SATs, and contents of 2010 editions and 2012 SAT were found to be 0.39 and 0.33, respectively. These figures indicate the presence of weak strength between the contents of the textbooks and SATs. To know the representation of the language skills and areas in terms of their sub-skills, all major content areas were sub-divided into their respective sub-skills and language areas. The result of this analysis also indicated that some sub-skills and language areas were still dominant and others were given little attention. To triangulate the result obtained from the document analysis, teachers of Harar Secondary School and testing experts of Harari Bureau of Education were questioned and their responses were analysed as per the demands of the research objectives. Reflections of the respondents were found to be casting their doubts on the standard and content validity of these exams. Generally, the computed values of the χ^2 & V indicate that the SAT items were not adequately represented from the contents of the textbooks of the stated grades.

5.2. Conclusions

The researcher has been analysing and interpreting the data gathered using the aforementioned tools. As per the objectives of the syllabi the expected and observed number of questions should be identical, but the findings of the study revealed that majority of the contents of the textbooks and SATs of English exams were divergent. Based on the research study and summary, the researcher has arrived at the following conclusions.

1. In all the SAT papers, synonym has got a significant representation which is quite appreciable as it has a positive washback effect on the students' behaviour, but some prestigious components of vocabulary items were left not assessed. This might orient students to give less attention on vocabulary skills resulting into weak competence in vocabulary.

2. Syntactic structures were well represented in the later three tests. The statistical result was quite satisfactory as these items have consistency across the test papers of the EGSEC English examinations. The washback effect of these test items has found to be positive as students give due attention to concord in their textbooks and class hour. In line with this, semantics has also received a fair representation only in the 2012 test paper. This indicates how much it was forgotten and now it seems to be reviving and is able to construct positive washback effect in the future language teaching and learning.

3. The result of the document analysis indicates the sample SATs were not adequately sampled from the contents of the textbooks. The language areas and skills were not proportionally incorporated in four of the SATs. This means most language areas and skills are fairly treated in the textbooks but not proportionally represented in the SATs and even like listening go unheeded.

4. The response of the teachers and testing experts also confirmed that what they teach and test is not properly reflected in the EGSEC English SATs. Some important components of the language were less treated and others like phonological structures, paraphrasing, sentence completion, etc. were totally ignored in the testing system of the NAEAE.

5. The findings clearly show the violation of content validity of important language areas and skills in the exams. Such inappropriate inclusion of textbooks contents in the SATs seems as there is a biased testing practice of test makers and this may result in the biased habit of students' which deprives them optimal learning.

6. With regard to the NEAEA's SATs of English of the present study, due attention was not given towards preparing content valid standardized national exams. The content invalidity of the exams has been orienting students and teachers to rely on what appears most in the exam.

7. Overlooking of preparing well sampled SAT papers prove the poor quality of education which figured out the country for many years. Such biased testing system of the NEAEA proves the wrong perception we have of the role of tests to the teaching process. The backwash effect is affecting many secondary school students negatively. Tests that favour to certain contents of the textbooks in no way can develop learners' skills and communicative abilities in their study habits. This shows that the sample SATs remained weak at assessing the learner’s knowledge of language proficiency in each area of the syllabus.

8. The findings also confirmed that majority of teachers have low level of understanding with regard to content validity. This low level of understanding negatively influenced the teachers' attitude of language testing and concentrating only into the limited language test items, particularly grammar, reading and vocabulary sections by giving little attention to other areas and skills. Such kind of testing pressurise students to develop a habit of grammar oriented studying practice. Therefore, it can be concluded that four of the SATs have low content validity and weak strength. So, the discussion on content validity of EGSEC English examinations can be recapped as the test items were simply taken from certain area of contents regardless of the weight given in the syllabi.

5.3. Recommendations

- NEAEA or concerned body in the Ministry of Education needs to come forward to look into the stated issues and do the needful by conducting further studying on the standards of the exams in terms of their content validity, their impact on quality of language teaching and learning and able to rectify it.
-
- Since examinations -among other things - are tools of quality control through which professionals in the area of teaching can judge the success of their pedagogical and educational objectives, positive steps should be taken to foster quality of teaching and learning. Examinations can put both the teacher and the learner on the right track of the whole education process.
- Selected school teachers should address the problem and contribute questions to NEAEA, so that experts could have an item bank like what the school teachers and testing experts in Harar do.
- Attention should be drawn towards making a balance between textbook contents and test contents. For this, the respective bodies of examinations should be given in-service training to tackle the problem.

- Testing should be given due attention as of teaching because teaching without proper testing holds no value. Awareness creation training and workshops on assessment mechanisms in the areas of language testing should be given to both school teachers and the test makers continuously by highly specialized testing experts. So, these test makers should validate the tests before they administer them.
-
- Increase the number of items that ranges from 110 up 120 so that it would be possible to adequately mirror the objectives and contents in the syllabi. The implication of this concept is longer tests tend to be more valid and reliable than shorter tests.
-
- To keep exam papers consistent, reshuffling the contents of the exam should be avoided to minimize students' frustration when difficult test items are set up at the outset.

5.3.1. Implications regarding further studies on these examinations

In this study, the researcher tried to understand the low level of content validity of the EGSEC English exams. However, this study only focused on the assessment of the content validity. Besides, the research was also limited to four SATs and fails to see other SATs of EGSEC English examinations. Thus, the study recommends further researches can be conducted on the 'level of difficulty (item analysis)' and 'item discrimination' of the EGSEC English examinations

5.3.2. Implications for Teachers

1. For a teacher with a low level of understanding the concept of content validity, it is difficult for him/her to understand whether the overall textbooks objectives are achieved in his/her exam or not and remained unable to assess the students mastery of the four language skills and areas effectively.

2. The biased testing practice of teachers leads students into the biased learning and studying habits. Such biased testing practice of the teachers can potentially demean the quality of education in general and students’ global language ability in particular.

5.3.3. Implication for Test Makers

As shown in the study, low level of content validity of SATs of EGSEC English exams might play their role in negatively affecting the students’ communicative competence in that these exams made them more to focus on the dominant sections in the exam and ignore the other skills and language areas. Therefore, test and policy makers, administrative staff in the NEAEA and at various levels should exert their best level for creating conducive atmosphere for the preparation of quality tests in general and contently valid tests in particular as part of creating conducive atmosphere in schools for effective execution of the communicative language teaching and learning.

6. REFERENCES

Abebe Bekelle, 1986. "An Insight into the Concept of the Curriculum."*The Ethiopian Journal of Education, X (1):33-58*

Abraham Kebede, 2009. An Inquiry into Content validity of English Language Classroom Teacher Made Tests with particular reference to some selected secondary Schools in Wolayta Zone: Awareness, Practices and consequences, Haramaya University, (Unpublished MEd Thesis).

Adams, R.J., P.E. Griffin and L. Martin, 1987. A Latent Trait Method for Measuring a Dimension in Second Language Proficiency. *Language Testing*

Adugnaw Ayele and Desalegne Chalchisa, 1998. Teacher's opinion on the ESLCE and Content Validity Analysis of The 1998 grades 12 and 8 National examinations. NOE, Research, Addis Ababa.

Aggrawal,Y.A., 1998. Statistical Methods Concepts, Application and Computational. Third Revised and Enlarged Ed. Sterling Pub Priv Limited New Delhi-India.

Alemu Tsegaw, 1983. Assessment of Grades Six and Eight English National Examinations, MA Thesis. Addis Ababa: Addis Ababa University. (Unpublished).

Alderson, J.C., 1981. "Report of discussion on Communicative Language Testing", The British Council (ELT) Documents.

Alderson, J.C., 1981. "Report of on Communicative Language Testing", The British Council.

Alderson, J.C., C. Claphan and D. Wall, 1996. Language test construction and evaluation. Cambridge university press.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 1974. Standards for Educational and Psychological Tests. Washington DC: *American Psychological Association*.

Anastasi, A., 1976 & 1990. Psychological Testing. New York, Macmillan publishing company.

Atkins, J., B., Hailom and Nuru Mehammod, 1996. Skills Development Methodology. (Part 2). Addis Ababa: Addis Ababa University Press.

Bachman, L.F., 1990. Fundamental Consideration in Language Testing, Oxford, Oxford University Press.

Bachman, L.F., 1991. What does language testing have to offer? TESOL Quarterly, 25(4) University of California.

Bachman, L.F. and A.S. Palmer, 1996. Language Testing in Practice: Designing and Developing Useful Tests. Oxford University Press.

Bachman, L., 2000. Modern Language Testing at the Turn of the Century: Assuring that what We Count Counts. Language Testing.

Baker, D., 1989. Language Testing, A Critical Survey and Practical Guide, Edward Arnold

Benson, J.A., 1981. A Redefinition of Content Validity. Educational and Psychological Measurement. 41: 793-802.

Berhanu Moges, 2004. “Teachers Assessment of students Performance: Continuous Assessment” (Unpublished MA. Thesis) Addis Ababa University.

Brown, F.G., 1976. Principles of Educational and Psychological Testing (2nd Ed.). New York: Holt, Rinehart & Winston.

Brown, H., 1994. Principles of Language Learning And Teaching. San Francisco State University: Prentice Hall Regents.

Canale, M. and M. Sulain, 1980. Theoretical Cases of Communicative Approach to Second Language Teaching and Testing, In Applied Linguistics, Vol. 1.N0 .1. USA.

Carroll, B.J., 1981. Specifications for an English language Testing Service. In Alderson, J.C. and Hughes, A., editors, *Issues in language testing: ELT Document 111*. London: The British Council.

Cronbach, L.J., 1971. Test Validation. In R.L. Thorndike (Ed), *Educational Measurement (2nd Ed.).* Washington DC: American Council on Education.

Cureton, E.E., 1951. Validity. In E.E. Lindquist (Ed), Educational measurement. Washington DC: *American Council on Education.*

Davies, A., 1988."Communicative Language Testing." Testing English for University Study. Modern English Publications and The British Council.

Davies, A., 1990. Principles of Language Testing. Oxford: Basil Blackwell.

Desalegne Chalchisa, 2001. Testing and Assessment Skill, Teaching Material for the Course Educ.632, Addis Ababa: Addis Ababa University. (Unpublished)

Dejene Leta, 1994. Testing and Its Practical Application: Testing Reading in Focus (Unpublished Seminar Paper).AAU: Department of Foreign Language and Literature.

Ebel, R.L. and D.A. Frisible, 1991. Essentials of Educational Measurement, Prentice-Hell of India, New Delhi.

Fallik, F. and B. Brown, 1983. Statistics for Behavioural Sciences. The Dorsey Press: Homewood.

Ferguson, 1981. Statistical Analysis in Psychology and Education, 5th Ed. Mc Graw-Hill Company.

Finocchiaro, M. and S. Sako, 1983. Foreign Language Testing: A Practical Approach, New York: Regents pub. Comp, Inc.

Fitzpatrick, A. R., 1983. The Meaning of Content Validity, Applied Psychological Measurement Vol. 7. 1983. No. 1: PP.3-13. University of Massachusetts, Amherst.

Girma Lemma, 1997. Representativeness and coverage adequacy of Physics and Mathematics ESLCE items, “*Institute of Educational Journal” V.IV, 1997*. Addis Ababa University.

Gronlund, N.E., 1990. Measurement and Evaluation in Teaching. 6th Ed: Macmillan
Published Company. New York, London.

Hableton, R.K., & D.R. Eignor, 1979. A Practitioner's Guide to Critrion-referenced test Development, Validation &Test Score Usage (Report No.70). Amherst MA: University of Massachusetts, School of Education, Laboratory of Psychometrics & Evaluative Research.

Harmer, J., 2001. The Practice of Language Teaching, (3rd Ed). Longman: Parsons Education Limited.

Harris, D. P., 1979. Testing English as a Second Language. McGraw-Hill and India offset press, New Delhi.

Harrison, A., 1989. A language Testing, Handbook, London. Macmillan Published, Ltd.

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

Heaton, J.B., 1990. Classroom Testing. New York: London.

Henning, G., 1987. A Guide to Language Testing, Development Evaluation Research, Cambridge, Massachusetts Newbury House Publisher.

Hughes, A., 1988. Testing English for University, Oxford: Modern English publication and The British comp.

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

Hughes, A., 2003. Testing for Language Teachers. (2nd Ed.). Cambridge University Press, UK.

ICDR, 2000. English Language Syllabus for Grade Nine and Ten, Addis Ababa. (Unpublished)

Johnston, P., 2009. Enhancing Validity of Critical Tasks Selected for College and University Program Portfolios, The University of Tampa, *NATIONAL FORUM OF TEACHER EDUCATION JOURNAL*, 19(3): 3-6.

Jump, C. N., 1970. Introduction to Psychological Measurement, New York: McGraw Hillman.

Kerlinger, F.N., 1986. Foundation of Behavioural Research, New York: Holt, Rinehart and Winston.

Kifle kebede, 1992. An Assessment of Content –Related validity of High School: Addis Ababa, Addis Ababa University (Unpublished Thesis).

Lennon, R.T., 1956 & 1980. Assumptions Underlying the Use of Content Validity. *Educational and Psychological Measurement*. New York: Collier MacMillan.

Lewkowicz, J.A., 2000. Authenticity in Language Testing: Some outstanding questions in Language Testing.

Lin, R.L., 1974. Issues of validity in Measurement for Competency-based programmes. In A.R. Fitzpatrick, 1983. The Meaning of content Validity. University of Massachusetts, Amherst. *Applied psychological Measurement.* 7(1):3-13.

Lin, R.L., 1980. Issues of validity for Criterion-referenced Measures. Applied Psychological Measurement. 4:547-561. in A.R. Fitzpatrick, 1983. The Meaning of content Validity. University of Massachusetts, Amherst. *Applied psychological Measurement.* 7(1):3-13.

Loveinger, J., 1957. Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, 3: 635-695.

Madsen, H.S., 1983. Techniques in Testing. New York: Oxford University Press.

Meherens and Lehmann, 1991. Measurement and Evaluation in Education and Psychology, (2 th Ed.). New York: Holt, Rinehart & Winston

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

Mekonnen Mazengia, 1982. Content Analysis of Senior High School English Textbooks In Terms of Performance Objectives, Addis Ababa University (Unpublished MA Thesis).

Messick, S., 1975. The Standard Problem: Meaning and Values in Measurement and Education. *American Psychologist, 30:955-966*

Ministry of Education, 1994. New Education and Training Policy. Addis Ababa. EMPEDA

Ministry of Education, 1996. English for Ethiopia (Grade- 9 and 10): Students' Book (1st Ed) Addis Ababa: EMPDA.

Nitko, J.A., 1996. Educational Assessment of Students Ohio, A Simon and churner Company Englewood Cliff prance Hall, Inc.

Nuru Mohammed, 1992. "Level of Questions: A Description of Textbook and Examination Questions in Higher Secondary Schools" (Unpublished MA. Thesis). Addis Ababa University.

Ogunniyi, M.B., 1991. Educational Measurement and Evaluation. Lagos: Longman Nigeria.

Oller, J., 1988. "Practical Ideas for Language Teachers from a Quarter century of LanguageTesting Forum".

Preacher, K. J., 2001. Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software]. Available from http://quantpsy.org.

Popham, W.J., 1978. Criterion-referenced Measurement. Englewood Cliffs NJ: Prentice-Hall.

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

River, W.M., 1981. Teaching foreign Language Skills, The University of Chicago Process Ltd, London.

Rozeboom, W.W., 1966. Foundations of the Theory of Prediction. Homewood IL: Dorsey.

Siddiek, A.G., 2010. The Impact of Test content validity on language Teaching and Learning, Shaqra University, *Asian social science journal*, 6(12): 133-140.

Siddiek, A.G., 2010. Standardization of the Saudi Secondary School Certificate Examinations and their Anticipated Impact on Foreign Language Education, *International journal of humanities and social science, VOL. 1 NO. 3.*

Siegel, S., 1956. Non-parametric Statistics for the Behavioural sciences, McGraw Hill, inc. Tokyo.

Spolksy, B., 1978. Advances In language Testing. The Centre for Applied Linguistics, USA.

Stephen, G.S., 2007. Comments on Lissitz and Samuelsen on Validity Theory and Test validation. Educational Researcher, vol. 36. No. 8: PP.477-481

Stoddart, J., 1986. The Use and Study of English in Ethiopian Schools: Report for Ministry of Education. Addis Ababa (unpublished Manuscript).

Tirusseaw Taddesse, Desalegne Chalchisa and Getachew Yimer, 1992. A Compressive Approach to the Ethiopian Schools Leaving Certificate Examination; "Institute of Educational Research", Addis Ababa. Addis Ababa University.

Thorndike, M.R., 1997. Measurement and Evaluation in Psychology and Education (6th Ed). Prentice Hall Inc, USA.

Thorndike, R.L. and E.P. Hagen1977. Measurement and Evaluation in Psychology and Education. (4th Ed.). New York: John Wiley and Sons.

Teshome Demisse, 1995. “The Construction and Validation of Tests in English for Tertiary Education.” (Unpublished Ph.D. Dissertation). Addis Ababa University.

Tibebe Alemayehu, 1992. “The Predictive Validity of Ethiopian School Leaving Certificate Examination English and the Integrative Tests: Comparative Study.” (EJE, Vol. XIII, No 1).

Tylor, R., 2000a. English for Ethiopia. Addis Ababa: Ministry of Education.

Tylor, R., 2000b. English Syllabus for Grade 12. Addis Ababa: Ministry of Education.

Underhill, N., 1987. Testing Spoken Language a handbook of Oral Testing Techniques, Cambridge: University press. London.

Underhill, N., 1991.Testing Spoken Language. Cambridge: Cambridge University Press.

Venkateswaran, S., 1995. Principles of Teaching English. Delhi: Vikas Publishing Pvt. Ltd.

Walelign Admasu, 2006. Educational Measurement and Evaluation (Epsy 312). Department of Pedagogical Science, (Unpublished Handout), Haramaya University.

Weir, C.J., 1990. Communicative Language Testing, London: Prentice Hall internal Ltd.

Weir, G. J., 1995. Communicative Language Testing. Prentice Hall, New York.

Weirsma, W., 1995. Research Methods in Education. (6th Ed.) USA; A Simon and Schuster Company.

Wu, J., 2009. Insights in Language Testing: An interview with Jessica Wu. The University of Melbourne, Australia. *Shiken: JALT Testing & Evaluation SIG Newsletter. 13* (2) (p. 9 - 14)

7. APPENDICES

Appendix Table 1. The content coverage of the major language items of grade nine and ten old textbooks

Item	Coverage in number and percentage							
	9th		10th		Total		Average	
	Pds	%	Pds	%	pds	%	pds	%
Reading	31	18.23	25	18	56	36.23	28	18.115
Speaking	28	16.47	29	21	57	37.47	28.5	18.735
Listening	26	15.29	16	11	42	26.29	21	13,145
Writing	28	16.47	25	18	53	34.47	26.5	17.235
Grammar	29	17.06	25	18	54	35.06	27	17.53
Vocabulary	28	16.47	20	14	48	30.47	24	15.235
Total	170	≈ 100	140	100	310	≈ 200	155	≈ 100

Source: Grades 9 &10 English language plasma guides (old plasma guides)

Appendix Table 2 .The content coverage (frequencies) of the major language items in practice of grade nine and ten (new textbooks)

Item	Coverage in frequencies percentage and average							
	9th		10th		Total		Average	
	F	%	F	%	F	%	F	%
Reading	61	19	48	15	109	34	54.5	17
Speaking	81	25	80	25	161	50	80.5	25
Listening	28	9	34	11	62	20	31	10
Writing	52	16	56	18	108	34	54	17

Grammar	55	17	63	20	118	37	59	18.5
Vocabulary	47	14	34	11	81	25	40.5	12.5
Total	324	100	315	100	639	200	319.5	100

Source: English for Ethiopia, Student Textbooks (2010 editions) of grades 9 and 10. (F= frequency

Appendix Table 3. Chi-square distribution table.

Probability level (alpha,α)

(Df)	χ^2_c- value (critical chi-square value)										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	**11.07**	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
P -value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	**0.05**	0.01	0.001
	Non significant								Significant		

Source: From Wikipedia, the free encyclopaedia

Appendix Table 4. Response of Questionnaire

1. Teachers' Response for Testing the Following Test Items at Class

	Yes	No
Word formation	15(100%)	
Part of speech	15(100%)	
Pronunciation	8(53.3%)	7(46.7%)
Derivatives	7(46.7%)	8(53.3%)
Dictionary use	5(33.3%)	10(66.7%)
Antonyms	15(100%)	
Synonyms	15(100%)	

2. Degree of Teachers' response for testing the following test items

	Every semester	Once a month	Twice a semester	Rarely/once a year	Never	
Word formation	10(66.7%)		5(33.3%)			
Part of speech	7(46.7%)	4(26.7%)	4(26.7%)			
Pronunciati	3(20%)	3(20%)		2(13.3%)	6(40%)	
Derivatives	2(13.3%)		4(26.7%)	5(33.3%)	4 (26.7%)	
Dictionary			2(13.3%)	9(60%)	4(26.7%)	
Antonyms	6(40%)	5(33%)	4(26.7%)			
Synonyms	6(40%)	5(33.3%)	5(33.3%)			
Definition	5(33.3%)	2(13.3%)	6(40%)	3(20%)		
Gap filling	6(40%)	3(20%)				

3.Degree of teachers response for testing the following writing test items

	Every time	Almost every time	Occasionally	Almost never	Never
Letter witting		1(6.7%)	12(80%)	2(13.3%)	
Narration		1(6.7%)	12(80%)	2(13.3%)	
Description		3 (20%)	12(80%)		
Report writing		4(26.7%)	6(40%)		5(33.3%)
Summary writing			11(73.3%)		4(26.7%)
Note-making		3(20%)	8(53.3%)		4(26.7%)

4. Priority teachers give for testing the following grammar test items.

		High priority	Moderate priority	Somewhat priority	Low priority	Not a priority
Paraphrasing			3(20%)	4(26.7%)	8(53.3%)	
Completion		8(53.3%)	3(20%)	4(26.7%)		
Modified cloze	Testing prepositions	3(20%)	8(53.3%)	3(20%)	1(6.7%)	
	Testing articles	4(26.7%)	7(46.7%)	2(13.3%)	2(13.3)	
	Testing a variety of grammatical structures	9(60%)	5(33.3%)	1(6.7%)		
	Testing sentence	6(40%)	6(40%)	2(13.3%)	1(6.7%)	

5. Teachers response(in frequency) for testing the following sub-skills of reading

	Frequently	Occasionally	Rarely	Very rarely	Never

Scanning	12 (80%)	3(20%)				
Skimming	11(73.3	4(26.7%)				
Identifying stages of an		5(33.3%)	5(33.3%)	2(13.7%)	3(20%)	
Identifying examples presented in support of an argument		8(53.3%)	2(13.3%)	4(26.7%)	1(6.7%)	
Identifying reference of	13(86.7	2(13.7%)				
Using context to guess meaning of unfamiliar words	15(100%)					
Understanding relation between parts of text by recognizing indicators in discourse		9(60%)	4(26.7%)	2(13.7%)		

6. Teachers' response (in frequency) of testing the following sub-skills of listening

	Always	Usually	About half the time	Rarely	Never
Listening for specific		2(13.3%)		4(26.7%)	9(60%)
Obtaining gist of what is being said		2(13.3%)		4(26.7%)	9(60%)
Following direction	2(13.3%)			5(33.3%)	8(53.3%)
Following instruction	2(13.3%)			4(26.7%)	9(60%)
Note-taking		3(20%)		4(26.7%)	8(53.3%)
Partial dictation		3(20%)		5(33.3%)	7(46.7%)
Interpretation of intonation patterns		2(13.3%)	2(13.3%)	4(26.7%)	8(53.3%)
Recognition of function of structure				2(13.3%)	9(60%)

7. Teachers' response to the extent of testing their students oral ability at class

	To a great extent	To some extent	Very little	Not at all
Dialogue	10	5(33.3%)		
Expressing	5(33.3%)	3(20%)	3(20%)	4(26.7%)
Narrating	5(33.3%)	3(20%)	5(33.3%)	2(13.3%)
Eliciting	6(40%)	7(46.7%)		2(13.3%)
Directing	7(46.7%)	6(40%)		2(13.3%)
Reporting	6(40%)	3(20%)	2(13.3%)	4(26.7%)

8. Teachers response to the extent of including the phonological structures of English in their test.

	To a great	To some extent	Very little	Not at all
Stress		5(33.3%)	5(33.3%)	5(33.3%)
Intonation		5(33.3%)	5(33.3%)	5(33.3%)
Syllable		5(33.3%)	5(33.3%)	5(33.3%)
Length		5(33.3%)	5(33.3%)	5(33.3%)

9. Degree of testing students' knowledge of the following morphological structures of English

	Always	Often	Some times	Rarely	Never	
Prefix	11(73.3%)	4(26.7%)				
Suffix	11(73.3%	4(26.7%)				
Lexis	8(53.3%)	7(46.7%)				

10. Degree of testing students knowledge of syntactic structure and semantics

	Always	Often	Sometimes	Rarely	Never
Phrase	9(60%)	6(40%)			
Clause	9(60%)	6(40%)			
Semantics	9(60%)	6(40%)			

11. Teachers response to the extent of testing the following test items.

	To a great extent	To somewhat	Very little	Not at all
Matching	5(33.3%)	8(53.3%)	2(13.3%)	
Gap filling	5(33.3%)	7(46.7%)	3(20%)	
MCIs	15(100%)			
True/false		4(26.7%)		11(73.3%)
Essay writing		3(20%)	12(80%)	

Appendix Table 5. Questionnaire

HARAMAYA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

COLLEGE OF SOCIAL SCIENCES AND HUMANITIES

DEPARTMENT OF ENGLISH

Questionnaire to be filled by teachers and testing experts

Place: ________________________________, Date: ____________________________

Title: **Assessment of the Content Validity of the Ethiopian General Secondary Education Certificate (EGSEC) of English Examinations.**

Instruction: The purpose of this questionnaire is to gather information on the title indicated above. Dear respondents, since the reliability of this survey depends on the objectivity of your response, you are kindly requested to offer your response based on the factual and genuine information.

Direction: (1) You don't need to write your name.

(2) Encircle the letter of your option or tick [√] your option in the box for closed ended questions

(3) When written response is required, please make a brief comment.

(4) Respond all questions precisely and genuinely.

Thank you for your cooperation in advance!!

1. Have you ever tested your students the following vocabulary test items in your classroom test?

	Yes	No
Word formation		
Part of speech		
Pronunciation		
Derivatives		
Dictionary use		
Antonym		
Synonym		

2. How often do you test your students the following vocabulary test items in your test? (Put a tick mark [√] in each of the following rows)

	Every semester	Once a semester	Twice a semester	Rarely /once a semester	Never
Word formation					
Part of speech					
Pronunciation					
Derivatives					
Dictionary use					
Antonym					
Synonyms					
Definition					
Gap filling					

Comment if you would like to: ____.

3. How often do you include the following essay (letter writing, narration, description, simple report writing, summary writing, note making, etc) test items in your test? (Tick one box in each of the rows).

	Every time	Almost every time	Occasionally	Almost never	Never
Letter writing					
Narration					
Description					
Simple report writing					
Summary writing					
Note-making					

Any reason for "Almost never" and "Never" responses: ______________________________.

4. When you prepare grammar test items, put the priority you give to each of the following items. (Put the numbers 1-5 in each of the rows).

1= high priority 2= Moderate priority 3= somewhat priority 4= Low priority 5= Not a priority

Paraphrasing		
Completion		
Modified cloze	Testing prepositions	
	Testing articles	
	Testing a variety of grammatical structures	
	Sentence linking	

5. How frequently do you test your students the following **sub-skills of reading?**

(Put numbers 1-6 in each of the row).

1= frequently 2= Occasionally 3= rarely 4= very rarely   5= Never

Macro skills	Scanning text to locate specific information	
	Skimming text to obtain the gist	
	Identifying stages of an argument	
	Identifying examples presented in support of an argument	
Micro skills	Identifying reference of pronounce	
	Using context to guess meaning of unfamiliar words	
	Understanding relation between parts of text by recognizing indicators in discourse, especially for the introduction, development, transition and conclusion of ideas	

Comment if you would like to: ____

6. Put the numbers 1-5 in each of the rows for the frequency of testing the **sub-skills of listening** in your test? 1=always   2=usually   3=about half the time   4=rarely   5=never

Macro skill of listening	Listening for specific information	
	Obtaining gist of what is being said	
	Following direction	
	Following instruction	
	Note-taking	
	Partial dictation	
Micro skill of listening	Interpretation of intonation patterns(recognition of sarcasm(irony), etc)	
	Recognition of function of structure (such as interrogative as requests for example, could you pass the salt?)	

7. To what extent do you test your students' oral ability of the following sub-skills of speaking? Put the numbers 1-4 below in each of the rows as:

1= to a great extent 2= to some extent 3= very little   4= Not at all

Dialogue : conversation between or among people	
Expressing : thanks, requirements opinions, comment, attitude, confirmation, apology, want/need, information, complaints, reasons, justification	
Narrating: Sequence of events	
Eliciting: information, direction, service, clarification , help, permission (& all areas above)	

Directing: ordering, instructing (how to), persuading, advising, warning	
Reporting : description, comments, decision	

8. If your response to question number 7 is "Very little" or "Not at all", could you write your reason/s? ____.

9. To what extent do you include the following phonological structures of English in your English language test? (Tick one box in each of the rows)

	To a great extent	To some extent	Very little	Not at all
Stress				
Intonation				
Syllable				
Length				

10. If your response to question number 9 is "Very little" or "Not at all", give your reason/s? ____.

11. How often do you test your students' knowledge of the following morphological and lexical structures? (Tick one box in each of the rows)

	Always	Often	Sometimes	Rarely	Never
Prefixes					
Suffixes					
Lexis					

12. If your response to question number 11 is "Rarely" or "Never", could you write your reason/s? ____.

13. How often do you test your students' knowledge of the following syntactic structures and semantics? (Tick one box in each row).

	Always	Often	sometimes	Rarely	Never
Phrases					
Clauses					
Sentences (semantics)					

14. If your response to question number 13 is "Rarely" or " Never", please list your reason/s? ____

15. To what extent do you use matching, gap filling, multiple choice items (MCIs), true/false and essay writing items in your test? (Tick one box in each row).

	To a great extent	To somewhat	Very little	Not at all
Matching				
Gap filling				
MCIs				
True/false				
Essay writing				

Comment if you would like to: ____

Appendix Table 6. The summary of interview questions that were interviewed with teachers and testing experts.

1. Do the EGSEC English examinations cover all the four skills?
2. Are the EGSEC English examinations communicating language testing in comparison with the syllabi?
3. How often do you prepare table of specifications before test construction?
4. How do you decide the test items of your test?
5. What is your opinion about the contents of the EGSEC English examinations in accordance with the weight given in the syllabi or periods allotted?
6. Are the instructions of the EGSEC English examinations clear?
7. Is the test so constructed that the students begin with easier items and proceed to the more difficult ones?
8. Do you think that the EGSEC English examinations are standardized?
9. What looks like the relationship of English language teachers with regard to the construction of SATs? Do you prepare/design tests and contribute Questions to the NEAEA?
10. Can you tell me the major content areas of the books currently you are teaching in?
11. Can you tell me the major contents of your test? What about contents of EGSEC?
12. From the major content areas of the books, is/are there anyone that you do not include in your exam? Why?
13. If you are not testing all the contents of your textbooks, could you tell me the reasons why you are not testing all the content areas of the textbooks?
14. Do you believe the EGSEC English examinations are the representative sample of the content coverage of the books?
15. Have you ever taken any short or long term training in the areas of designing and analysing effective content valid language tests?
16. To which contents do you give emphasis when you prepare your English language tests?
17. Are you familiarized with the term content validity?
18. Do you include test items like listening skill, stress? Intonation and syllable? If you do not why?
19. Could you tell me the factors that reduce the content validity of EGSEC English examinations or questions?

Appendix Table 7. Analysis of the test papers by their sub-skills

	Items	2009	2010	2011	2012
Vocabulary	Word formation				
	Part of speech				
	Pronunciation				
	Derivatives				
	Dictionary use				
	Antonyms				
	Synonyms	9 (12%)	5 (6.67%)	5 (6.25%)	7 (8.75%)
	Definitions				
	Gap filling				
Writing	Letter writing	2 (2.67%)			
	Narration	2 (2.67%)	1	1 (1.25%)	1 (1.25%)
	Description		2 ()2.67%	2 (2.5%)	2 (2.5%)
	Report writing				
	Note-making				
	Mechanics	3 (4%)	3 (4%)	3 (4%)	4 (5%)
Grammar	Paraphrasing				
	Completion				
	Modified cloze				
	Preposition		1 (1.33%)		2 (2.5%)
	Articles				1 (1.25%)
	Grammatical structure	17	20	22 (27.5%)	24 (30%)
	Sentence linking	6 (8%)	3 (4%)	6 (7.5%)	
Macro-skills of reading	Scanning	7 (9.33%)	10	13	12 (15%)
	Skimming	11	11	7 (8.75%)	11 (13.75%)
	Identifying stages of an argument				
	Identifying examples presented in support of an argument				
Micro-skills of reading	Identifying reference of pronouns	4 (5.33%)	4 (5.33%)	6 (7.5%)	2 (2.5%)
	Deducing	1 (1.33%)	4 (5.33%)	1 (1.25%)	7 (8.75%)
	Understanding relation between parts of a text by recognizing indicators in discourse				1 (1.25)
Speaking	Dialogue	13	11	14 (17.5)	6 (7.5%)
	Expressing				
	Narration				
	Directing				
	Reporting				
Total		75	75	80	80