Global Scholars Journal of Linguistics and Communication Studies Vol.1 (1), pp 1-13 May, 2014
Available online http://www.globalscholarsjournals.org/SS/GSJLCS/index.php

Full Length Research Paper

Assessment of Content Validity in Ethiopian General Secondary Education Certificate of English Examinations (2009-2012)

Mekonnen Yibrah Kahsay[1], Dr. Julia Devardhi[2]* and Dr. Deepika Nelson[2]

[1]Ethiopian Ministry of National Defence, Eastern Command, Peacekeeping English Project P.O.Box: 19, Harar, Ethiopia.
[2]School of Foreign Language Studies, Haramaya University, Post Box -53, Dire Dawa, Ethiopia.

Accepted 28 April, 2014.

The purpose of this study was to assess whether the Ethiopian General Secondary Education Certificate (EGSEC) English exams administered by the National Educational Assessment and Examinations Agency fairly represent the content coverage and adequate sampling of the objectives stated in the syllabi. To attain the objective of assessing whether the EGSEC SATs of English exams possess content validity or not, quantitative and qualitative research designs, that is Mixed Method Approach was used. The qualitative methods were used to justify, clarify and interpret the data collected through interview and questionnaire. The quantitative method was used to show the outcomes of Pearson's chi-square test of independence and Cramer's coefficient of association. 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 used 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 divergent and weak in strength of association. This study only focused on the assessment of the content validity and was limited to four SATs and failed to see other SATs of EGSEC English examinations. Thus, the study recommends further researches to be conducted on the 'level of difficulty (item analysis)' and 'item discrimination' of the EGSEC English examinations

Key words: Content validity, assessment, language learning, backwash, relationship, testing.

INTRODUCTION

Harrison (1983) and Alderson et al., (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 (1990:244) defines "Content Validity is the extent to which the tasks required in the test adequately represent the behavioural domain in question." Moreover, Weir (1990) focuses on contents of standardized achievement tests that must reflect what has been in the syllabus and treated nationwide to arrive at the relevant decisions about students' performances.

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

* Corresponding Author email: devardhi.julia@gmail.com. Phone No: +251-915746481

such as face validity, construct validity and concurrent validity. This idea is 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 all 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.

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).

Importance of Content Validity in Achievement Tests

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 way 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 practicing those language areas and skills which do not appear or appear less in tests. They give considerable attention to those language areas which are going to be tested. In favour of this, Weir (1990) notes:

> *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 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 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.

Guidelines for 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 considerably care for writing tests. Anastasi (1975 & 1990) suggests 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 go 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 would forget to include relevant language elements into a test. Even the included ones would lack proportionality and some language areas will be dominant in the test whereas others will get little attention (Ibid).

According to Walelign (2006) the construction of a content valid test should take into consideration the major topics or contents and types of behavioural changes to be measured separately, the weight for various subject matter/topics and behavioural changes in terms of their relative importance, build TOS from the weighted lists of topics and behavioural changes, select/construct test items in line with the TOS’s. The closer the test

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

The Ethiopian Context

Validity, reliability and practicality have a mutual contribution to the good quality of a test; however, this study focuses on the content validity of the grade ten Ethiopian General Secondary Education Certificate **(**EGSEC) English examinations, which are prepared by the National Educational Assessment and Examinations Agency (NEAEA). This 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 the country. Hence, 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 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 conducted their studies on testing areas. Some of them were conducted on validity aspects. For example, Nugussie (2002) 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. He found that the content validity of the EGSEC English examination items in 1993 E.C. was very weak and divergent from the contents of the syllabus. The statistical result showed that the test items of the EGSEC English examination were not relatively valid and they did not match with the syllabus contents but his research was confined to only one year test items and it failed to see test items of other years' to confirm the invalidity of the EGSEC English examination.

Another research by Asmare (2008) conducted his research on the 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 textbooks or syllabi and sample test papers' contents. This was done with the purpose of assessing whether the coverage of English language tests administered in the college fairly represented the coverage of the textbooks. The result of the study showed that the contents of the sample test papers did not adequately represent the coverage of the textbooks.

Kifle (1995) also conducted a study on content validity of grade ten English language tests with reference to the former textbooks-English for Ethiopia- which was based on structural approach. 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 all the above mentioned studies showed that the examinations lacked content validity.

Rationale

The main reasons for studying the content validity of the EGSEC English language SATs were the observation of test papers when one of the researchers was in Axum Secondary School as a secondary school teacher. The treatment of all contents of the textbooks was not adequately 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.

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. However, some researchers in the area of testing expressed their fears 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, it was crucial to assess the content validity of the EGSEC English language exams in response to the aforementioned divergent views. The question remains open, who is at fault, whether the problem is on the part of the test designers or students are weak and remained unable to score well in their exams.

The present study thus, focused on assessing content validity of English language tests at GEQAEA of EGSEC level. It aimed at assessing 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. For this reason, the concern of this study was to assess the content validity of three EGSEC English examinations (2009-2011) in comparison with the 1996 textbook editions and one EGSEC English exam in comparison with the 2010 textbook editions and their syllabi. The current study has tried to answer the question whether these SATs are truly a representative sample of the contents of the textbooks.

Lack of Content Validity and Backwash

Backwash is an important concept that refers to the impact and influence the test can have on the teaching and learning process. This influence can be positive or negative. The backwash concept is very much connected with the test content validity. If the test lacks this quality, it consequently yields negative backwash on both the teacher and the learner and vice versa (Siddiek, 2010).

The key concept of content validity sometimes known as logical or rational validity can be briefly defined as the extent or adequacy with which the test items adequately

and representatively sample the content areas in the syllabi or students' book, then we can say that the test is content valid, and if the interaction is less, we can say that there is low content validity. In the teaching and learning process, 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, 1988).

In other words, teaching which is based on student-centred approaches mainly stresses the use of the four skills to the maximum that leads to the preparation of better test items that requires high level of mental effort. 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 it is intended to measure, reliable - scores should be consistent when repeated, and practical- a test should be economical, easy to administer and score. This process is known as backwash effect. Hughes (1988:1) defined the word as "the effect of testing in teaching and learning.

METHODOLOGY

Mixed research design was used. Qualitative data analysis techniques were used to interpret, clarify, and justify the data collected through interview and questionnaire. Quantitative approach was also used to show the outcome of the inferential statistics of the chi-square test of independence and Cramer's V coefficient of association to test the goodness-of-fit and strength of association between the contents of the textbooks and contents of the SATs, respectively.

To enrich the data obtained from the content analysis, questionnaire and interview questions were also formulated. The collected data were analysed in a quantitative and qualitative way. 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 were also included to enhance the reliability of the research. This research was based on the following hypothesis.

Null hypothesis (H_o) of this research is that there is no difference between the observed and expected variables and can be stated as follows: H_o: f_o-f_e=0 which states that the difference between the two variables is equal to zero.

Alternative hypothesis (H_a) H_a: f_o-$f_e \neq 0$ states that the difference between the observed and expected variables is not equal to zero.

Sample

Four years of EGSEC English Examinations prepared by the NEAEA from 2009-2012 in Harari Regional State were assessed 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 included.

Instruments

First document analysis was used to classify the contents of the textbook into six major headings or sections as reading comprehension, listening, speaking, writing, vocabulary, and grammar items. Then the total number of periods allocated or frequency of tasks for each major heading in both grade levels their average were 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.

To obtain the expected amount of questions (frequencies) for any cell in cross-tabulation in which the two variables are assumed independent, the row and column totals were multiplied for that cell, and the product was divided by the total number of cases. Thus, ten reading comprehension questions were expected and this figure was compared with observed reading comprehension questions in the actual test paper. This process was then applied to all major headings of the textbooks until 100 questions were 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. This value was interpreted based on the null (there is no difference between the expected and observed variables) and alternative hypothesis (there is difference between the expected and observed number of questions).

Next, a questionnaire was used to gather qualitative data by integrating both closed and open ended questions (20 questions). Some of these items included 1-4/5 Likert scale, containing rating scales. Views from 12 Harar Secondary school teachers regarding the content validity of the EGSEC English exams prepared by the NEAEA were also taken into consideration. Finally, an 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 major headings of the textbooks. They were also asked about their attitude towards teaching the neglected skills in testing (often speaking and listening) and phonological structures.

DATA ANALYSIS

The major headings of the textbooks were divided into six categories reading comprehension, listening comprehension, speaking, writing, grammar, and vocabulary. Using these prompts as functional exponents,

Table 1: Results of chi-square test analysis with regard to 2009 EGSEC English test items

	Coverage in numbers and percentages							
Item	Expected(E)		Observed(O)		O-E		$(O-E)^2$	$(O-E)^2/E$
	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
Vocabular	12	16	9	12	-3	4	9	0.75
Total	75	100	75	100	$\Sigma(O-E)=0$			$\sum \frac{(O-E)^2}{E}$=28.97

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

the amount of expected questions from all major headings were computed and those number 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. The period's allocation of the contents of the textbooks, and observed values of the test papers were used as independent variables. The dependent variable was the chi-square result. 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 of independence statistical analysis formula.

The chi-square test of independence tells us whether the two nominal categorical variables are related or not but it does not tell us the strength of that relationship. Thus, to know the strength of association between contents of the textbooks and sample SATs of this study, Cramer's V coefficient of contingency was preferred among the χ^2 based measures of association. The data collected using the unstructured interview and questionnaire were analysed qualitatively through evaluating, coding and tabulating using percentages accordingly.

RESULT AND DISCUSSION

Four of the SATs of EGSEC English examinations constructed, administered, scored and announced by the NEAEA from 2009-2012 were thoroughly analyzed to identify whether the tests constructed in the agency of the aforesaid years possessed content validity or not. To know the goodness- of- fit test between the observed and expected number of questions, chi-square test was used. It was noted that "observed values" refer to frequencies that 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.

Analysis and results of content validity of 2009 EGSEC English exam

In the 2009 EGSEC English examination test items shown in table 1, it is difficult to observe any correspondence between the observed test items and expected test items or fit values. 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. 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 proportionate to the amount of their coverage in the textbooks. The over treatment of reading and grammar, and maltreatment of the other skills in such SATs indicates that the students will neglect the practice of such skills, both in and out of the classroom. In line with this, Thorndike (1997) explained content validity is 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.

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

	Coverage in numbers and percentages							
Items	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	$\Sigma(O-E)=0$			$\sum \frac{(O-E)^2}{E}$ =47.49

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

The ''bad-fit'' of the observed and expected number of questions was statistically determined using Pearson's chi-squared test of independence. The calculated χ^2 value in table 1, 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). The two tailed p-value is $<$0.001, and the p-value of the χ^2 28.97 is =0.00002345. By conventional criteria, this difference is considered to be 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 have significant relationship to each other.

Analysis and results of content validity of 2010 EGSEC English exam

Table 2 indicates that the 2010 EGSEC English examination has a great disparity between the numbers of test items constructed by the GEQAEA when 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. The exam clearly showed the dominance of reading and grammar at the expense of other test items. It was twice the expected amount.

The representation of writing and vocabulary were reduced in number, almost by half than the expected representation of these items in the exam whereas speaking has got relatively fair representation. The testing of grammar (32%) was favoured next to the receptive skill of reading (38%). This implies that grammar was considered a mechanism of effectively evaluating the students' language proficiency. Finally, listening skill did not receive any representation at all. This biased testing practice can lead to the biased or inappropriate way of studying and learning the ignored skill or area of language in the testing system. This can potentially affect their global language progress. According to Abraham (2009), testing practice of testers by itself forces students to decide on what to study or not.

The statistical result of Pearson's chi-squared test of independence 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

Table 3: Results of chi-square test analysis with regard to 2011 EGSEC English test items

	Coverage in numbers and percentages							
Items	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	$\Sigma(O-E)=0$		-	$\sum \frac{(O-E)^2}{E}$ =45.79

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

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. This table entails that, there is a great disproportion between the fit-values and observed number of questions and this may result in negative washback effect on the students' methods of learning the language.

Analysis and results of content validity of 2011 EGSEC English exam

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. In favour of this, Jessica Wu (2009) noted that over emphasizing on the grammatical accuracy in an exam brings general lack of ability to communicate in English due to "old-fashioned" approaches in English education. 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 received less than half the 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 periods. Despite the coverage indicated in the syllabi, it was totally ignored in all the EGSEC English language testing system. This may affect students' attention towards learning listening activities in the classroom. 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) added that students expect exams to appear from what has been emphasized in the class.

There is a disparity in treating the language skills/areas of the test items. The syllabus gives weight to the receptive skills that is to prepare 14 reading and 11 listening questions but the test items do not weigh them as expected. Tamirat (2002) notes that emphasis in test construction and preparation must be according to the weight of the given syllabus. The general implication of this test construction leads to negative impact on students' receptive skills of English competency due to the harmful washback effect.

The result of the chi-squared test of independence 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 statement shows the mismatch between the expected (fit values) and observed number of questions in the test paper. Since the statistical result of the chi-squared test is greater than the value of

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

	Coverage in numbers and percentages							
Items	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
Vocabular	10	12.5	7	8.75	-3	3.75	9	0.90
Total	80	100	80	100	$\Sigma(O-E)=0$		-	$\sum \frac{(O-E)^2}{E}$ =56.86

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

the χ^2_c, the null hypothesis is rejected. A p-value of 0.05 or less is usually regarded as statistically significant. This means the two variables are statistically dependent or related to each other.

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 concludes that the procedure in 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.

Analysis and results of content validity of 2012 EGSEC English exam

Table 5 above 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 (6), 0 for listening, half for writing (7) and almost half for vocabulary (7) representations than expected to be appeared in the examination. The implication of the disproportion of these sections in this exam 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.

The chi-square test of independence illustrates that the 2012 EGSEC English examination was significant at 0.05 significant levels with a given 5 degree of freedom. The χ^2 value has to be at least $\geq$11.070. Therefore, the null hypothesis of no association between the two variables is rejected. 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. 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 items. 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.

This leads us to the conclusion that there are differences 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. Generally speaking, all the above reasons lead us to the conclusion that all the papers from 2009-2012 EGSEC English

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).

	Frequency by test year and content areas							
Content area of sample SATs	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	25.6
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

examinations lack content validity. 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 showed that four of the SATs of EGSEC English examinations were statistically significant when compared 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.

Strength of Association between the Textbooks and Sample SATs

The strength of association between contents of the textbooks and sample SATs of this study was calculated using Cramer's V among the χ^2 based measures of association because it is useful for comparing multiple χ^2 test statistics and is generalisable across contingency tables of varying sizes or is independent of the size of the variables.

Cramer's V always takes value in the interval [0.00, 1.00]. The coefficient ranges from 0 (no association) to 1(perfect association). Generally, the value for Cramer's V is $>$0.5, it can be considered to have a stronger relationship between the two variables, and with a smaller value V$<$0.5 indicates a weaker relationship (weak association) between the two observations (contents of the textbooks and sample SATs of EGESC English examinations). In this regard, Underhill (1991), Alderson, Claphan and Wall (1995) 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 textbooks and the frequency of test items in each content areas of the test paper. Hence, the sample SATs in which this study focused 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.

As can be seen from the above table 5, the sum of test items of each section of the three years along with its percentage is determined is 230. This figure will serve later as total frequencies of items in the test papers.

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	%	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 (grand sample size)

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 extents of relationship between the two observations were computed using data from the textbooks, and syllabi. In the same way, data from sampled test papers (former three exams) analysis are presented in table 5. The total frequencies of items for both observations were put into figures and percentages. Table 6

The contingency table is 2-by-6 (2 columns and 6 rows), q =2, which is the smaller of the matrix. The χ^2 result was 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) were computed applying Cramer's coefficient of contingency where n=385 (grand sample size) and q=2 and χ^2=59.69.

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?" was 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 $0.5<V<1$, it indicates the presence of high or strong association or strong relationship whereas if coefficient value reads $0<V<0.5$, it indicates the presence of weak association. 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.

Table 7: Total frequencies of test items of 2012

Content area	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

Summary of total frequencies of test items in sample test paper of 2012

The analysis of the distribution of different content areas of the sample standardized achievement test papers in relation to the contents of its sections is summarized in figures and percentages as shown in table 7 above.

Table 8: Total frequencies of content areas of 2010 editions of English for Ethiopia and sample test papers of EGSEC English examination (2012)

Content area	Frequency of practice tasks in textbooks		Frequency 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)

Total item frequencies of 2010 textbook editions versus EGSEC of 2012

The content coverage (frequencies) of the major language items in practice tasks of grades 9 and 10 and summary of total frequencies of test items in various content areas in the sample test paper by content area were determined. The total frequencies of items for both observations were merged and put into figures and percentages in table 8 above.

To calculate the relationship or strength of association between the textbooks (editions of 2010) and the 2012 test paper, Cramer's contingency coefficient statistical formula was applied where, χ^2 = chi-square test, n = grand sample size, q= minimum (r-1, c-1) and it was found to be V≈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. This "weak" association means that this sample test paper did not adequately represent the practice items and task formats in the textbooks. In other words, there was a sampling bias in the test paper, a significant deviation between the empirical data (values) and expected (fit) values or frequencies. The "bad-fit" can also be easily observed 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.

Testing micro and macro-skills of reading

The data showed 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 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 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, reading skill questions were two times greater than the expected. This distribution directs the students to concentrate on these sub-skill and teachers to focus their teaching only on these sub-skills ignoring the others as they are not or less important for the exam. Sub-skills of speaking, listening, components of morphology were also assessed, and these were mistreated. 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. Parts of reiteration and collocations which are the nuts and bolts of lexical structures were totally ignored. Nitko, (1996), 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 negative wash back effect.

The representation of syntactic structures was also part of the textbooks and, therefore, assessed. Four of the exam papers were thoroughly inspected. Except for the year 2009, test items of 2010, 2011, and 2012 exams focused 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 questions every year. The wash back effect of these test items was found to be positive as students gave due attention to concord in their textbooks. Further, semantics, phonological structures and intonation were ignored in the testing system of the EGSEC English exams. 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 wash back effect because SATs should fairly represent to all content areas in a syllabi or textbooks.

Analysis of teachers' views on content validity of EGSEC English exams

Every year before the national exams, teachers express their dissatisfaction regarding the format of the EGSEC English examinations that do not actually reflect what is given in the textbooks and what they teach. All the respondents shared the idea that the exams did not encompass all the four skills. This indicated that 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 prepared without taking into account the weight given in the syllabi or allotment of periods and frequency of practice tasks.

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 prevent cheating. In this method, some of the students received papers with the difficult items at the beginning, whereas others received papers with the easier ones. This frustrated the students and made them nervous. Such a method of testing contradicts the objective of the syllabi of grades 9 and 10, 2008 edition on page 2 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 the testing system of the agency was contrary with this idea. Such factors minimize the content validity of the SATs.

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. The teachers were asked if they are familiar with the term *'content validity'* and their cautious replies indicated that they and the 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.

All the school teachers do not use table of specifications (TOS) before their test construction. They simply prepare their exams looking at the major contents of the textbooks ignoring the weight given to each content. Such habits of the school teachers indicate their test lacks proportionality. The misuse of the TOS led to harmful washback effect and thus, low content validity. Supporting this view, Siddiek (2011) 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.

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. The exams do not possess content validity and have weak association/ strength with their respective textbooks as well. 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. Finally, the study showed that teachers' limited awareness concerning content validity affects the quality of language teaching and learning.

CONCLUSION

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. In all the SAT papers, use of synonym was given significant representation which has a positive washback effect on the students' behaviour, but some important components of vocabulary items such as dictionary use, antonyms, hyponyms, homonyms, etc were not assessed. The result of the statistical 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.

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 that there is a biased testing practice, and this may result in depriving the students of optimal learning.

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. 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. Such biased testing system of the NEAEA proves the wrong perception of the role of tests to the teaching process. The backwash effect is affecting many secondary school students negatively. Tests that favour certain contents of the textbooks cannot develop learners' skills and communicative abilities. This shows that the sample SATs remained weak at assessing the learner's knowledge of language proficiency in each area of the syllabus.

REFERENCES

Abraham K (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).

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

Alderson JC, Claphan C and Wall D (1995). Language test construction and evaluation Cambridge university press.

Anastasi A (1975 & 1990). Psychological Testing. New York, Macmillan publishing company.

Asmare M (2008). An Assessment of the Content Validity of English Language Tests: the case of Awassa College of Health Sciences. Addis Ababa: Addis Ababa University Press.

Bachman LF (1990). Fundamental Consideration in Language Testing, Oxford, Oxford University Press.

Ebel RL and Frisible DA (1991). Essentials of Educational Measurement, Prentice-Hell of India, New Delhi.

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

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

Heaton JB (1990). Classroom Testing. New York: London.

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.

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

Kerlinger FN (1973). Foundation of Behavioural Research, New York: Holt, Rinehart and Winston.

Kifle K (1995). An Assessment of Content –Related validity of High School: Addis Ababa, Addis Ababa University (Unpublished Thesis).

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 JA (1996). Educational Assessment of Students Ohio, A Simon and churner

Company Englewood Cliff prance Hall, Inc.

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

Nugussie T (2002). The Content Validity of the Ethiopian General Secondary

Education Certificate English Examination. (Unpublished MA Thesis). Addis Ababa University.

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

Siddiek AG (2011). Standardization of the Saudi Secondary School Certificate Examinations and their Anticipated Impact on Foreign Language Education, *International j of humanities and social science, 1(3):57*-64

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

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

Tibebe A (1992). "The Predictive Validity of Ethiopian School Leaving Certificate Examination English and the Integrative Tests: Comparative Study." (*EJE*, 13 (1)).

Underhill N (1991).Testing Spoken Language. Cambridge: Cambridge University Press.

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

Weir GJ (1990). Communicative Language Testing, London: Prentice Hall internal Ltd.

Weir GJ (1993). Communicative Language Testing. Prentice Hall, New York.

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):9-14.