JIMMA UNIVERSITY

COLLEGE OF SOCIAL SCIENCE AND HUMANITIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

A COMPARATIVE STUDY OF THE VALIDITY OF UNIVERSITY ENTRANCE EXAM AND CLASSROOM TESTS, AND THEIR LEVEL OF ACCURACY: *BIRBIRSA* SECONDARY SCHOOL, GRADE TWELVE IN FOCUS

BY: MEKETE MEKONNEN

A RESEARCH PAPER SUBMITTED TO DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE FOR PARTIAL FULFILLMENT OF M.A DEGREE IN TEACHING ENGLISH AS A FOREIGN LANGUAGE (TEFL)

ADVISORS: ENDALFER MELESE (Ass. Pro)

CO-ADVISOR: ASNAKECH DEMISSIE (PhD)

SEPTEMBER, 2021

JIMMA, ETHIOPIA

A Comparative Study of the Validity of University Entrance Exam and Classroom Tests, and Their Level of Accuracy: *Birbirsa* Secondary School, Grade Twelve in Focus

(A Comp. Study of the VUEE and VCT and their LA)

BY: Mekete Mekonnen

A Research Paper Submitted to Department of English Language and Literature for Partial Fulfillment of M.A Degree in Teaching English as a Foreign Language (TEFL)

Jimma University

College of Social Science and Humanities

Department of English Language and Literature

September, 2021

Jimma, Ethiopia

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university.

Name: Mekete Mokonnin Dirirsa

Signature: _____

Place: Jimma University

Date of submission: September 15, 2021

JIMMA UNIVERSITY

COLLEGE OF SOCIAL SCIENCE AND HUMANITIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

This is to certify that the thesis prepared by Mekete Mekonnen Dirirsa entitled **A Comparative Study of the Validity of University Entrance Exam and Classroom Tests, and Their Level of Accuracy: *Birbirsa* Secondary School, Grade Twelve in Focus** submitted in fulfillment of the requirements for M.A Degree (Teaching English as a Foreign Language) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

By: Mekete Mekonnen

Approved By:

Advisor: Endalfer Melese (Ass. Pro.)

_____ _____ _____

Co - advisor: Asnakech Demissie (PhD)

_____ _____ _____

Head Department of English Language

_____ _____ _____

Internal Examiner

_____ _____ _____

External Examiner

_____ _____ ___

Acknowledgements

I must first thank God the Almighty without His will and help I could not accomplish my study.

I would like to extend my heartfelt gratitude to various people for their useful assistance during my M.A work. Without the help of these people, this work would not have been realized.

I am grateful to my advisors, ***Endalfer Melese*** (Ass. Pro**)** and ***Asnakech Demissei*** (PhD) for their comments to the betterment of this work.

My warmest thanks are due for the hospitality/generosity of ***my close friends***, ***teachers and students***, particularly, of ***Birbirsa*** Secondary School who kindly worked together with me during the collection of documents and questionnaire implementation.

My special thanks go to my wife ***W/roTadelechTeshome*** and my children (***Mesin Mekete*** and ***Jitu Mekete***) for their overall support and care through the work.

Abstract

The main purpose of this study was to compare the validity of university entrance exam and classroom tests and their level of accuracy at ***Birbirsa*** *Secondary School (grade 12 in focus). The study employed comparative method research design. The participants of the study were 4 teachers and 33 students who were comprehensively selected. The study variables were comparing the university entrance exam and classroom tests validity and their level of accuracy. To collect data on these variables, instruments like document analysis, content analysis, questionnaire and interview were used. In this regard the 2020 First Semester English Language Classroom Tests, the 2020 University Entrance Exam, the Students' Average Result Indicating Document and A Grade 12 Syllabus were collected. The collected data were analyzed quantitatively and qualitatively. Findings of the study reveal that the classroom tests represented 46.5% of the instructional of objectives found in the syllabus, so the classroom tests holds below average content validity evidences. The average deviation of the result of the classroom test from the entrance exam is 38.13%, and the students rank is different in both tests result; so both tests have no concurrent validity, and the classroom tests has no predictive validity evidence. The average/mean of the classroom tests and the university entrance exam was 3.2 and 2.8 respectively, so both tests hold moderate face validity evidence. The researcher have recommended as the teachers should discharge one of their professional responsibilities through conducting a test in line with the test validation procedures; and the National Educational Assessment and Examinations Agency, under the Ministry of Education, should make reformations in the examination that would encourage improvement on validity of the tests, based on test validation procedures. The future researchers will have to conduct a study on construct validity of classroom tests and university entrance exam, and on the content and predictive validity of university entrance exam.*

Key terms: Validity, content validity, predictive validity, concurrent validity, face validity and accuracy

Table of Contents

Contents **Page**

Tables

CHAPTER ONE: INTRODUCTION

1.0. Introduction

This chapter deals with the introduction of the issue under study. This chapter incorporates issues like: the Background of the Study, Problem Statement, General and Specific Objectives of the Study, Basic Research Questions, Significance of the Study, Delimitation of the Study, Organization of the Study, Operational Definitions of Key Terms and Abbreviations.

1.1. Background of the Study

This study was initiated in response to the concerns about the validity of the university entrance exam (VUEE) and classroom tests (CT), and their level of accuracy (LA) at ***Birbirsa*** Secondary School. CT result should be equivalent with the UEE; the CT at this school seems as they were loosely conducted when checked for some validity evidences, and this was an indicator for invalidity of the tests**.** Different scholars have reached on different findings on test validity and also defined the key terms of this study as follows.

Aided Madurai (2016), defined evaluation, testing and examination as follows: "Evaluation is the collection, analysis and interpretation of information about any aspect of a program of education or training as a part of a recognized process of judging its effectiveness, its efficiency and its any other outcomes it may have. Test is used to examine someone's knowledge of something to determine what he/she knows or has learned. Testing measures the level of skills or knowledge that has been reached. Examination is a test to show the knowledge and ability of a student." In case of this study a grade 12 classroom test by the year 2020 was used as a data source due to the absence of second semester final examination, due to the wide spread of CORONA VIRUS (COVID-19) by this year.

Angus MC Donald (2000), defined an entrance examination as it is an examination that educational institutions conduct to select prospective students for admission. It may be held at

any stage of education, from primary to tertiary stage. The 2020 grade 12 UEE was used in case of this study.

Practicality, reliability, validity, authenticity and wash-back are the five useful guidelines for both evaluating an existing assessment and designing assessment tasks. The term validity refers to whether or not the test measures what it claims to measure. On a test with high validity the items will be closely linked to the test's intended focus. There are several ways to estimate the validity of a test including content validity, concurrent validity, predictive validity, construct validity and face validity, Browns, 2004,p:31. From the above five useful guidelines of evaluating the assessment validity; CT content validity and predictive validity, and concurrent and face validity of the CT and UEE were examined/checked, in comparison with external standards in case of this study.

Accuracy is closely related to the statistical term "validity." *However, establishing validity requires statistical analysis.* "Validity" is not simply a statistical term, although statistical methodologies are useful in some forms of validity evidence, "Accuracy," as a technical term, is the state of being precise or correct according to a traceable reference standard. It is always computed as a statistical process based on known and defined units. Assessments are accurate when they measure what they purport to measure. To this end the assessments should be aligned with the *standards and/or learning proficiencies* that they are designed to measure, Cronbach, L. J. (1949). Depending on this the VCT and the VUEE were examined based on *theoretical arguments/traceable reference standards,* and their level of accuracy (LA) is computed by using descriptive analysis statistical process.

The main focus of this study was checking the VCT and the VUEE, through the use of comp. research design, and comparing their LA. The four validity estimate items: *content validity, predictive validity, concurrent validity and face validity* were checked, and the results gained for both exams were compared only in case of concurrent and face validity of the exams. However the issue under study is interesting and researchable, no one has conducted a study on the issue previously and this was why the researcher has become more interested in conducting this study.

1.2. Statement of the Problem

Different scholars have defined testing, defined and classified test validity and test validity types, defined accuracy of tests, and comparative research methods in different ways. Different individual researchers and organizations had also conducted a study on the areas of test quality. And reached on different investigations; however, test validly is not still studied in encompassing all the validity items. Some commonly agreed up on definitions of the researcher's issue and his study variables, and the investigations of other researchers were presented.

English Language Test is used to examine someone's knowledge of something to determine what he/she knows or has learned. Testing measures the level of skill or knowledge that has been reached, Ebel, 1979. Test validity is the extent to which a test really measures what it is supposed to measure. If it does not meet that purpose then testing could be useless or misleading. Five types of validity are emphasized when looking at early readings: *content validity, concurrent validity, criterion-related validity, construct validity and face validity,* (Gipps, 1994). Test Accuracy," as a technical term, is the state of being precise or correct according to a traceable reference standard, Cronbach, L. J. (1949). From these five types of validity evidences three of them (content, criterion-related: concurrent and predictive, and face validity) of the classroom test, and two of the types of validity evidences (concurrent and face validity) of UEE were checked in comparison. The two of the types of validity evidences (concurrent and face validity) of CT and UEE were compared based on the result of comparison.

Content validity assesses whether a test is representative of all aspects of the content. To produce valid results the content of a test, survey or measurement method must cover all relevant parts of the subject it aims to measure. *Criterion-related validity evidence* is classified in to two kinds: *concurrent validity* and *predictive validity*. *Face validity* refers to whether the test predicts future performances accurately or well. A test holds *concurrent validity* evidence, if it is administered along with the established/well known/traditional test to a group of students. The scores of the two tests are correlated to determine the relationship. *Face validity* refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it. The face validity of the tests

determined not only by students but also by other stakeholders, Brown, 2004. This study has investigated, whether a test is representative of all aspects of the content in relation to content validity; if a test looks right, if it appears to measure the knowledge or abilities it claims to measure, and whether the test predicts future performances of the learners accurately or well in order to check face validity.

It is unquestionable as teachers at every school prepares tests themselves to determine what their pupil has learned or knows; it also undoubted as not all teachers refer to test validity criteria to validate their tests, and these in turn made the accuracy level lower. As a result of the students take invalid tests their knowledge is not measured well and the result/grade they score may become not ultimate substitute of their knowledge. So, the invalidity of tests may affect the students learning. Therefore, conducting a study on this issue seems vital.

The study which is conducted by, Bond (2003) on "***Exploring How Objects Can Influence The Level of Construct Validity of a Picture Vocabulary Test***", had reached on the finding as a test has a high level of construct validity when the items perform the same way across different groups.

Another study which was conducted by, David Ewing (2010), on "***Using Test-takers' Feedback To Enhance Quality and Validity in Language Testing***" by using a mixed methods approach had reached on the findings as language test candidates not only have strong opinions (both positive and negative) about the tests they take, but they also have a strong desire to share those opinions with test developers, and this type of feedback can then be used to substantially improve future tests, thereby helping to enhance the validity of the test system.

According to ***International Conference on Education & Educational Psychology (ICEEPSY)*** 2013, The EFL test results showed high reliability coefficients, however the test found invalid with regards to the content validity. Thus, almost half of the test items were identified as being in need of revision. In addition to revision of test items, the instructions for some sections were found to be necessary to edit.

A comparative study by Badia Muntazir Hakim, (2017), on ***validity of paper-based tests and computer-based tests in the context of educational and psychological assessment among Arab students*** result showed that (a) the pretests improved the results by providing experience for the tests themselves, (b) participants in computer-based testing (CBT) Group showed better test performance. In fact, CBT is known to be an efficient tool for assessment.

Melkamu Abate, 2007, had conducted a study on **"*The Wash back Effect of Grade Ten English Language (EGSEC) Examination*"** had investigated that the Grade 10 EGSEC English Language Examinations were found to have inadequate coverage of the contents (objectives) of the courses and high proportion of the examinations items was found to be relevant to few most important objectives of the program.

Simachew Gashaye, 2012,had conducted a study on **"*Wash-back of the University Entrance English Exam (UEEE) on Teachers' and Students' Practices*"** and investigated as the students were relying largely on exam-related materials other than the textbooks. The awareness that teachers had about the content and format of the exam geared their teaching to be exam-targeted. Different contextual factors contributed for the exam to influence students' practice to be exam-oriented.

The work of Fromse as Education and Training /FEAT/, 2015, on "***Evaluation of Learning Achievement in Selected Woredas in Amhara and Addis Ababa Sub-Cities*"**, shows the quality and usability of the test items vary significantly between subjects and teachers were found to demonstrate quite a high understanding of the theory of continuous assessment.

The works of the above researchers had been relied on ***test-takers' feedback to enhance quality and validity in language testing, construct validity of a picture vocabulary test, content validity, wash-back of the University Entrance English Exam, the teachers' awareness of content and format of the exam, inadequate coverage of the contents (objectives) of the courses and the quality and usability of the test items***; but this study mainly concerned on checking/examining four test validity types: ***content validity, concurrent validity, predictive validity and face validity*** of grade 12 CT and UEE, and comparing these two exams LA; which was the gap that the study is going to fill. From the five test's validity types construct validity is not directly included in the

study. Because, according to Alderson, 1995, construct validity is the most difficult of all validity types. This is partly because the concept is very complex to explain, and it encompasses all types of validity. Since test validity is not only the concern for grade 12 to be studied the result may have beneficial effect to all levels of educational settings. To compare the VCT and UEE data was collected by instruments such as: document analysis, content analysis, interview and questionnaire.

The major indicator for being applicable of tests with a lower validity was the exceeding of CT result, when compared with the UEE result at ***Birbirsa*** Secondary School; as the researcher observed from the school students' result register document. In other term the higher the students' CT result and the lower the result they scored in UEE was the best indicator of test validity problem; since valid tests should be always reliable, and has to yield consistent results. However a criticism for its being invalid needs a deep investigation.

The problem itself than what others have done and what makes my study different makes the study different as observed from the above overall presentation. What test validation criteria say and what are practically on application significantly vary. So, this encouraged the researcher to conduct a study.

1.3. Objectives of the Study

1.3.1. General Objectives

The general objective of this study was to compare the VUEE and VCT, and their LA at ***Birbirsa*** Secondary School (grade 12 in focus).

1.3.2. Specific Objectives

- To examine the VCT comparatively;
- To check the VUEE comparatively, and
- To compare the CT and the UEE level of accuracy.

1.4. Research Questions

In line with the theoretical justifications/the criteria for test validation and raw data gathered from the universe of the study this study must respond for the following questions:

1. To what extent, are the CT valid when it is examined in comparing with external standards?
2. To what extent, is the UEE valid when it is checked in comparing with external standards?
3. Which test (the UEE or the CT) is more accurate when compared to each other?

1.5. Significance of the Study

The investigation of the test validity level works for the future test validity improvement/to keep it up when it is highly valid. When the invalidity of the English language exams, whether it is the CT or the UEE is investigated, the recommendations may have beneficial effects to improve the validity of exams, and this in turn has values for developing different language tests to different educational level learners. The result of the study may have beneficial effect/values to the beneficiaries like:

- The primary beneficiaries from the result of the study are the grade 12th students at ***Birbirsa*** Secondary School, and grade 12 students at other schools in our country due to they sit for valid tests.
- Next to them, the English Language Teachers, and teachers in other discipline are also directly benefited from the result, to prepare a valid test for their learners being based on the recommendations given.
- It will enable the future researchers to only work on filling the gap between this study and his/her study rather than conducting his/her research newly.
- It also provides basic data for educational planners, and exam conductors.
- In addition to the above benefits it also gives information for educational leaders and managers.

1.6. Delimitations

This study was conducted in geographically delimited area at ***Birbirsa*** Secondary School, which is found at a distance of 698km from the capital of Ethiopia, 698km from the regional town and 98km from the zonal town in ***Nono Sele Woreda, Birbirsa Kebele*.** It was confined to grade 12 students whose total number was 81, and ***Birbirsa*** Secondary School grades 9-12 English

Language Teachers whose total number was 4. It was also theoretically delimited to making cross check of VCT and the VUEE to compare the LA of both exams. It only deals with the three tests' validity types, namely: ***content validity, criterion-related validity (concurrent and predictive validity) and face validity*** in case of the CT; but in case of the UEE concurrent and face validity were checked, and the two exams face and concurrent validity were compared. Because, other test and exam related evidences were not incorporated due to it makes the study beyond the capacity of the researcher to be managed. Hughes, 2003.p.30 stated that ***investigation of test's content validity and criterion-related validity provide evidence for its' overall, or construct validity***. In a sense of Hughes definition this study had also responded for construct validity, However construct validity was not studied in the researcher's case. Construct validity was therefore purposely rejected from the study. Therefore, the three validity evidences except construct validity of the tests from the four were studied.

1.7. Organization of the Study

This study was entitled on comparing the VUEE and VCT, and their LA at ***Birbirsa*** Secondary School: grade 12 in focus. Chapter one is an introduction part under which background of the study, statement of the problem, research questions, objectives of the study, significance of the study, scope of the study, and organization of the study were discussed; In chapter two review of related literatures is presented; Chapter three is about research design and methodologies; In chapter four the study has given the finding and procedures of data interpretations and analysis is presented, and chapter five is summary, conclusion and recommendations of the study.

1.8. Operational Definition of Key Terms

Classroom Tests: Are English language tests conducted by a grade 12 English language teacher at ***Birbirsa*** Secondary School.

University Entrance Exam: Is an English language exam given to grade 12 students for joining University by the year 2020

Validity: Is the extent to which a grade 12 classroom test and entrance exam really measures what it is supposed to measure.

Accuracy: Is the classroom test's and entrance exam's state of being precise or correct according to a traceable reference standards.

1.9. Abbreviations

CT: Classroom Tests

UEE: University Entrance Exam

VCT: Validity of Classroom Tests

VUEE: Validity of University Entrance Exam

LA: Level of Accuracy

CTLA: Classroom Tests Level of Accuracy

UEELA: University Entrance Exam Level of Accuracy

TV: Test Validity

SS: Secondary School

ESL: English as a Second Language

ICEEPSY: International Conference on Education & Educational Psychology

EFL: English as a Foreign Language

ESL: English as a Second Language

CVI: Content Validity Items

EGSEC: Ethiopian General School Examinations

ALTE: Association of Language Testers in Europe

VRIP: validity, reliability, impact and practicality

SILT: Studies in Language Testing

CPE: Certificate of Proficiency in English

CELS: Certificates in English Language Skills

BEC: Business English Certificates

TOEFL: Test of English as a Foreign Language

CIS: Candidate Information Sheets

GPA: Grade Point Average

TPTC: Table for Program Test Comparison

MOET: Ministry of Education and Training

IELTS: International English Language Testing System

FCE: First Certificate in English

UEE: University Entrance Examination

FEAT: Fromse as Education and Train

CHAPTER TWO: LITERATURE REVIEW

2.0. Introduction

This chapter was devoted to describe the literature reviewed for the study. The literature began with reviewing the roles of evaluation, particularly that of language tests and their influence on the teaching and learning process. Next, the literature was reviewed or the terms that were related to validity of tests. This was followed by identifying types of validity. Thirdly, methods of seeking and analyzing test validity evidences were presented and the review ended with summarizing the result of different researchers on the related areas of the researchers' concern.

2.1. The Role of Evaluation in Education

Evaluating the educational processes and outcomes has been practiced for a long time in the educational settings. Through the evaluation process, judgment is made about the performances of students and institutions, the processes of the classroom practices of teachers and students, and the usability of the instructional materials and the curriculum at large. The judgment on the processes and outcomes of the educational activities is made on the basis of the information obtained through the different assessment tools, Ebel, 1979.

According to Ebel (1979:22), Evaluation is an indispensable part of educational processes. Unless the effectiveness of instructional processes and the achievement of students, the efficiency of educational institutions, and usefulness of the designed curriculum with the supposed instructional materials and teaching methodology are evaluated, the success of the educational outcome would be doubtful. Since education is a giant enterprise, it needs regular evaluation. As long as there is teaching, there is a need to evaluate its processes and outcomes. Thus, evaluation in the educational system becomes an integral part of the teaching-learning process to determine the achievement of the desired educational objectives.

2.2. Language Tests

Language tests, as is the case in the general education, have a range of functions for they are used to measure a sample of traits or attributes. The tests administered at different levels in the educational system provide dependable data either qualitatively or quantitatively about the learners' progress or achievement, the suitability of the curricula, the effectiveness of the methodology and instructional materials. Moreover, tests, specially the public ones, are used to avoid nepotism and favoritism in the educational opportunities given through keen competitions to candidates, to encourage development of talent, and to upgrade school performance. Different parties such as policy makers, teachers, and employers use the data that are obtained from the outcomes of tests. This information saves wastage of time, money, and energy of the concerned stakeholders and the country at large. This is the diagnostic function of tests. In addition to their diagnostic function, test scores have been extensively used for selection and placement purposes of the candidates into the various academic programs. Moreover, test scores have extrinsic power of motivating students, teachers, and institutions to work better in the teaching and learning processes. Anticipating the forthcoming examination, students learn and practice the lessons better, (Bachman, 1990).

The recurrent use of test scores for the selection and placement of candidates in the educational system leads tests to produce impact backwards on the different aspects of teaching and learning practices. This is due to the tradition that high score in the tests is associated with a good candidate, (Bachman, 1990).

2.3. The Test Development Process

Bachman and Palmer (1996, p. 85) provide a conceptual framework for the development process which is organized into three stages. Accordingly, test usefulness is the most important quality of any test and should be taken into account at all stages of the development process; usefulness includes reliability, construct validity, authenticity, instructiveness, impact and practicality, which test developers should balance appropriately to optimize the usefulness of their test.

Cambridge ESOL have provided a basis for assessing usefulness consisting of four key examination qualities, validation, reliability, impact and practicality (Hawkey,2009, p. 127). This process of validation is operational using Weir's (2005) cited in Hawkey p.173) socio-cognitive approach to test validation. It views any testing activity as "a triangular relationship between three critical components: the test taker's cognitive abilities, the task and context, and the scoring process". Weir's framework sees construct validity as consisting of three symbiotic elements: cognitive, context and scoring validity. By separating context validity from scoring and cognitive validity the framework allows for adjustments to be made depending on the skill being tested.

2.4. Validity of Tests

The most traditional definition of validity is the extent to which a test really measures what it is supposed to measure. If it does not meet that purpose then testing could be useless or misleading. Five types of validity are emphasized when looking at early readings: content validity, constructs validity, concurrent validity, predictive validity and face validity.

- Construct validity refers to whether the test is actually adequate to what is being assessed.
- Predictive validity refers to whether the test predicts future performances accurately or well.
- Content validity covers the more appropriate and necessary content which is necessary for a good performance.
- Criterion related validity evidence is whether a test correlates with or gives nearly the same result as another similar test of the same skill.
- Face validity refers to the extent to which a test looks effective in terms of what it is expected to measure.

However, emphasis on these different types of validity has led to a situation where evidence might point to only one or perhaps two of these various validity types when developing tests (Gipps, 1994).

2.4.1. Content Validity

Content validity assesses whether a test is representative of all aspects of the construct. To produce valid results the content of a test, survey or measurement method must cover all relevant

parts of the subject it aims to measure; if some aspects are missing from the measurement (if irrelevant aspects are included), www.thegraidenetwork.com.

Example

A mathematics teacher develops an end=of-semester algebra test for her class. The test should cover every form of algebra test for her class. If some types of algebra are left out, then the results may not be an accurate indication of students' understanding of the subject. Similarly, if she includes questions that are not related to algebra, the results are no longer valid measure of algebra knowledge, www.thegraidenetwork.com

Content validity includes any validity strategies that focus on the content of the test. To demonstrate content validity, testers investigate the degree to which a test is a representative sample of the content of whatever objectives or specifications the test was originally designed to measure. Content validation which plays a primary role in development of any new instrument provides evidence about the validity of any new instrument by assessing the degree to which the instrument measures the targeted construct. All elements of the instrument (e.g. items, stimuli, codes, instructions, response formats, scoring) that can potentially impact the scores obtained and the interpretations made should be subjected to content validation. There are three key aspects of content validity: domain definition, domain representation, and domain relevance, Sireci, 1998a.

In order to have a valid test, the course contents should be matched with the instructional objectives. This is possible if a table of test specification is used, EQUIP,2008:16.

2.4.2. Construct Validity

Construct-related evidence is widely called construct validity. But the term construct validity is the most difficult of all validity types to understand. This is partly because the concept is very complex to explain and partly, it encompasses all types of validity (a super ordinate form of validity to which content and criteria related validity contribute), Alderson, 1995.

According to Brown (2004, p. 25), construct may or may not be directly or empirically measured, their verification often requires inferential data and a test has construct validity evidence if its' relationship to other information corresponds well with some theory. To have evidence on the construct validity of a test, a tester has to assess "… to what extent the test is successfully based up on its' underlying theory…. The experts are selected, given some definitions of underlying theory and asked to make judgments after an inspection of the test as to its construct validity." (Anderson,1995, p. 183).

2.4.3. Criterion-related Validity Evidence

Criterion-related validity evidence refers to the evidence that, it is obtained by correlating test scores with an external standard or criterion. Criterion-related evidence is classified in to two kinds: concurrent validity and predictive validity. Test holds concurrent validity evidence, if it is administered along with the established/well known/traditional test to a group of students. That is to say both tests are administered at the same time and the scores of the two tests, the new and the established one, are correlated to determine the relationship. This is done to obtain empirical evidence about their relationship (Brown, 2004, p. 24).

Kubiszyn & Borich, 2003:301, point out that, it is possible to determine the test's concurrent validity evidence by administering two tests to one group of students, one basic test that requires 60 minutes, which has been used for a long time and the shorter and the better test that has been developed recently and that requires 20 minutes for the administration to the same group of students and then the computation of a correlation coefficient of the two results is required to obtain empirical evidence about their relationship.

2.4.3.1. Concurrent Validity Evidence

A test holds concurrent validity evidence if it is administered along with the established test to a group of students. …the scores of the two tests are correlated to determine the relationship. …the validity of a high score on the final exam of foreign language course will be substantiated by actual proficiency in language, Brown, 2004.p.24.

2.4.3.2. Predictive Validity Evidence

Predictive validity evidence refers to how well the test predicts the taste takers future performance. This form of validity evidence is required for the educational institutions that give entrance exam for the candidates who want admission to college or University. Placement test … predicts success in college and make decision who should be admitted to the University. To determine the ability of final exam to predict success after the exam, the scores and final grades of the same students were obtained and correlated with their final exam scores. Brown, 2004.

Predictive validity evidence refers to how well the test predicts the taste takers future performance. This form of validity evidence is required for the educational instructions that give entrance exam for the candidates who want admission to colleges or Universities. The proficiency test, placement taste or the aptitude test administered to the candidates should have predictive validity so as to predict success in college and make decision who should be admitted to the University. To determine the predictive validity evidence of a test, one has to administer the taste to a group of subjects and correlate the results with the tests results which the subject takes after a certain period of time. Predictive validity was estimated by correlating the scores of the midterm achievement tests with those of the final exam, as well as by correlating the final exam scores, Kukuk Walters, 2009.

2.4.4. Face Validity

According to Brown, 2004,p.27, Face validity refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers. The face validity of the tests determined not only by students but also by other stakeholders. In fact of all the stakeholders, students would like to express their opinions about the taste taken. Form the students' perspective to what extent the test does its job or is the test easy or difficult? Does it look realistic or trivial? The students' responses to such questions will help to decide the face validity of a test. As suggested by Brown, a test will likely have high face validity if students take:

- A well-constructed, expected format with familiar tasks,
- A test that is clearly doable within the allotted time limit,
- Items that are clear and uncomplicated,
- Tasks that relate to their course work/content validity, and
- A difficulty level that presents a reasonable challenge

Face validity was determined by asking both the instructors and students about how well they felt the content of the course were represented on the achievement tests.

2.5. Evidence in Validity

Weir (2005), states that the satisfactory evidence of validity is highly necessary for any serious test. Then he addresses two different concepts when he describes the importance of validity.

Concept 1, Validity resides in test scores. By this he means that validity in this case might be better defined as the extent to which a test could produce proper data, i.e., test scores which are accurate in their representation of what level a student is at regarding their skills or knowledge of the language. He is more concerned about the scores produced by a particular administration of a test on a particular sample of candidates. Then over time, cases can be made that different tests are valid if various versions of a test or administrations of a test show similar results.

Concept 2, Validity is multifaceted. Weir explains this concept by saying that to support any claims for the validity of scores on a test, there is a need of different types of evidence. As an evidential basis for test interpretation, these are complimentary aspects and not alternatives. One single validity aspect may not be looked upon as better or superior to another. If there is deficit in any one, then it might raise questions regarding how well-founded the interpretations of test scores are.

As we can see from the presented research above, one might think that when we use the term validity in testing, it might seem as if we should consider validity as a checklist procedure. However, that is not the case. As Lane, 1999, points out, when accumulating validity evidence, it

should be treated as more than a checklist procedure. He point out that the validation process involves the development and evaluation of a coherent validity argument for and against proposed test score interpretations and uses. In the validity argument, each inference is based on a proposition or assumption that requires support.

Messick. S, 1989 (in Lane,1999) states that the process of validation involves gathering evidence for and looking into possible threats to the validity of test score interpretations. Furthermore he argues that "...the most attention should be given to the weakest part of the interpretative argument because the overall argument is only as strong as its weakest link" (Lane, 1999, p.1). Moreover, Lane argues that to establish what validity evidence is necessary for a particular purpose of testing, analysts should define a set of propositions that would support the proposed interpretation. For each proposition evidence should then be collected as support.

2.6. Good Practice and Test Validation

Saville (2003:65–78) summarizes the implications of such good practice as the need to pursue test *validation*, namely to make every systematic effort to ensure that a test or exam achieves:

- appropriateness to the purposes for which it is used
- the ability to produce very similar results in repeated uses (Jones 2001)
- positive influence 'on general educational processes and on the individuals who are affected by the test results' (Saville 2003:73), and
- Practicability in terms of development, production and administration.

In the theory and practice of Cambridge ESOL test research and development, these four exam targets are labeled ***validity, reliability, impact*** and ***practicality*** (VRIP for short). The overlap with Bachman and Palmer's 1996six test ***usefulness*** qualities, ***reliability, construct validity***, ***authenticity, interactivity, impact*** and ***practicality*** *(*see previous*)* is neither insignificant nor coincidental given the close relationship of Bachman with UCLES. On the2005 Cambridge ESOL website, reference is still made to activities planned as a follow up to the work of Lyle Bachman and colleagues, on what was known as the Cambridge-TOEFL Comparability Study, carried out between 1987–9.According to Saville,2003, impact studies cover three major groups

of stakeholders: the examination developer, the examination taker, and the examination user, that is someone 'who requires the examination for some decision-making or other purpose.'
Although there are four components to VRIP, it is clear that they are by no means independent. Test validity, for example, in the unitary construct proposed by Messick (1989:16), subsumes reliability, impact and practicality.

Individual examination qualities cannot be evaluated independently. Rather the relative importance of the qualities must be determined in order to maximize the overall usefulness of the examination; Weir (2004), in his socio-cognitive framework for validating tests has a similar perspective. For him, test ***validity*** is the super ordinate category to ***theory-based validity***(covering internal language ability processes), ***context validity***(the appropriateness of the communicative and administrative context in which the test takers are called upon to perform) and ***scoring validity*** (the dependability of test results, subsuming the conventional category of reliability). Test validity also subsumes the two post-test validities, ***concurrent*** **and** ***consequential***, the latter, of course, including the study of the impacts of the test on stakeholders. But Weir reminds us that test 'validity is perhaps better defined as the extent to which a test can be shown to produce data, i.e. test scores, which are an accurate representation of a candidate's true level of language knowledge or skills'. Weir adds, 'over time if various versions of a test or administrations of the same test provide similar results then synthetically a case may be made for X or Y test being valid overtime and across versions and population samples.' Mc Namara, (2000:138), suggested that validity is about 'the relationship between evidence from test performance and the inferences about candidates' capacity to perform in the criterion that are drawn from the evidence.

2.7. Methods of Seeking and Analysing Validity Evidence

According to Messick, (1994) one of the main assumptions to be tested is that the University Entrance Examination(UEE) test users might have used a trait-based approach to measurement when they added up the correct answers, divided the resulted raw score by 8 and then reported the result as one final score. Therefore, both general techniques for collecting validity evidence and those specific to the latent trait -measurement perspective will be discussed.

2.7.1. Content Relevance, Representativeness and Technical Quality

Regarding the UEE English test, in stating that the content of the test will come entirely from the high school English program as stipulated by the Ministry of Education and Training (MOET), (Vu, 2005), the test designers made a commitment that the test would relate to and represent the content of the English curricula in use in the three-year high school program in Vietnam.

Also, in using the test for college selection, the test users also assumed that the test content would also be relevant to and representative of the criterion domain, that is the English program at the college. To evaluate the content relevance, content representativeness and technical quality of the test, Messick (1994) acknowledged the importance of expert judgment. Seven experts were thus recruited to analyze the 2008 English test paper. Four of them were teachers of English from different high schools in Hanoi, the capital city, and one of them was from a remote province in the centre of Vietnam. Two others were teachers of English and applied linguistics from the CFL that admitted successful candidates. They all had prior training in test design and from six to 20 years of English teaching and testing experience. Besides, none of these participants was involved in the production of the test in question. Hughes (1989, p. 22) considered this an important criterion for fair judgment. To begin with, these teachers were requested to conduct individual content analysis of all the 80 test items. First, they identified the specific knowledge and skills needed to get each item right, specified the testing points of each distracter (i.e. the wrong answer choice alternative) and recorded their independent judgments in an Item Content Analysis Form designed by the researchers. They were then asked to analyze the test using Hambleton's 1980 Item Bias Review Form (reproduced with the author's permission) to detect possible bias in items, and Item Technical Review Form (adapted with permission from Hambleton (1984, p. 227)'s form to accommodate features of an English test) to judge items' technical quality. This first stage of working with individual experts is important because the content aspect is not "the surface content of test items or tasks but the knowledge, skills, or other pertinent attributes measured by the items or tasks". The chance for individual analysis is also essential to ensure that each expert analyzed the test thoroughly by him/herself, (Samuel Messick, 1989, p. 39).

Messick, 1989, pointed out that in reality, experts sometimes do not prepare carefully before the group meeting and their judgment in such cases is often influenced by the group dynamics. Upon completion of individual content analysis, the six experts who were available for further participation were invited to join a group discussion on issues concerning the test content under the facilitation of one researcher. The questions directed toward high school teachers were (1) to what extent the knowledge and skills measured by the test as analyzed were representative of the high school English language curriculum, (2) whether there was any important area of knowledge and skills not covered in the test, and (3) to what extent the test tasks, the test format, the administration conditions, the scoring criteria were relevant to what candidates had learned at high school. Similarly, the questions to CFL teachers were (4) whether the knowledge and skills measured by the test as analyzed were representative of the knowledge and skills required for college studies, (5) whether there was any important area of knowledge and skills not covered in the test, and (6) to what extent the test tasks, the test format, the administration conditions, the scoring criteria were relevant to what candidates would be learning at the college. The six experts were also asked what recommendation they would like to make to improve the content relevance, representativeness and technical quality of the test. Answers to all these questions were tape-recorded for subsequent qualitative data analysis. As can immediately be seen, this comprehensive review, though named content analysis, would yield information that fell into the overlapping area between content-related and construct-related evidence. Besides expert judgment, empirical analysis can also lend valuable information about the test content.

To assess the item technical quality empirically in the Rasch measurement, Smith (2004. p.107) suggested using item fit statistics to evaluate the extent to which items tap into the same construct and place test-takers in the same order. He argued that test-takers should be ranked consistently by items measuring the same construct. If not, the miss-fitting items to the Rasch model, i.e. the items that measure a different construct compared to other items in the test, should be subject to revision or elimination.

The representativeness of the content can be evaluated empirically in the Rasch measurement via the inspection of the spread of the items along the person ability – item difficulty scale and their

individual standard errors in the item calibration. If gaps are found in any regions of the variable, the content is said to be under-represented, and new items need to be designed to fill them up. The aim is to have a variable that cover test-takers' ability range (Smith, 2004, pp. 105-106). Since content-related and construct-related inferences are practically inseparable, the analysis of the knowledge and skills required of the examinees to do the test successfully necessitates empirical construct validation (Cronbach, 1971, p. 452) as in the substantive and structural aspects that follow.

2.7.2. Cohesion of the Test Items

To investigate substantive and structural aspects of validity and answer the question regarding the soundness of the test as a measure of English language ability, the test scores and the item responses to the 42-item test of the whole population of 33 classroom test takers and university entrance exam takers of 2020 were collected from exam record. As stated earlier, the test users must have made an assumption that the English ability measured by the UEE English test was dimensional construct when they reported only one single score. This assumption could be best tested using the simple logistic model or the Rasch model (Rasch, 1960) to calibrate all items. This is a process that uses a logistic function to establish the common reference linear equal interval scale that expresses both item difficulty and person ability. On the basis of these two parameters, the probability of a person succeeding on an item can be determined. The Rasch model was selected for this study due to its strengths over typical test theory.

The most important advantage is that the Rasch modeling permits the person ability parameter and the item difficulty parameter to be estimated independently from each other, thus enabling complete objective measurement of items and persons (Wright, 1977). Another advantage is that the Rasch model analysis yields item and person separation indices and an accompanied picture of the item and person map that shows the relationship between the item difficulty distribution and the person ability distribution. The item separation index along with the hierarchy of item difficulties is valuable for defining what is measured by the test, thus constituting a measure of construct validity. The person separation index along with the hierarchy of person ability is useful for establishing concurrent validity. Besides, the linearity of the Rasch measures and the

capacity of the calibration process to calculate measurement error for every item and person estimate make the calculation of reliability or the precision of measurement more accurate (Wright &Masters, 1982).

If all the items of the 2008 UEE English test were found to fit the Rasch model, then it would be the evidence for the underlying hypothesized construct of English language ability that supports the current practice of reporting one single score. Statistics from the Rasch item analysis (e.g. fit statistics, item difficulty and person ability estimates, measurement errors, item and person hierarchy and information from differential item functioning analyses) could then be sought as evidence of the structural and substantive aspects of validity. Indices of test reliability such as item separation index and person separation index would also be checked to evaluate the degree of stability, consistency and accuracy of the test results. In addition, experts' opinion on the quality of item distracters and the descriptive statistics of how many test-takers chose each option were further information that helped answer this second question. On the contrary, if the data were found not to fit the model, then the extent of misfit would be evaluated to investigate possible causes. For example, the content analysis by experts and the empirical statistics of the miss-fitting items might reveal that the items were technically flawed or that they measured something else rather than English language ability. The former possibility could be the evidence against the inclusion of such items in the test, and the latter could work against the assumed undimensionality of the test items. Whatever the outcome of these statistical and judgmental analyses, one definite gain is a better understanding of the items making up the UEE English test, Wright &Masters, (1982).

2.7.3. Prediction

The judgment of a test's predictive power would involve a predictor and a minimum of one criterion to be predicted. The predictor in this study was naturally the UEE English test score. The criterion was selected to satisfy certain prerequisites: the criteria have to be relevant to the final aim of test score use, reliable, unbiased and practical in terms of effort, time and cost (Thorndike, 1949). Typically, the predictive power of admission tests is judged by the degree to which test scores can predict immediate criterion such as the first-year college grade point

average (GPA) (Zwick, 2002). According to Zwick, (2002), the reason for this is that subsequent collection of grades may risk seeing more students dropping out or transferring to other schools.

It was reported in most predictive validity studies that predictive power was higher in semester 1 than in semester 2. The arrows in the model indicate expected causal connections among the three variables, each of which has an error term going in to it. The details of how these variables are measured are as follows. First, the UEE English score was measured using the Rasch model because, as mentioned earlier, the Rasch measures have better measurement properties than the raw scores (and their rescaled versions) while they still rank test-takers in the same way as the raw scores do. In fact, the correlation between the raw scores and the Rasch logic measures is close to 1 (Wu & Adams, 2005).

With respect to the English achievement scores, a mean score of the results of the two English subjects Oral and Written communication was obtained in both semesters following the common practice of using GPA as the criterion measure of achievement in predictive validity studies. Besides, this is also consistent with the current fashion in English language testing to report the composite score (in the case of the Test of English as a Foreign Language) or the mean score (in the case of the International English Language Testing Service Test).

To be supporting evidence for the claimed predictive power, these values should be as high as possible. In reality, there is no definite or rule-of-thumb value that coefficients should be, so validity coefficient is often judged relatively against “typically obtained” coefficient of similar tests,(Gronlund, 2006, p. 207).

2.7.4. Consequences

The data needed to answer this question will come from the content analysis of the test, the empirical statistics of test scores and expert judgment. All types of evidence (content bias, the overall test quality, reliability, construct validity, and expert opinions, etc.) will be integrated to form an overall evaluation of the consequence of the test. To be a good test with positive consequences, the test should be free from sources of invalidity like construct under presentation and construct irrelevance variance that may put test-takers at an unfair disadvantage by lowering

their scores. It should also be free from bias in scoring and interpretation or unfairness in test use. It is our belief that the above range of evidence we collected is sufficient and the combination of judgmental and statistical approaches to answering the four questions is adequate for the intended validation argument.

2.8. Research Comparing FCE and another Examination

A further possibility would be to compare the results of the FCE test to another test. One interesting comparison for Switzerland would be to compare the rank order of FCE results against the rank order of "Berufsmatura" results as well as the correlation coefficients of the two examinations. Having this information would allow students and future employers to firstly see if the examinations measure the same thing and also to see whether one test is more difficult than the other, Deborah Grossmann, (2010:28).

2.9. Research Findings

Different studies had been conducted on the nearly the same areas with on what the researcher is going to conduct a study. Their investigation is discussed as follows:

Simachew Gashaye, 2012, who had conducted a study on related issue has presented on pp:186-189 different contextual factors contributed for the exam to influence students' practice to be exam-oriented which is shortly set from 1-5 here under.

1. Teachers' practice was found influenced by the wash-back of the exam in that their focus area of teaching was on form-related language aspects mainly on grammar, vocabulary, and technical aspects of speaking and writing skills.
2. Students' practice was also found influenced by the wash-back of the exam.
3. Different factors were identified that contributed for the exam to influence teachers' practice to be exam-oriented. Teachers' personal and contextual factors were found mediating their practice to be exam-oriented.
4. Different contextual factors contributed for the exam to influence students' practice to be exam-oriented. The first factor that strongly influenced the students' practice to be exam-focused was their perceived ambition for success in the exam.

5. The other factor that influenced students' practice to be exam-targeted was the pressure they felt in their learning for score-gain from parents and school administrators

According to Melkamu Abate, (2007, pp: 76-77.), the results of the data collected through these different approaches and instruments have been analyzed. After the necessary data analysis, the investigator came up with the following major findings:

1. The reactions of the Grade 10 students and their English language teachers to the EGSEC English Language Examinations were positive.
2. The Grade 10 students and their English language teachers welcomed favorably the role and/or effect of the examination in the students' learning.
3. There is a perceived wash-back effect of certain aspects (format. . .) of the EGSEC English Language Examination on the students' learning by the Grade 10 English language teachers.
4. There is a perceived wash-back effect of certain aspects (formant. . .) of the EGSEC English Language Examinations on the students' learning by the Grade 10 students.
5. The Grade 10 EGSEC English Language Examinations were found to have inadequate coverage of the contents (objectives) of the courses.
6. High proportion of the examinations items was found to be relevant to few most important objectives of the program.
7. Only low proportion of the items contained in the examinations were found pertinent in measuring the program's highest priority objectives, Melkamu Abate, (2007, pp: 76-77.)

Fromse as Education and Training /FEAT/, (2015. pp:133-136), had reached on an investigation after conducting a study on language testing.

1. It is difficult to assess the extent of progress against targets for numbers of beneficiaries, as these were not set by the project.
2. The project has made a good start on developing useful test items for teachers and with teachers.

3. Teachers seem to know what is right to be done in terms of pedagogy and there seems to be some positive efforts by some teachers to implement promising practices in the classroom, but overall the uptake is low and does not seem to be fully embedded.
4. Teachers were found to demonstrate quite a high understanding of the theory of continuous assessment, the role it plans in the learning process and the valid reasons behind the approach, but the uptake of the actual practice is still low.
5. Learner scores are still relatively low in many subjects and wide variations were found to exist within the target grades. There is little difference between students' performance between comparison and program schools and from the baseline data.
6. Learner scores are still relatively low in many subjects and wide variations were found to exist within the target grades.

CHAPTER THREE: RESEARCH METHODOLOGY

3.0. Introduction

The objective of this study was to compare the validity of the VUEE and VCT, and their LA at ***Birbirsa*** Secondary School (grade 12 in focus). In order to achieve this objective designing and use of methods and methodic (methodologies) were mandatory. So, this chapter was designed in consisting of the research design and methods used in carrying out this study, the subject of the study, sample selection and sampling techniques, instruments used for gathering data, data collection procedures, techniques of data analysis, validity of instruments, reliability of instruments and ethical issues.

3.1. Research Design

The main purpose of this study was to compare the validity of the VUEE and VCT, and their LA at ***Birbirsa*** Secondary School (grade 12 in focus).

Comparative research differs from non-comparative work in that it attempts to reach conclusions beyond single cases and explains *differences and similarities* between objects of analysis and relations between objects against the backdrop of their contextual conditions. In addition to these general benefits, comparison also has specific scientific advantages. …to fully exploit these benefits, it is essential that the objects of analysis are compared on the basis of a common theoretical framework and that this is performed by drawing on equivalent conceptualizations and methods.…no comparative research without an extensive theoretical argument underlying it, or without a methodologically adequate research design to undertake it, Frank Esser (n.d. p: 1-2). In this case theoretically proved procedures for test validity evaluation were used for both the CT and the UEE to be compared with, and the statistical value gained through comparison with theoretical arguments underlying the TV were compared to examine their LA.

In order to investigate the relationship (similarity and difference), between the VUEE and VCT the comparative research method was used. This method was employed to compare the content,

concurrent, predictive and face validity of the CT, and the concurrent and face validity of the UEE.

3.2. Sources of Data and Subjects of the Study

In order to respond for this study's basic research questions, data was collected from primary and secondary sources. Two males and two females grade 9-12 English language teachers, 18 males and 15 females grade 12 students in the academic year 2020 were the subject of the study, from where the sample was taken in order to reduce the number into manageable level.

3.3. Sample Size and Sampling Techniques

3.3.1. Sample Size

It should be ensured in the sampling process itself that the sample selected is representative of the population, Flick, (2002). Based on this theory the researcher has decided to sample the subjects of the study. Data was collected from the total population of the study. From the total of 33 students in grade 12 (100%) or 33 students in 2 sections were taken as a sample to be studied. From the total of 4 English teachers at this school all of them whom are the grade 9-12 English language teachers (100%) were used as a sample. The reason why the researcher was intended to use these populations was because of these populations are decisive to be studied and to be taken as a sample to achieve the objective of the study. The following table shows the respondents that the researcher has been selected from the total population of the research in the universe of the study.

Table 1: ***Sample size which was selected from the total population***

No	Respondents	Total Population in Number	Sample Size in Number	Sample Size in Percent (%)
1	Grade 12 Students	33	33	100%
2	Grade 9-12 English Language Teachers	4	4	100%
	Total	37	37	100%

3.3.2. Sampling Techniques

In case of teacher respondents, 4 (100%) English language teachers from 4 total teachers at ***Birbirsa*** Secondary School were selected in comprehensive sampling technique. From the 33 grade 12 total students at ***Birbirsa*** Secondary School (100%) of them were selected in comprehensive sampling method to be admitted for the questionnaire. Grade 12 UEE and CT of the year 2020 were sampled as a data source, through the use of purposive sampling technique. The reason why these two exams were taken as a sample is: due to the second semester final exam (classroom tests) by this year were rejected due to the existence and wide spread of the CORONA-VIRUS, and the 2021 UEE was extended to November, 2021, but due to the study was conducted earlier to the 2021 UEE, the researcher found these two exams/tests very recent to be used. The other documents like a grade 12 syllabus, UEE result reporting document and CT average result reporting document were also purposely sampled.

3.4. Data Collection Instruments

With the regard of data collection instruments, Dowson (2007:15) says "Research methods are the tools you use to gather your data." In the current study four instruments were used. These tools were: document analysis, content analysis, interview and questionnaire. These instruments were considered to be important to triangulate the data and/or to examine and compare the results.

3.4.1. Questionnaire

Dowson (2007:62) indicated that the questionnaire is the most widely used type of instrument in educational research. The basic tool that the researcher used to collect data from the respondents was questionnaire. In case of collecting evidences for ***face validity*** of the tests, ***closed-ended questionnaire was*** used. These questions were taken from Brown, 2004.p.27. In this questionnaire the ***Likert-scale*** (1=strongly disagree, 2=disagree, 3= undecided, 4=agree, 5= strongly agree) was used to collect and present the data to check for the school level test and entrance exam face validity, and Brown, 2004, scale range which implies as 4.5 5 = very high; 3.5-4.4. = high; 2.5-3.4 = moderate; 1.5-2.4 = low, and 1-1.4 = low (never) was used for

analyzing and uncovering the trends in the data set. This enabled the researcher to partially achieve the three objectives of the study**.**

3.4.2. Document Analysis

In order to supplement the data that will be collected through observations, survey questionnaires, interviews and checklist, …relevant documents will analyzed in the study as Alderson and Wall, (1993:115-129), have recommended. In this case, five types of documents were collected and analyzed. They were grades 12 syllabus for English Language, classroom teacher made of the year 2020 first semester test, the 2020 UEE sheet, a mark list of CT average results and UEE results reporting document.

Grade 12 syllabus for English Language was collected and analyzed for the purpose of counting the course content specification objectives and figure out them under one of the six objective types: ***knowledge, comprehension, application, analysis, synthesis and evaluation objectives*** to see the extent to which the content of the test represents the course objectives. The CT was collected to check the content validity and face validity of the test. The UEE sheet was collected to check its face validity based on the teachers subjective judgments use it as evidence. The students' CT average result reporting document and UEE result reporting document were collected to check the concurrent validity of both tests (CT and UEE), and predictive validity of the CT. The predictive validity of UEE was not checked due to prediction is future oriented, and it may require gathering evidences of exams after taking UEE at University level students achievement, and this was not the concern of the researcher.

3.4.3 Content Analysis

Content analysis is a research tool used to determine the presence of certain words or concepts within texts or sets of texts. Researchers quantify and analyze the presence, meanings and relationships of such words and concepts, then make inferences about the messages within the texts, the writer(s), the audience, and even the culture and time of which these are a part. To conduct a content analysis on any such text, the text is coded or broken down, into manageable categories on a variety of levels word, word sense, phrase, sentence, or theme and then examined using one of content analysis' basic methods: conceptual analysis or relational analysis,

Palmquist's, 1990. In order to check for the CT content validity, instructional objectives specification and test content specifications were prepared and the number of objectives of the syllabus and the test objectives were analyzed and figured under six columns on separate sheets. The categories of the objectives are: knowledge, comprehension, application, analysis, synthesis and evaluation; then the numbers were correlated to each other in a two by two correlation of the variables to check the content validity of the classroom test.

The content validity of UEE was not checked, due to it incorporates different grade levels, and this makes the correlation impossible to relate the variables with a grade 12 course objective specification.

3.4.4. Interview

This data collection instrument was used to get information directly from participants of the research. Therefore, the researcher was used the structured interview so as to get information from teachers. Four teachers were interviewed. The reason why the researcher was used structured interview for teachers was that it has an interview guide that would be asked during conversation, it has open- ended questions and discussion, and respondents (teachers) had freedom to express their view in their own terms.

3.5. Data Collection Procedure

There are five steps in the process of quantitative data collection. This process involves more than simply gathering information; it includes interrelated steps. It involves the steps of determining the participants to study, obtaining permissions needed from several individuals and organizations, considering what types of information to collect from several sources available to the quantitative research, locating and selecting instruments to use that will net useful data for the study, and finally, administering the data collection process to collect data, Creswel, 2011: 137. The researcher has considered these steps, and has taken them in account during the data collection.

3.6. Methods of Data Analysis

Creswel, 2011:177, stated that the following ideas on the methods of data analysis. He said that, after you prepare and organize the data, you are ready to analyze it. You analyze the data to address each one of your research questions or hypotheses. Questions or hypotheses in quantitative research require that you compare two or more groups on the independent variable in terms of the dependent variable.

Regarding the document analysis, the students' result of the UEE and the result of the CT were collected and compared to one another to compare the CT concurrent and predictive validity, and to compare the concurrent validity of the UEE. At the end the concurrent validity of the UEE and the CT were compared to each other to examine their relationship.

The content analysis was used to analyze the contents of the syllabus. In this regard the course/instructional objectives in the syllabus and the test objectives specification table which is adapted from "Reader on students assessment", EQUIP November 2008, p.16 was used to specify the objectives. After the objectives were specified and summed, the sum of the test objectives specification were compared with the sum of instructional objectives to compare the content validity of the CT.

In order to analyze the questionnaire data, mean analyses were used to compare the face validity of UEE and the CT.

The interview responses were qualitatively analyzed to support the claims in quantitative analysis.

There are several interrelated steps used in the process of analyzing quantitative data. The first step is to *prepare the data for analysis*. This involves determining how to assign numeric scores to the data, assessing the types of scores to use, selecting a statistical program, and inputting the data into a program, and analyzing. The second step *begins the data analysis*. Typically you conduct a descriptive analysis of the data reporting measures of central tendency and variation.

Then you *conduct more sophisticated inferential analysis* to test hypotheses and you *examine confidence intervals and effect sizes*. The next step is to *report the results* that are found using tables, figures, and a discussion of the key results. Finally, you *interpret the results from the data analysis*. This consists of summarizing the results, comparing the results with past literature and theories, advancing the limitations of the study, and ending with suggestions for future research.

All the above steps were implemented to conduct a valid and reliable data analysis. The explanatory and comparative data analyses were used to analyze the collected data.

3.7. Validity of Instruments

According to Johns Hopkins, 2007, validity is the degree to which any measurement approach or instrument succeeds in describing or quantifying what it is designed to measure. Validity reflects those errors in measurement that are systematic or constant.

Measurement is accepted by those concerned as being logical (face validity also expert validity). The items included in the measure should adequately represent the universe of questions that could have been asked. The new measure should agree with an external criterion, e.g., an accepted measure (criterion-related validity). The measure should be consistent with the theoretical concept being measured (construct validity).

Factors to be considered to validate the instruments:

- How much time and money do you have to carry out your own tests?
- How small a difference in the measurement do you expect?;
- Can you use a previously validated measure? and
- Does the previous measure work within the context of your setting?

All tests of validity ultimately designed to support/refute the instruments were given due emphasis to ensure the validity of the instruments.

3.8. Reliability of the Instruments

According to Johns Hopkins, 2007, Reliability of a Measure is the degree to which a measurement technique can be depended upon to secure consistent results upon repeated application. Wiersita (1986), Komb and Tromp (2006), have also defined reliability refers to the consistency of research and the extent to which studies can be replicated. An instrument is reliable when it can measure a variable accurately and consistently and obtain the same result under the same condition over a period of time. In order to achieve the reliability of the instruments, the instruments' reliability were checked through pilot test (replicated use of instruments).

3.9. Ethical Issues

In conducting this study, emphasis was given to every important ethical issue. First, before entering into the actual data collection, a formal letter was received from ***Werada*** Education office. Then, the letter was given to the school principals by the researcher and good rapport on their agreement was created at the same time. In addition, people were participated with their full permission. Every effort was made to keep participants anonymous and confidentiality. Moreover, every source that was used in this study were acknowledged.

The researcher has kept all the secrets which were gained during the investigation of the study and it announced the result only when it is necessary. The researcher also kept/protected the right of the target group/the subject of the study to know the objective of the study and keep their interest whether they want to participate in the study or not. The final aim of the research purpose was not to blame the people who participate in the study. In general the researcher worked in harmony/agreement with the participants of the study.

CHAPTER FOUR: DATA PRESENTATION, ANALYSIS AND INTERPRETATION

4.0. Introduction

Under this chapter, the data gathered from the field of study were presented dominantly in tabular format, rather than the responses for the interview questions. The analysed data, the way in which it was analysed, and the interpretation of the analysed data was presented under this chapter.

4.1. The Validity of Classroom Tests

In order to respond for the first basic research question: "To what extent, are the CT valid when it is examined in comparing with external standards?", and achieve the first specific objective of the study, data was collected from the field of study by using *document analysis*, *content analysis and questionnaire*.

Data gathered for examining the content, concurrent, predictive and face validity of CT comparatively were quantitative data.

4.1.1. The Content Validity of Classroom Tasts

Content validity includes any validity strategies that focus on the content of the test. To demonstrate content validity, testers investigate the degree to which a test is a representative sample of the content of whatever objectives or specifications the test was originally designed to measure, Sireci, 1998.

In order to verify how far the CT represents the instructional objectives of the syllabus; the total number of instructional objectives in the syllabus were classified under sub categories of skills by using the table which is adapted from "Reader on students assessment", EQUIP November 2008, p.16 (table 2). Table 3 indicates the total number of the CT objectives specification and the category of each item and what type of learning objectives the tests intend to achieve were presented in table of tests content specification. Table 4 shows how far the test content represents the intended instructional objectives to be achieved. This was analyzed in percentage in each category and it also contains the total/cumulative percentage.

Table 2: Table of Instructional Objectives Specification

No.	Content	General Objectives						
		Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
1	Unit 1	3	3	2	2	2	1	13
2	Unit 2	0	4	4	3	3	2	16
3	Unit 3	2	2	4	3	2	2	15
4	Unit 4	2	2	2	2	1	2	11
5	Unit 5	4	2	3	2	2	2	15
6	Unit 6	4	2	3	2	2	3	16
	Total	15	15	18	14	12	12	86

Table 3: Table of Tests Content Specification

No.	Content	General Objectives						
		Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
1	Unit 1							0
2	Unit 2							0
3	Unit 3							0
4	Unit 4							0
5	Unit 5							0
6	Unit 6							0
	Total	10	10	6	11	2	1	40

Table 4: Analysis of Classroom Tests Content Validity

No.	Content	General Objectives						
		Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
1	Unit 1							0
2	Unit 2							0
3	Unit 3							0
4	Unit 4							0
5	Unit 5							0
6	Unit 6							0
	Total	66.666667	66.66666667	33.333333	78.5714	16.66667	8.3333333	46.51

The above data shows: as the *analysis-related* objectives were covered *78.57%* and primarily paid attention than the other objectives in being covered by the CT; as the *knowledge* related objectives were covered *66.6%;* the *comprehension-related* objectives were covered *66.6%*; the *application-related* objectives were *33.3%* covered; the *synthesis-related* objectives were covered *16.6%; and evaluation-related* objectives were *8.3%* covered by the CT. This implies that three of the general objectives of the course (analysis-related, knowledge-related and comprehension-related objectives) were covered by the tests in above average; and the remaining three of the general objectives specification (application-related, synthesis related and evaluation-related objectives) were covered in below average. The total result (the degree to which the *overall* objectives of the course are covered by the test) is *46.5%*.

The extents to which the CT are prepared in line with the instructional objectives were significantly varied category to category. Some of the instructional objectives of the course like: *analysis-related, knowledge-related and comprehension-related objectives* were represented in the test in 78.56%, 66.6%, 66.6% respectively, so this implies that as these instructional objectives were represented above average by the tests. The *application, synthesis and evaluation related objectives* were represented in the CT in 33.3%, 16.6%, 8.3% respectively, so these three objectives were represented below average in CT.

In general term, the CT contents at ***Birbirsa*** Secondary School represented the objectives of the syllabus in 46.51% and this is below average. So, the CT content was poorer in being representative sample of the instructional objectives.

4.1.2. Criterion-related Validity of the Classroom and University Entrance Exam

Criterion-related validity evidence refers to the evidence that is obtained by correlating *test scores* with an external standard or criterion. Criterion-related evidence is classified in to two kinds: *concurrent validity* and *predictive validity*. A test holds *concurrent validity* evidence if it is administered along with the established/well known/traditional test to a group of students. …the scores of the two tests, the new and the established one, are correlated to determine the relationship. This is done to obtain empirical evidence about their relationship, Brown, 2004, p. 24.

A. Concurrent Validity of Classroom Tests and Entrance Exam

The classroom test results and the UEE results were collected from the source to compare the two results gained by the same students in order to decide on the CT concurrent validity. The empirical evidence shows the relationship between the two tests. Below there is a table which shows the scores in CT and scores in UEE and their rank.

Table 5: ***The Analyzed Data for Comparing Concurrent Validity Evidence***

No.	Name	Scores on UEE out of 100%	Rank in UEE Result	Scores on CT Out of 100%	Rank in CT Result	The variation of CT Result from Scores in UEE	Sum of the two 200%	Av. Of the two 100%
1	EZB	40	1	99	1	-59	139	69.5
2	YGM	35	2	84.5	3	-49.5	119.5	59.75
3	TAM	35	3	81.5	4	-46.5	116.5	58.25
4	KEG	34	4	70.5	8	-36.5	104.5	52.25
5	CIM	33	5	74.5	7	-41.5	107.5	53.75
6	KKA	33	6	65.5	12	-32.5	98.5	49.25
7	EWF	32	7	59.5	19	-27.5	91.5	45.75
8	LKL	31	8	51.5	32	-20.5	82.5	41.25
9	MTG	31	9	62	13	-31	93	46.5
10	MSG	30	10	56.5	25	-26.5	86.5	43.25
11	BAL	29	11	67	9	-38	96	48
12	TSM	29	12	77	5	-48	106	53
13	BAG	29	13	52	31	-23	81	40.5
14	HTG	28	14	60	18	-32	88	44
15	KYJ	27	15	89.5	2	-62.5	116.5	58.25
16	BCB	27	16	58	24	-31	85	42.5
17	TAM	27	17	54	30	-27	81	40.5
18	UTD	26	18	67	10	-41	93	46.5
19	DTD	25	19	56	26	-31	81	40.5
20	TKG	24	20	62	13	-38	86	43
21	DTM	23	21	55.5	27	-32.5	78.5	39.25
22	TYG	23	22	59	20	-36	82	41
23	YTA	23	23	61	16	-38	84	42
24	LNM	23	24	55.5	27	-32.5	78.5	39.25
25	SAD	22	25	58.5	22	-36.5	80.5	40.25
26	TTD	21	26	50.5	33	-29.5	71.5	35.75
27	YKG	20	27	61.5	15	-41.5	81.5	40.75

28	BBL	20	28	76.5	6	-56.5	96.5	48.25
29	SBK	20	29	61	16	-41	81	40.5
30	GHM	20	30	59	20	-39	79	39.5
31	GAI	19	31	68	9	-49	87	43.5
32	NMH	15	32	58.5	22	-43.5	73.5	36.75
33	NKM	15	33	55.5	27	-40.5	70.5	35.25
	Total	869		2127.5		-1258.5	2996.5	1498
	Class Av.	26.33		64.46		-38.13	90.8	45.4

The above table (table 5) shows: the scores that the students have scored in UEE, their rank in UEE result, the score they gained in CT and their rank in CT result, the UEE result deviation from the CT result, sum of the two exams and average of the two tests.

The comparison of the two results shows deviation. All the students have scored higher in their CT; but in UEE they scored poorer; The result of CT deviates from the UEE result which ranges from *-24*the lowest to-*62.5* the highest; In average the result of CT of grade 12 students at *Birbirsa* Secondary School deviates in average of *38.13%;* The student who stands/comes first have scored *99* in CT and ranked 1^{st} but in UEE he scored *40* and ranked 5^{th}*;* The students who ranked last in UEE by scoring 15 has scored *55.5* and ranked *27* out of *33* students in CT result; The average result that the students gained in the CT is *64.46;* The average result the students gained in UEE is 26.3 when compared as a whole class, and average of the two tests to the whole class is *45.4%*.

The tests (both the CT and UEE) were given to the same group of the students, but when the results of the two tests are compared they have no relationship with each other in terms of rank of the students and the result they gained with both tests were highly deviating. In specific terms comparison of the two tests shows as: all the students have scored higher in their CT and lower in UEE; the result of CT deviates from the UEE *positively* and in reverse the UEE result deviates *negatively* from the CT result; the average deviation of the result of the CT from the UEE as a whole is very high (*38.13%);* the rank of the students is significantly vary between the two tests, (all the students CT rank is different from UEE rank except the first student; the average result that the students gained in the CT is more than the average, which is *64.46%* and this implies as all students have scored beyond a pass mark in classroom tests; the average result the students gained in UEE is lower, which is (*26.3%)* when compared as a whole class and this shows as all

the students have performed poorly and scored below a pass mark, and average of the two tests to the whole class is *45.4%*, and this implies that as the whole students are not successful within average result of the two tests as a whole, and in turn this implies that as students are not successful in both the CT and UEE how far they all scored beyond 50% and succeeded in CT.

In general terms, the data gathered and analyzed to respond for to what extent the CT has *criterion-related validity evidence (concurrent validity)* implies as the CT and UEE have no concurrent validity evidence.

B. Predictive Validity of Classroom Tests

Predictive validity evidence refers to how well the test predicts the taste takers future performance. This form of validity evidence is required for the educational instructions that give entrance exam for the candidates who want admission to colleges or Universities. To determine the predictive validity evidence of a test, one has to administer the taste to a group of subjects and correlate the results with the tests' results which the subject has taken after a certain period of time. Predictive validity was estimated by correlating the scores of the midterm achievement tests with those of the final exam, as well as by correlating the final exam scores, Kukuk Walters, 2009.To determine the ability of classroom test to predict success after the test, the scores and final grades of the same students were obtained and correlated with their final exam scores. Brown, 2004.

In order to check for the predictive validity evidence of the grade 12 English Language CT by comparing with the UEE data is gathered from the field of study by using document analysis. In order to see how far the CT predict a success in UEE, the two exams results were compared.

As shown above in *table 5* the CT result doesn't predict the result in UEE. The result shows: All the students have not scored the same or nearly the same results in CT and in UEE results; The average result of the CT is 64.4*%* and the average result of the UEE result is *26.3%*, and this shows a higher disparity; This comparison of the two exams shows as there is a disparity between the two exams individual students' scores and average score of the whole class.

In general as the above data implies the CT have no predictive validity evidence, since it has not predicted the future performance of the taste takers. In other terms the result and rank of the students in CT is not in line/in congruence with the result and the rank of the students in UEE.

4.1.3. Face Validity of the Classroom Tests

According to Brown, 2004,p.27, face validity refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers. The face validity of the tests determined not only by students but also by other stakeholders. In fact of all the stakeholders, students would like to express their opinions about the taste taken. From the students' perspective to what extent the test does its job or is the test easy or difficult? Does it look realistic or trivial? The students' responses to such questions will help to decide the face validity of a test. Face validity is determined by asking both the instructors and students about how well they felt the content of the course were represented on the achievement tests, Brown, 2004,p.27.

In having its root in this theory the researcher have collected evidences from the field of study to check how far the classroom Test looks right, and appears to measure the knowledge or abilities it claims to measure. In order to collect the subjective judgment of the students on the school level test and the subjective judgment of the teachers on entrance exam the following data on table 6 was collected for validation of both tests. The questions for checking the test's face validity was taken from Brown, 2004.p.27.

Table 6: Students *Questionnaire for checking face validity of the classroom tests*

No.	Questions	Total No. of Respondents	Very High (5)	Total = No. of ss * 5	High (4)	Total = No. of ss * 4	Moderate (3)	Total = No. of ss * 3	Low (2)	Total = No. of ss * 2	Never (1)	Total = No. of ss * 1	Sum = the aggregate of totals	Mean = sum/81
1	The content of the main course book was represented in the tests and exams sufficiently.	33	2	10	16	64	15	45	0	0	0	0	119	3.6060606
2	The content of the grammar book was represented in the test sufficiently.	33	7	35	11	44	11	33	4	8	0	0	120	3.6363636
3	The content of the writing courses was represented in the test sufficiently.	33	0	0	7	28	15	45	4	8	7	7	88	2.6666667
4	The content of the reading courses was represented in the test sufficiently.	33	0	0	0	0	20	60	5	10	8	8	78	2.3636364
5	The content of the speaking course was represented in the test sufficiently.	33	0	0	2	8	4	12	6	12	21	21	53	1.6060606
6	Grammar taught in the courses was represented in the test sufficiently.	33	13	65	8	32	12	36	0	0	0	0	133	4.030303
7	The vocabulary taught in course was represented in the test sufficiently.	33	22	110	7	28	4	12	0	0	0	0	150	4.5454545
8	The listening practices focused on in the courses were represented in the test sufficiently.	33	0	0	12	48	11	33	8	16	2	2	99	3
9	The exercises made in the courses were represented in the test sufficiently.	33	13	65	4	16	16	48	0	0	0	0	129	3.9090909
10	The content of the laboratory courses was represented in the test sufficiently.	33	6	30	18	72	9	27	0	0	0	0	129	3.9090909
11	In general, the contents of the courses were represented in the test sufficiently.	33	0	0	9	36	12	36	12	24	0	0	96	2.9090909

As indicated in the above table 6 the subjective judgment of the students were collected on 11 test validity checking questions. The students rate 1-5 in order to show how far the test looks right, and appears to measure the knowledge or abilities it claims to measure.

According to Brown, 2004, scale range 4.5 5 = very high; 3.5-4.4. = high; 2.5-3.4 = moderate; 1.5-2.4 = low, and 1-1.4 = low (never). The mean of their rating was used to decide on in which category their rating is falling.

As the analyzed data in the above table shows: in checking whether content of the main course book was represented in the test sufficiently, the mean of the students is *3.6*, and this is categorized in *high* ; in checking whether content of the grammar book was represented in the test sufficiently, the mean of the students is *3.6*, and this is categorized in *high*; in checking whether the content of the writing courses was represented in the test sufficiently., the mean of the students is *2.6*, and this is categorized in *moderate*; in checking whether the content of the reading courses were represented in the test sufficiently.., the mean of the students is *2.3*, and this is categorized in *low*; in checking whether the content of the speaking course was represented in the test sufficiently, the mean of the students is *1.6*, and this is categorized in *low*; in checking whether the Grammar taught in the courses was represented in the test sufficiently., the mean of the students is *4*, and this is categorized in *high*; in checking whether the vocabulary taught in course was represented in the tests sufficiently., the mean of the students is *4.5*, and this is categorized in *very high*; in checking whether the listening practices focused on in the courses were represented in the test sufficiently, the mean of the students is *3*, and this is categorized in *moderate*; in checking whether the exercises made in the courses were represented in the tests sufficiently, the mean of the students is *3.9*, and this is categorized in *high*; in checking whether the content of the laboratory courses were represented in the tests sufficiently., the mean of the students is *3.9*, and this is categorized in *high*, and in checking whether the in general, the contents of the courses were represented in the tests sufficiently., the mean of the students is *2.9*, and this is categorized in *moderate*.

The interpretation of the above data implies that: as the content of the main course book was highly represented in the tests; as the content of the grammar book was highly represented in the tests; as the content of the writing courses were moderately represented in the test; as the content

of the reading courses was lowly represented in the test; as the content of the speaking course was lowly represented in the test; as the Grammar taught in the courses was highly represented in the test; as the vocabulary taught in course was highly represented in the test; as the listening practices focused on in the courses were lowly represented in the test; as the exercises made in the courses were highly represented in the test; as the content of the laboratory courses was highly represented in the test, and generally, as the contents of the courses were lowly represented in the test.

4.2. Validity of University Entrance Exam

In order to check to what extent the UEE measures what is intended to be measured and how far it looks right in order to respond for the question "To what extent the UEE valid when checked" and achieve the objective "To check the validity of UEE comparatively with the theoretical frameworks", data was gathered from the universe of the study by using a table taken from Brown, 2004.p.27. The data were collected from 4 ***Birbirsa*** Secondary School English Language Teachers. In case of UEE only the face validity of the exam was checked.

4.2.1. Face Validity of University Entrance Exam

The researcher have collected data from the field of study to check how far the UEE looks right, and appears to measure the knowledge or abilities it claims to measure. In order to collect the ratings and subjective judgment of the teachers through the use of close ended questions and analyzed quantitatively. Here is the evidences collected in ratings by the use of Brown's table of test validation criteria's.

Table 7: ***Questionnaire for teachers on face validity of university entrance exam***

No.	Questions	Total No. of Respondents	Very High (5)	Total = No. of teachers * 5	High (4)	Total = No. of teachers * 4	Moderate (3)	Total = No. of teachers * 3	Low (2)	Total = No. of teachers * 2	Never (1)	Total = No. of teachers * 1	Sum = the aggregate of totals	Mean = sum/4
1	The content of the main course book was represented in the tests and exams sufficiently.	4	1	5	2	8	1	3	0	0	0	0	16	4
2	The content of the grammar book was represented in the test sufficiently.	4	1	5	2	8	1	3	0	0	0	0	16	4
3	The content of the writing courses was represented in the test sufficiently.	4	1	5	0	0	0	0	0	0	3	3	8	2
4	The content of the reading courses was represented in the test sufficiently.	4	2	10	1	4	1	3	0	0	0	0	17	4.25
5	The content of the speaking course was represented in the test sufficiently.	4	0	0	1	4	1	3	1	2	2	2	11	2.75
6	Grammar taught in the courses was represented in the test sufficiently.	4	1	5	2	8	0	0	1	2	0	0	15	3.75
7	The vocabulary taught in course was represented in the test sufficiently.	4	0	0	0	0	3	9	1	2	0	0	11	2.75
8	The listening practices focused on in the courses were represented in the test sufficiently.	4	0	0	0	0	0	0	0	0	4	4	4	1
9	The exercises made in the courses were represented in the test sufficiently.	4	0	0	0	0	1	3	3	6	0	0	9	2.25
10	The content of the laboratory courses was represented in the test sufficiently.	4	0	0	0	0	0	0	4	8	0	0	8	2
11	In general, the contents of the courses were represented in the test sufficiently.	4	0	0	0	0	1	3	3	6	0	0	9	2.25

The above data (table 7) has presented the subjective ratings of the teachers to the questions for checking face validity of the 2020 UEE. In case of this study the mean the teachers' ratings were used to categorize the ratings under one of the five intervals of Browns' scale interval. Based on the data in table 7 the analyzed data implies: in checking whether content of the main course book was represented in the UEE sufficiently, the mean of the teachers is *4*, and this is categorized in *high*; in checking whether content of the grammar book was represented in the UEE sufficiently, the mean of the teachers is *4*, and this is categorized in *high*; in checking whether the content of the writing courses was represented in the UEE sufficiently, the mean of the teachers is *2*, and this is categorized in *low*; in checking whether the content of the reading courses was represented in the UEE sufficiently.., the mean of the teachers is *4.25*, and this is categorized in *high*; in checking whether the content of the speaking course was represented in the UEE sufficiently, the mean of the teachers is *2.5*, and this is categorized in *moderate*; in checking whether the Grammar taught in the courses was represented in the UEE sufficiently., the mean of the teachers is *3.75*, and this is categorized in *high*; in checking whether the vocabulary taught in course was represented in the UEE sufficiently., the mean of the teachers is *2.75*, and this is categorized in *moderate*; in checking whether the listening practices focused on in the courses were represented in the UEE sufficiently, the mean of the teachers is *1*, and this is categorized in *never*; in checking whether the exercises made in the courses were represented in the UEE sufficiently, the mean of the teachers is *1.25*, and this is categorized in *low*; in checking whether the content of the laboratory courses was represented in the UEE sufficiently., the mean of the teachers is *2*, and this is categorized in *low*, and in checking whether the in general, the contents of the courses were represented in the UEE sufficiently., the mean of the teachers is *2.25*, and this is categorized in *low*.

The interpretation of the above data implies that: as the content of the main course book was highly represented in the UEE; as the content of the grammar book was highly represented in the UEE; as the content of the writing courses was lowly represented; as the content of the reading courses was highly represented; as the content of the speaking course was lowly represented; as the grammar taught in the courses was highly represented; as the vocabulary taught in course

was moderately represented; as the listening practices focused on in the courses were never represented; as the exercises made in the courses were lowly represented; as the content of the laboratory courses was lowly represented in the UEE, and generally, as the contents of the courses were highly represented in the UEE.

4.2.1. The University Entrance Exam's Face Validity

In order to check for the face validity of grade 12 UEE, data was gathered from the teachers who have comprehensively selected and checked the exam due to the students left the school as immediate as they took the exam and difficult and even impossible to access them back to gather data from. This is why the researcher collected this data from the teachers only. The evidences or personal judgments that the teachers provided are the following.

The content of the main course book was moderately represented in the UEE. The exam focuses on both grade 11 and 12, and even 9 and 10. Some of the contents are also extracted from out of these grade levels.

The content of the grammar book was represented in the UEE. The grammatical contents included are not a representative sample of those in the student's text book. Most of the grammatical contents were taken from somewhere out of the text. They are also beyond the students' level of understanding.

The content of the writing courses were not represented in the university entrance exam. However some exercises were given in teachers guide, and the students are given some project works of writing to deal with the writing skill.

The content of the reading courses was represented in the UEE, but it seems overloaded when compared with the student's language skill. The time allotted for accomplishing the reading/comprehension questions is shorter, and the students assign answer to the reading questions randomly in our experience.

The vocabulary taught in course was not represented in the UEE. The vocabularies are taken from some materials with which the students are not familiar with. The students face with challenges in responding for the vocabulary questions.

The UEE concentrates on the two language skills: reading and grammatical aspect, and left the remaining three skills: writing, speaking and listening.

Based on the result the students gained/scored as a result of the exam have not represented the language skill of the students. It only serves to measure/evaluate the learners in order to assign with grades and certify their success to be used in their future higher education joining, or use it in their work life.

4.3. The Classroom Tests and the University Entrance Exam Level of Accuracy

Accuracy, as a technical term, is the state of being precise or correct according to a traceable reference standard. It is always computed as a statistical process based on known and defined units. Assessments are accurate when they measure what they purport to measure, Cronbach, L. J. (1949). Depending on this the validity of the CT and the UEE were examined based on theoretical arguments and their LA is computed by using statistical process.

4.3.1. Face Validity of Classroom Test and Entrance Exam Comparatively

Since checking the VUEE is delimited at only checking its face validity and data is collected only on its face validity, the face validity of UEE and the face validity of CT were compared.
Comparative research attempts to reach on conclusions beyond single cases and explains ***differences and similarities*** between objects of analysis and relations between objects.… Frank Esser (n.d:1-2). In this case the data collected by using the Brown, 2004 table for face validation is used to compare LA of both tests.

Table 8: ***The Mean of Teachers and Students Ratings on University Entrance Exam and Classroom Tests***

No.	Questions	Total No. Students Respondents	Total No. Teachers Respondents	Mean of the students on CT	Mean of the teachers on UEE
1	The content of the main course book was represented in the tests and exams sufficiently.	33	4	3.606060606	4
2	The content of the grammar book was represented in the test sufficiently.	33	4	3.636363636	4
3	The content of the writing courses was represented in the test sufficiently.	33	4	2.666666667	2
4	The content of the reading courses was represented in the test sufficiently.	33	4	2.363636364	4.25
5	The content of the speaking course was represented in the test sufficiently.	33	4	1.606060606	2.75
6	Grammar taught in the courses was represented in the test sufficiently.	33	4	4.03030303	3.75
7	The vocabulary taught in course was represented in the test sufficiently.	33	4	4.545454545	2.75
8	The listening practices focused on in the courses were represented in the test sufficiently.	33	4	3	1
9	The exercises made in the courses were represented in the test sufficiently.	33	4	3.909090909	2.25
10	The content of the laboratory courses was represented in the test sufficiently.	33	4	3.909090909	2
11	In general, the contents of the courses were represented in the test sufficiently.	33	4	2.909090909	2.25
	Sum	363	44	36.18181818	31
	Average	*33*	*4*	*3.2892562*	*2.818181818*

Table 8 presented the relationship between the CT validation mean from the students rating, and the UEE rating mean of the teachers. This data serves to *compare the similarity and difference* between the CT and the UEE. The analysis of the collected data indicates: in checking whether content of the main course book was represented in the CT and in UEE sufficiently, the mean of the students is *3.6,* and the mean of the teachers' ratings is *4,*The students' rating is categorized in "*high*", and the teachers' rating is categorized in "*high*"; in checking whether content of the grammar book was represented in the UEE sufficiently, the mean of the students is *3.6,* and the mean of the teachers' ratings is *4,* both ratings are categorized in "*high*"; in checking whether the content of the writing courses was represented in the CT and in the UEE sufficiently., the mean of the students is *2.6,* and the mean of the teachers' ratings is *2.* The students' rating is categorized in "*moderate*", and the teachers' rating is categorized in "*low*"; in checking whether the content of the reading courses was represented in the CT and in UEE sufficiently.., the mean of the students is *2.3,* and the mean of the teachers' ratings is *4.25.* The students' rating is categorized in "*low*", and the teachers' rating is categorized in "*high*"; in checking whether the content of the speaking course was represented in the CT and in UEE sufficiently, the mean of the students is *1.6,* and the mean of the teachers' ratings is *2.5.* The students' rating is categorized in "*low*", and the teachers' rating is categorized in "*moderate*"; in checking whether the grammar taught in the courses was represented in the CT and in UEE sufficiently, the mean of the students is *4,* and the mean of the teachers' ratings is *3.75.* The students' rating is categorized in "*high*", and the teachers' rating is categorized in "*high*"; in checking whether the vocabulary taught in course was represented in the CT and in UEE sufficiently., the mean of the students is *4.5,* and the mean of the teachers' ratings is *2.75.* The students' rating is categorized in "very *high*", and the teachers' rating is categorized in "*moderate*"; in checking whether the listening practices focused on in the courses were represented in the CT and in UEE sufficiently, the mean of the students is *3,* and the mean of the teachers' ratings is *1.* The students' rating is categorized in "*moderate*", and the teachers' rating is categorized in "*low*"; in checking whether the exercises made in the courses were represented in the CT and in UEE sufficiently, the mean of the students is *3.9,* and the mean of the teachers' ratings is *2.25.* The students' rating is categorized in "*high*", and the teachers' rating is categorized in "*low*"; in checking whether the content of the laboratory courses was represented in the CT and in UEE sufficiently, the mean of

the students is *3.9*, and the mean of the teachers' ratings is *2*. The students' rating is categorized in "*high*", and the teachers' rating is categorized in "*low*"; in general, in checking whether the contents of the courses were represented in the CT and in UEE sufficiently., the mean of the students is 3.2 and the mean of the teachers' ratings is *2.8*. The students' rating is categorized in "*moderate*", and the teachers' rating is also categorized in "*moderate*".

The interpretation of the above data which is presented in table 8 shows: as the content of the main course book was *highly* represented in the CT and in UEE; as the content of the grammar book was *highly* represented in the CT and in UEE; as the content of the writing courses were *lowly* represented in the CT and in UEE; as the content of the reading courses were represented in the CT *moderately*, and in UEE *highly*; as the content of the speaking course was *moderately* represented in the CT and *lowly* in UEE. as the grammar taught in the courses was *highly* represented in the CT and in UEE; as the vocabulary taught in course was represented in the CT *highly*, and in UEE *lowly*. as the listening practices focused on in the courses were represented in the CT *highly*, and *lowly* represented in UEE; as the exercises made in the courses were represented in the CT *highly*, and in UEE *lowly*; as the content of the laboratory courses was represented in the CT *moderately*, and in UEE *lowly;* in general, the contents of the courses were represented in the CT and in UEE moderately.

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.0. Introduction

In this chapter, the summary, the conclusions, and the recommendations of the study are presented. In the first section, the participants of the study, the instruments employed the major findings of the data analyses and the discussions made on the findings are summarized. In the next part, the conclusions drawn from the findings follow. In the final section, the recommendations suggested are given for more advancement.

5.1. Summary

The main purpose of this study was to compare the validity of Classroom English Language Tests and University Entrance English Language Exam, and their LA. The study was conducted at ***Birbirsa*** Secondary School, which is found in ***Oromia*** Regional State, ***Ilu Aba Bora Zone, Nono Sele Woreda***. The tools used to collect the data were document analysis, content analysis, questionnaire for 33 students and 4 English Language teachers and interview.

Before using the tools for the study, the validity and reliability of the tools were checked through reviewers and through pilot study. The reviewers have reviewed the face and content validity of the questionnaires. Then, these tools together were pilot tested. The data collected were analyzed to check the validity and reliability of the questionnaires and to obtain lessons for the study.

To collect data for the main study, different numbers of participants were selected according to the types of tools. In order to compare the validity of the CT (content validity and criterion-related validity) document analysis was used. Three documents were collected in this regard: the 2020 UEE, and the 2020 First Semester CT, the grade/result report of the UEE, the grade/result report of the CT were collected, and in order to check and compare the *face validity* of the CT and the UEE questionnaire was used.

To collect data regarding teachers' subjective judgment on the 2020 UEE, 4 teachers were selected in comprehensive sampling and filled in the survey questionnaire to check and compare the *face validity* of both CT and UEE. Similarly, to gather data regarding students' subjective

judgment on the 2020 First Semester CT, 33 comprehensively selected students from the whole 33 students filled in the survey questionnaire. To compare how far the CT matched with the syllabus objectives, the objectives of the syllabus were counted and categorized under *six* instructional objectives specifications to check and compare for its *content Validity*. In the same way the test content specification is used to categorize the 42 total test items and compared with the syllabus objectives to compare the CT content validity. To check how far the CT have *criterion-related validity evidence (concurrent validity and predictive validity)* the scores of the 2020 grade 12 students in CT and in UEE was collected and their results were compared to decide up on their concurrent and predictive validity.

The data collected were analyzed quantitatively. The questionnaire, content analysis and document analysis data were analyzed quantitatively. The data collected through the use of questionnaire were analyzed using mean percentage to compare the face validity of the CT and the UEE, and later their LA was checked in comparison of the two means/averages. The documents were analyzed with percentage and rank to check and compare the concurrent and predictive validity of the CT and the concurrent validity of the UEE.

Discussions were made by comparing the findings of the quantitative data which was supplemented by qualitative data analyses so as to answer the created/posed research questions. The discussions of the findings were made mainly to check and compare the extent to which the CT and the UEE were valid, and to compare the two tests face and concurrent validity to decide up on their LA. The summary of the discussions of the findings was made in line with the answers to the research questions. The discussions made in relation to the answers for the basic research questions are summarized as follows.

In checking the CT *content validity*; the extent to which the CT were prepared in line with the course content objectives were significantly vary. Some of the objectives of the course like: analysis-related objectives were *78.5%* covered by the tests, knowledge-related objectives and comprehension-related objectives were covered *66.6%* and *66.6%* respectively and this shows as they are highly paid attention in contrast with the others. The application, synthesis and evaluation related objectives were covered *33.3%, 16.6%* and *8.3%* respectively, and this shows

as they were paid of the least attention in being incorporated in the test; especially, evaluation related content of the course were with very least attention in being included in the tests. In general terms the CT contents at *Birbirsa* Secondary School were lower in representing the instructional objectives (in its content validity), in relation to the relative importance given to the objectives, *46.5%* of the six instructional objectives specification were covered by the CT from the objectives in syllabus of the subject; so, *the CT hold below average content validity evidences.*

However both the CT and UEE were given to the same group of the students, the results of the two tests *have no relationship* to each other in terms of rank of the students and the result they gained with both tests. In specific terms comparison of the two tests shows as: all the students have scored higher in their CT which ranges from *50.5% - 99%,* and lower in UEE which ranges from *15% - 40%;* the highest result of the CT deviates from the UEE positively in *59% (99-40=59),* and in reverse the UEE result deviates negatively which is *-59% (40-99=-59)* from the CT result; the average deviation of the result of the CT from the UEE as a whole is *38.13%*, which is gained in subtracting the mean of the UEE from the mean of the CT (*64.46 – 26.33 = 38.13*), *so both tests have no concurrent validity*.

The ranks of the students were significantly vary between the two tests, (all the students CT rank and their UEE rank are different except the one student; the average result that the students gained in the CT is more than the average, which is *64.46%* and this implies as all students have scored beyond a pass mark in CT; the average result the students gained in UEE is lower, which is *(26.3%)* when the mean is calculated as a whole class and this shows as all the students have performed poorly and scored below a pass mark in UEE.

The CT *have no predictive validity evidence,* since it has not predicted the future performance of the taste takers. In other terms the result and rank of the students in CT is not in line/in congruence with the result and the rank of the students in UEE, so the CT have no predictive validity evidence.

In checking the face validity of the CT: the vocabulary taught in course was highly represented in the test; the content of the main course book, the content of the grammar book, the grammar

taught in the courses, the exercises made in the courses, and the content of the laboratory courses were moderately represented in the test; the content of the writing courses, the content of the reading courses, the listening practices focused on in the courses, and the contents of the course were lowly represented in the test; and the content of the speaking course was never represented in the test. The average/mean of the CT was *3.2* and this falls in "moderate", so *the CT were moderately face valid*.

In checking the face VUEE the content of the main course book, the content of the grammar book, the content of the reading courses, the grammar taught in the courses, and the contents of the courses were highly represented in the UEE; the content of the speaking course, the vocabulary taught in course, the exercises made in the courses, and the content of the laboratory courses; and the content of the writing course and the listening practices focused on in the courses were lowly represented in the UEE, The average/mean of the UEE is *2.8* and this falls in "moderate", so *the UEE was of moderate face validity*.

The comparison of LA of the CT face validity and the UEE face validity shows as the content of the main course book, the content of the grammar book, the grammar taught in the courses, and the contents of the courses, were highly represented in the CT and in UEE; and the content of the writing courses were lowly represented in the CT and in UEE. This investigation shows as the similarity between the two tests on their face validity. The difference between the two tests lies in: the content of the reading courses were represented in the CT moderately, and in UEE highly; the content of the speaking course was never represented in the CT and lowly in UEE in the form of dialogue; the vocabulary taught in the course was represented in the CT highly, and in UEE moderately; the listening practices focused on in the courses were represented in the CT moderately, and never represented in UEE; the exercises made in the courses were represented in the CT moderately, and in UEE lowly; and the content of the laboratory course was represented in the CT moderately, and in UEE lowly. The mean of the students ratings on the face VCT is 3.2 and the mean of the teachers' ratings on UEE is *2.8*, and this implies as the CT were a bit more accurate than the UEE. But in general term *both tests were moderately accurate in their face validity evidences.*

5.2. Conclusions

Based on the discussions of the findings of the study, it is possible to conclude that the CT and the UEE were both "low" in showing validity evidences; however the mean/average of the two tests were different they fall in the same interval in most cases. The conclusions of the findings were drawn in terms of the dimensions of the four test validity evidences. That is, *the conclusions were drawn in terms of content validity, concurrent validity, predictive validity and face validity.*

Firstly, in terms of the CT *content validity, the contents of the test were lower in representing the instructional objectives.* The test was not prepared in line with the relative importance given to the six course objectives namely knowledge-related objectives, application-related objectives, comprehension-related objectives, analysis-related objectives, synthesis related objectives and evaluation related objective.

Secondly, the students' scores are higher in their CT than in UEE. The best indicator of this statement is, all the students have passed in the CT result, but they have failed in UEE within a higher average result difference. There for the CT and the UEE *have no concurrent validity.*

Thirdly, the taste takers/students have gained the same or nearly similar results in CT and also nearly similar in UEE when the two exams were checked separately, and they stand/comes differently in rank in the two tests when their ranks are compared together. *The CT have no power of predicting the future performance of the students*; since the result and rank of the students vary throughout the class in both tests. So, *the CT have no predictive validity evidence.*

Fourthly, the classroom tests have represented the writing and reading skills in small scale and never represented speaking skills, however some of the language skills like: vocabulary, grammar, exercises are somewhat given more emphasis and represented in the tests as the mean of the questionnaire filled by the students indicated. So, *the CT were with moderate face validity evidence.*

Fifthly, the UEE represented the content, grammar and reading skills highly. Writing and listening courses were never represented in the exam. The other skills are represented in the exam moderately as the correlated mean of the questionnaire filled by the teachers indicated. So, *the UEE was lower in holding face validity evidences*.

Lastly, however there were some sort of differences between the two tests as the comparison of their face validity evidence indicates; the similarity of the two tests holds the higher portion, since they have hold "moderate" face validity evidences. *Both, the CT and UEE hold moderate face validity evidences, thus they are moderately accurate in holding face validity evidences* only.

5.3. Recommendations

The practical uses of the current study may contribute to the area of conducting a valid tests to the learners in terms of the different test validity types. As a result of the investigation of the study, the researcher has recommended as different endeavors/activities should be made by different educational stakeholders to promote validity of the CT and the UEE. In line with the results and conclusion of the study, the following recommendations for practice and for further study were made.

- Much practical research is needed nation-wide to further our understanding and practices in conducting a valid tests.
- Firstly, teachers should discharge one of their professional responsibilities through conducting a test in line with serving the instructional objectives of the course in the syllabus. The objectives of the test should give value for the syllabus objectives to conduct a content valid test.
- Secondly, in order to improve the mismatch in result between the CT and UEE (improve the concurrent validity of CT and UEE) the teachers and national level exam conductors should conduct the CT and the UEE in a similar standard if the improvement of concurrent validity of tests were needed.
- Thirdly, the students should work hard in their UEE to yield similar or nearly similar result with their CT result; and this will in turn improve the concurrent and predictive validity of the tests.

- Fourthly, the teachers should incorporate the writing, reading and speaking skills in the CT they conduct. Some of the language skills like: vocabulary and grammar have to be paid more emphasis however they gain a bit more emphasis in the previous tests. Exercises on student texts should be also given attention. These all endeavors will have improved the face validity of the CT.
- Fifthly, the National Educational Assessment and Examinations Agency, under the Ministry of Education, should make reformations in the examination that would encourage improvement on the face validity of the UEE, based on test validation procedures. The exam should be capable of emphasizing on the listening, speaking, and writing skills; and the contents of the course, the exercises on student text and the laboratory course contents, to improve the face VUEE. However incorporating the speaking, listening and writing skills were difficult and requires technological advancement, they should conduct different researches to come up with remedies.
- Finally, the future researchers will have to conduct a study on construct validity of CT and UEE, and on the content and predictive validity of UEE.

References

Aided Madurai, 2016, ***Evaluation in Education***, Lulu Publication, United States.

Alderson, J., and Wall, D. (1993). 'Does wash back exist?".***Applied Linguistics***. *14/2*, 115-129.

Bachman, L. F. (1990). ***Fundamental Considerations in Language Testing.*** Cambridge: CUP.

Bachman, L., & Palmer, A. (1996).***Language Testing in Practice***. Cambridge: CUP.

Barker, F., McKenna, S., Murray, S., et al (2007, November). ***Overview of the FCE and CAE Review Project research activity***. *Cambridge ESOL Research Notes 30*

Bond (2003).***Exploring How Objects Can Influence The Level Of Construct Validity Of A Picture Vocabulary Test***. University of Pretoria

Brown, D., &Abeywickrama, P. (2010) ***Language Assessment Principles and Classroom Practices***, 2nd Ed, Pearson Education U.S.

Cambridge, ESOL (2009). ***Cambridge ESOL exams and the CEFR***. Retrieved August 6th, 2010, from Cambridge ESOL:

Cambridge, ESOL (2009a). ***Cambridge ESOL exams and the CEFR***. Retrieved August 6th, 2010, from Cambridge ESOL

Cambridge, ESOL (2008, November). ***FCE and CAE exam specifications 2008***.Retrieved August 4, 2010, from Cambridge ESOL Bulletin:

CfBT Education Trust in Association with Fromseas Education & Training /Feat/ S.C, 2015, ***Evaluation of Learning Achievement in Selected Woredas in Amhara and Addis Ababa Sub-Cities***, Addis Ababa, Ethiopia *Language and Education* 11.

Clapham, C (1996) ***The development of IELTS: A study of the effect of background knowledge on reading comprehension***, Studies in Language Testing 4, Cambridge:UCLES, Cambridge University Press.

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

Cronbach, L. J. (1949). ***Essentials of Psychological Testing.*** New York: Harper & Row.

Deborah Grossmann, 2010,***The Validity and Reliability of the Cambridge First Certificate in English,*** Centre for English Language Studies, University of Birmingham

David Ewing (2010). ***Using Test-takers' Feedback To Enhance Quality and Validity in Language Testing***. University of Veracruzana, Mexico

FRANK ESSER (n.d:1-2). ***Comparative Research Methods.*** *University of Zurich, Switzerland*

Gipps, C. V. (1994). ***Beyond testing***: Towards a theory of educational assessment. London: Falmer.

Gunderson, L. (2009). *ESL (ELL)* ***Literacy Instruction****: A Guidebook to Theory and Practice*, 2nd ed. Taylor & Francis US.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk (Ed.), ***A guide to criterion-referenced test construction***. Baltimore and London: The Johns Hopkins University Press.

Hawkey, R. (2009). ***Studies in Language Testing 28****: Examining FCE and CAE.* Cambridge: CUP. B

H. Douglas Brown, 2004. ***Language Assessment Principles and Classroom Practices***,

Hughes, A. (2003). ***Testing for Language Teachers,*** *2nd Edition.* Cambridge: CUP.

Hughes, A. (1989). ***Testing for language teachers***. Cambridge [England] ; New York: Cambridge University Press.

International Conference on Education & Educational Psychology 2013 (ICEEPSY 2013) ***Reliability and content validity of an English as a Foreign Language (EFL) grade-level test for Turkish primary grade students***, *School of Foundations, Leadership, and Administration, Kent*

Jani Lemmetti, 2014,***What makes a good language test in EFL?,***Goteborgs University

John Hopkins, 2007. Measurement: Reliability and Validity Measures. Blooming School of Public Health,

John W. Creswell,(2011)***Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research***, Fourth Edition***,*** *University of Nebraska–Lincoln*

Jones, N (2001) ***Reliability in UCLES Examinations***, Cambridge: UCLES internal report.

Lane, S. (1999). ***Validity evidence for assessment***. Reid Interactive Lecture Series. McNamara,

Melkamu Abate, 2007, ***The Washback Effect of Grade Ten English Language (EGSEC) Examination***, Addis Ababa University Institute of Language Studies

McKenzie, K., Gow, K., & Schweitzer, R. (2004).***Exploring first-year academic achievement through structural equation modeling***. Higher Education Research and Development.

McNamara, T. (2006).***Validity in Language Testing***: the challenge of Sam Messick's legacy. *Language Assessment Quarterly.*

Messick, S. (1989).Validity. In R. L. Linn (Ed.), ***Educational measurement***(3rd ed.). New York: American Council on Education/Macmillan.

Messick, S. (1994).***Alternative Modes of Assessment, Uniform Standards of Validity***, Research Report.

Palmquist. 1990. ***Text Analysis for the Social Sciences***: Methods for Drawing Statistical Inferences from Texts and Tanscripts. Hillsdale, NJ: Lawrence Erlbaum Associates.

Roger Hawkey, 2000,***Studies in Language Testing, Impact Theory and Practice,*** Cambridge University Press, Cambridge, The Edinburgh Building, Cambridge CB2 2RU, UK www.cambridge.org

Saville, N., & Hargreaves, P. (1999).***Assessing speaking in the revised FCE.***ELT *Journal Volume 53/1*.

Simachew Gashaye,2012, ***Washback of the University Entrance English Exam (UEEE) on Teachers' and Students' Practices***: The Case of Preparatory Schools in Amhara National Regional State, Addis Ababa University School of Graduate Studies

Smith, J (2004) ***IELTS Impact: a study on the accessibility of IELTS GT Modules to 16–17 year old candidates***, *Research Notes* 18.

Weir, C. J. (2005).***Language Testing and Validation: An Evidence-Based Approach***. Oxford: Palgrave.

Weir, C., & Shaw, S. (2006, November).***Defining the constructs underpinning the Main Suite Writing Test: a socio-cognitive perspective.*** *Cambridge ESOL Research Notes 26*

Weir, C (2004) ***Language Testing and Validity Evidence***, Basingstoke: Pal

APPENDIX A: Tables of Instructional Objectives and Test Content Specification

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

A. Instructional Objective Specification

No.	Content	General Objectives						
		Knowledge	**Comprehension**	**Application**	**Analysis**	**Synthesis**	**Evaluation**	**Total**
1	Unit 1	3	3	2	2	2	1	**13**
2	Unit 2	0	4	4	3	3	2	**16**
3	Unit 3	2	2	4	3	2	2	**15**
4	Unit 4	2	2	2	2	1	2	**11**
5	Unit 5	4	2	3	2	2	2	**15**
6	Unit 6	4	2	3	2	2	3	**16**
	Total	**15**	**15**	**18**	**14**	**12**	**12**	**86**

B. Test Objectives Specification

No.	Content	General Objectives						
		Knowledge	**Comprehension**	**Application**	**Analysis**	**Synthesis**	**Evaluation**	**Total**
1	Unit 1							**0**
2	Unit 2							**0**
3	Unit 3							**0**
4	Unit 4							**0**
5	Unit 5							**0**
6	Unit 6							**0**
	Total	**10**	**10**	**6**	**11**	**2**	**1**	**40**

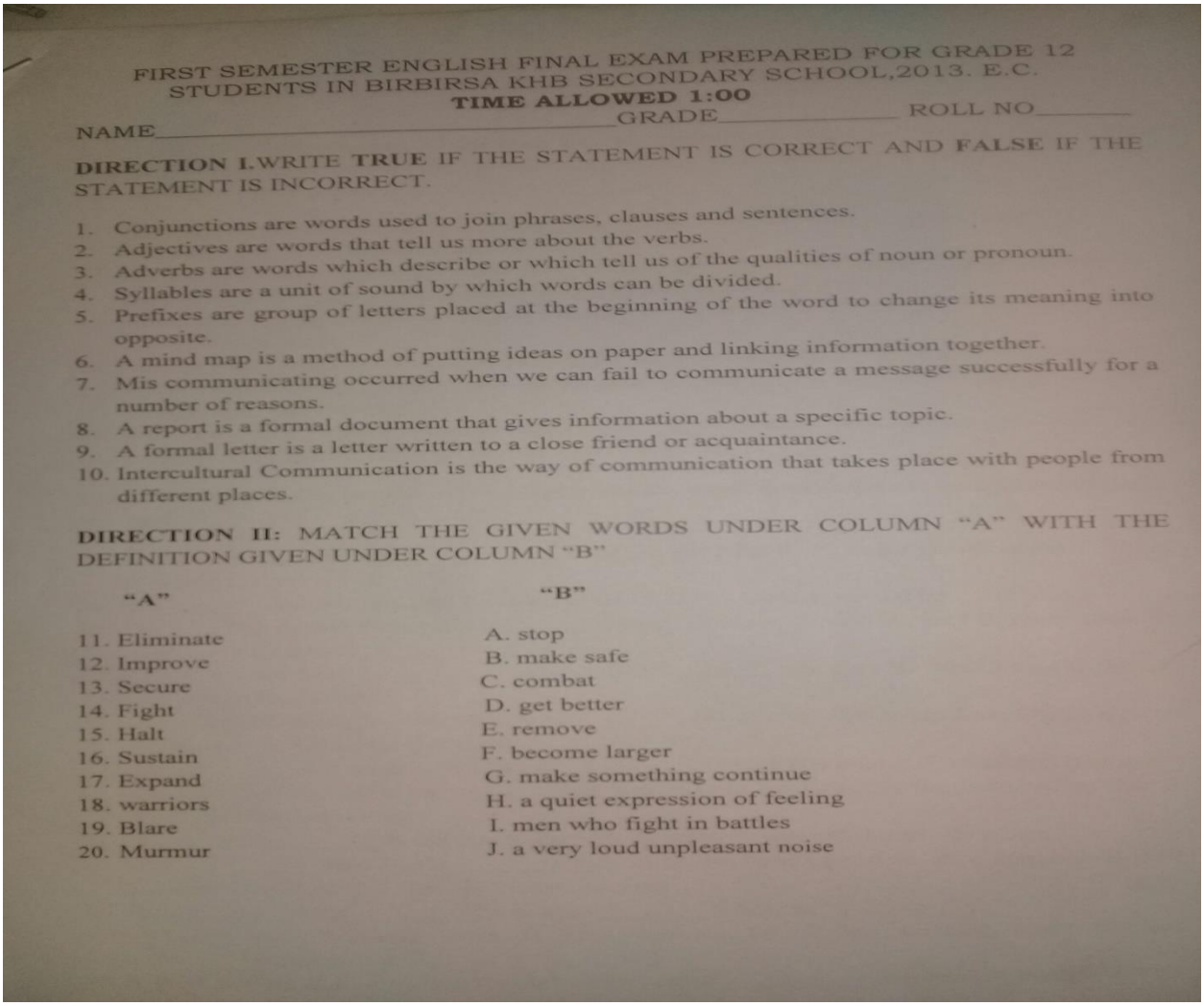
These table format were adapted from "Reader on students assessment", EQUIP November 2008, p.16

APPENDIX B: A Grade 12 English Language Classroom Test

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE



FIRST SEMESTER ENGLISH FINAL EXAM PREPARED FOR GRADE 12 STUDENTS IN BIRBIRSA KHB SECONDARY SCHOOL,2013. E.C.

TIME ALLOWED 1:00

NAME_____GRADE_____ ROLL NO_____

DIRECTION I.WRITE **TRUE** IF THE STATEMENT IS CORRECT AND **FALSE** IF THE STATEMENT IS INCORRECT.

1. Conjunctions are words used to join phrases, clauses and sentences.
2. Adjectives are words that tell us more about the verbs.
3. Adverbs are words which describe or which tell us of the qualities of noun or pronoun.
4. Syllables are a unit of sound by which words can be divided.
5. Prefixes are group of letters placed at the beginning of the word to change its meaning into opposite.
6. A mind map is a method of putting ideas on paper and linking information together.
7. Mis communicating occurred when we can fail to communicate a message successfully for a number of reasons.
8. A report is a formal document that gives information about a specific topic.
9. A formal letter is a letter written to a close friend or acquaintance.
10. Intercultural Communication is the way of communication that takes place with people from different places.

DIRECTION II: MATCH THE GIVEN WORDS UNDER COLUMN "A" WITH THE DEFINITION GIVEN UNDER COLUMN "B"

"A"	"B"
11. Eliminate	A. stop
12. Improve	B. make safe
13. Secure	C. combat
14. Fight	D. get better
15. Halt	E. remove
16. Sustain	F. become larger
17. Expand	G. make something continue
18. warriors	H. a quiet expression of feeling
19. Blare	I. men who fight in battles
20. Murmur	J. a very loud unpleasant noise

APPENDIX C: Students Names and Their Result in Both Exams

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

APPENDIX E

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

A TABLE FOR CORRELATION OF *CONCURRENT* VALIDITY EVIDENCE ON BOTH SCHOOL LEVEL AND ENTRANCE EXAM

A List of the Students **Name and Their Result in Both Exams at** *Birbirsa* **Secondary School**

In order to correlate the two results, and illustrate the concurrent validity evidence the following hypothetical table which is taken from Kubiszyn & Borich, 2003:301 will be used.

No.	Student Name	Scores on entrance exam	Scores on school level final exam
1	Birtukan Adisu / Abde Mohamed	29	52
2	Burka Chala	27	58
3	Desta Tilaye	23	55.5
4	Jindalkacho Geleta	32	59.5
5	Himot Tesfa	28	60
6	Kasahun Yadeta	27	89.5
7	Kume Ifhetu	34	70.5
8	Kusa Ketema	33	65.5
9	Leila Kedir	31	51.5
10	Lomite Negese	23	55.5
11	Mihiret Shontow	30	56.5
12	Nahim Mohamed	15	58.5
13	Simeny Betele	20	61
14	Tesnale Arega	27	54
15	Tolosa Yirdaw	23	59
16	Tsedale Tellu	21	50.5
17	Umer Teferi	26	67
18	Yalew Tesfaye	23	61

19	Yergalem	Kebede	20	61.5
20	Bontukan	Alemu	29	67
21	Bontukan	Birahu	20	76.5
22	Chalachew	Imiru	33	74.5
23	Debash	Boyine	35	81.5
25	Derartu	Tekuma	25	56
26	Ebisa	Terunde	40	99
27	Gemechu	Haile	20	59
28	Gezenat	Asmamaw	19	68
29	Kenesa	Getacho		68
30	Mesfin	Tefera	31	62
31	Negahe	Kebede	15	55.5
32	Zanete	Tefera		59
33	Samuel	Abdo	27	58.5
34	Tekaye	Kitesa	24	62
35	Tolera	Shibiru	29	77
36	Yaikob	Gezahegn	35	84.5
37				
38				
39				
40				
41				
42				
43				
44				

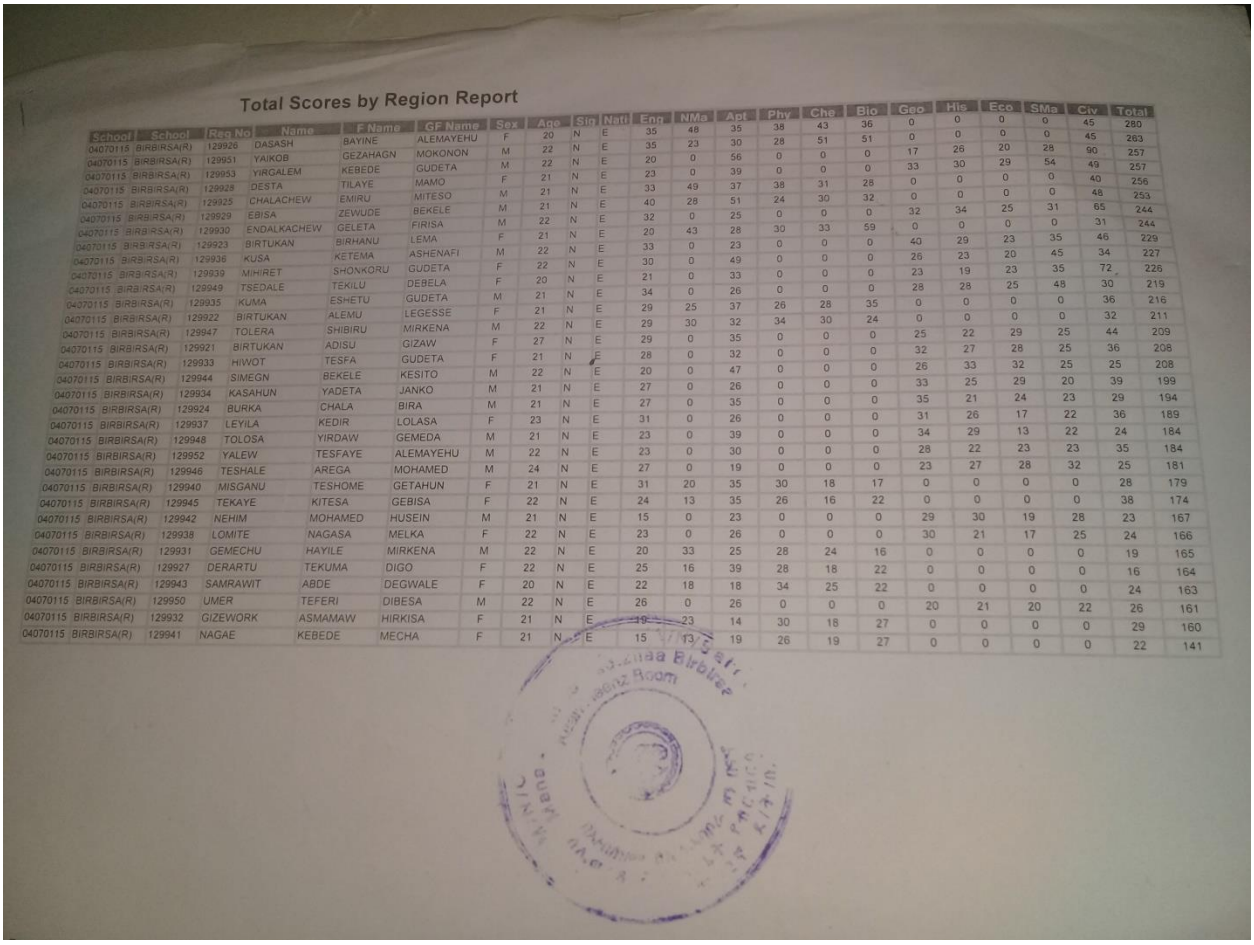
N.B. The score of the total sample (44 students) will be listed in this table, and will be c
to determine the concurrent validity.

APPENDIX D: A 2020 University Entrance Exam Result

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE



Total Scores by Region Report

School	School	Reg No	Name	F Name	GF Name	Sex	Age	Sta	Nati	Eqn	NMa	Apt	Phy	Che	Bio	Geo	His	Eco	SMa	Civ	Total
04070115	BIRBIRSA(R)	129926	DASASH	BAYINE	ALEMAYEHU	F	20	N	E	35	48	35	38	43	36	0	0	0	0	45	280
04070115	BIRBIRSA(R)	129951	YAIKOB	GEZAHAGN	MOKONON	M	22	N	E	35	23	30	28	51	51	0	0	0	0	45	263
04070115	BIRBIRSA(R)	129953	YIRGALEM	KEBEDE	GUDETA	M	22	N	E	20	0	56	0	0	0	17	26	20	28	90	257
04070115	BIRBIRSA(R)	129928	DESTA	TILAYE	MAMO	F	21	N	E	23	0	39	0	0	0	33	30	29	54	49	257
04070115	BIRBIRSA(R)	129925	CHALACHEW	EMIRU	MITESO	M	21	N	E	33	49	37	38	31	28	0	0	0	0	40	256
04070115	BIRBIRSA(R)	129929	EBISA	ZEWUDE	BEKELE	M	21	N	E	40	28	51	24	30	32	0	0	0	0	48	253
04070115	BIRBIRSA(R)	129930	ENDALKACHEW	GELETA	FIRISA	M	22	N	E	32	0	25	0	0	0	32	34	25	31	65	244
04070115	BIRBIRSA(R)	129923	BIRTUKAN	BIRHANU	LEMA	F	21	N	E	20	43	28	30	33	59	0	0	0	0	31	244
04070115	BIRBIRSA(R)	129936	KUSA	KETEMA	ASHENAFI	M	22	N	E	33	0	23	0	0	0	40	29	23	35	46	229
04070115	BIRBIRSA(R)	129939	MIHIRET	SHONKORU	GUDETA	F	22	N	E	30	0	49	0	0	0	26	23	20	45	34	227
04070115	BIRBIRSA(R)	129949	TSEDALE	TEKILU	DEBELA	F	20	N	E	21	0	33	0	0	0	23	19	23	35	72	226
04070115	BIRBIRSA(R)	129935	KUMA	ESHETU	GUDETA	M	21	N	E	34	0	26	0	0	0	28	28	25	48	30	219
04070115	BIRBIRSA(R)	129922	BIRTUKAN	ALEMU	LEGESSE	F	21	N	E	29	25	37	26	28	35	0	0	0	0	36	216
04070115	BIRBIRSA(R)	129947	TOLERA	SHIBIRU	MIRKENA	M	22	N	E	29	30	32	34	30	24	0	0	0	0	32	211
04070115	BIRBIRSA(R)	129921	BIRTUKAN	ADISU	GIZAW	F	27	N	E	29	0	35	0	0	0	25	22	29	25	44	209
04070115	BIRBIRSA(R)	129933	HIWOT	TESFA	GUDETA	F	21	N	E	28	0	32	0	0	0	32	27	28	25	36	208
04070115	BIRBIRSA(R)	129944	SIMEGN	BEKELE	KESITO	M	22	N	E	20	0	47	0	0	0	26	33	32	25	25	208
04070115	BIRBIRSA(R)	129934	KASAHUN	YADETA	JANKO	M	21	N	E	27	0	26	0	0	0	33	25	29	20	39	199
04070115	BIRBIRSA(R)	129924	BURKA	CHALA	BIRA	M	21	N	E	27	0	35	0	0	0	35	21	24	23	29	194
04070115	BIRBIRSA(R)	129937	LEYILA	KEDIR	LOLASA	F	23	N	E	31	0	26	0	0	0	31	26	17	22	36	189
04070115	BIRBIRSA(R)	129948	TOLOSA	YIRDAW	GEMEDA	M	21	N	E	23	0	39	0	0	0	34	29	13	22	24	184
04070115	BIRBIRSA(R)	129952	YALEW	TESFAYE	ALEMAYEHU	M	22	N	E	23	0	30	0	0	0	28	22	23	23	35	184
04070115	BIRBIRSA(R)	129946	TESHALE	AREGA	MOHAMED	M	24	N	E	27	0	19	0	0	0	23	27	28	32	25	181
04070115	BIRBIRSA(R)	129940	MISGANU	TESHOME	GETAHUN	F	21	N	E	31	20	35	30	18	17	0	0	0	0	28	179
04070115	BIRBIRSA(R)	129945	TEKAYE	KITESA	GEBISA	F	22	N	E	24	13	35	26	16	22	0	0	0	0	38	174
04070115	BIRBIRSA(R)	129942	NEHIM	MOHAMED	HUSEIN	M	21	N	E	15	0	23	0	0	0	29	30	19	28	23	167
04070115	BIRBIRSA(R)	129938	LOMITE	NAGASA	MELKA	F	22	N	E	23	0	26	0	0	0	30	21	17	25	24	166
04070115	BIRBIRSA(R)	129931	GEMECHU	HAYILE	MIRKENA	M	22	N	E	20	33	25	28	24	16	0	0	0	0	19	165
04070115	BIRBIRSA(R)	129927	DERARTU	TEKUMA	DIGO	F	22	N	E	25	16	39	28	18	22	0	0	0	0	16	164
04070115	BIRBIRSA(R)	129943	SAMRAWIT	ABDE	DEGWALE	F	20	N	E	22	18	18	34	25	22	0	0	0	0	24	163
04070115	BIRBIRSA(R)	129950	UMER	TEFERI	DIBESA	M	22	N	E	26	0	26	0	0	0	20	21	20	22	26	161
04070115	BIRBIRSA(R)	129932	GIZEWORK	ASMAMAW	HIRKISA	F	21	N	E	19	23	14	30	18	27	0	0	0	0	29	160
04070115	BIRBIRSA(R)	129941	NAGAE	KEBEDE	MECHA	F	21	N	E	15	13	19	26	19	27	0	0	0	0	22	141

APPENDIX E: Students and Teachers Questionnaire

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

Dear Student,

This format has been designed for the purpose of helping evaluate the face validity of school level English Language test. Please read and then rate for the ***extent to which a test looks effective*** in terms of what it is expected to measure. Your genuine response will contribute a lot to the study so that the result may help enhance effective language testing.

Thank you in advance!

Allow 5 if you strongly agreed upon the idea

Allow 4 if you agreed upon the issue

Allow 3 if you undecided upon the idea

Allow 2 if you disagreed upon the issue

Allow 1 if you disagree on the issue

Please put a tick mark [√] in the appropriate box.

1. The content of the main course book was represented in the tests and exams sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
2. The content of the grammar book was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
3. The content of the writing courses was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
4. The content of the reading courses was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
5. The content of the speaking course was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
6. Grammar taught in the courses was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
7. The vocabulary taught in course was represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐

8. The listening practices focused on in the courses were represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
9. The exercises made in the courses were represented in the test sufficiently.
   strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
10. The content of the laboratory courses was represented in the test sufficiently.
    strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐
11. In general, the contents of the courses were represented in the test sufficiently.
    strongly agree (5) ☐ agree (4) ☐ undecided (3) ☐ disagree ☐ strongly disagree(1) ☐

(… from Brown, 2004.p.27.)

RARRAATUU F: Bar-gaaffii Barattootaa
UNIVARSIITII JIMMAA

MUUMMEE BARNOOTA AFAAN INGILIFFAA FI OG-BARRUU

GUCA SIRRUMMAAN QORMAATA DAREE ITTI QABXEESSAMU

Kabajamoo b/taa/ttuu;

Uunki kun kan qophaa'eef sirrummaa qormaata Afaan Ingiliffaa madaaluuf akka isin gargaaruufi. Sirriitti erga dubbisteen booda qormaatichi hangam bu'a qabeessa akka fakkaatuu fi hangam waan madaaluuf yaadame akka madaaluu danda'u qabxeessi. Deebiin ati kennitu qoarannoo kanaaf bu'aa ni buusa; kanaanis bu'aan argamus dandeettii afaanii karaa bu'a qabeessa ta'een qoruu ni fooyyeessa.

Alaalatti galata qabda!

Qormaanni ati qoramte bu'a qabeessa baayyee olaanaadha kan jettu yoo ta'e, ***baayyee olaanaaykn 4*** kenni.

Qormaanni ati qoramte bu'a qabeessa olaanaadha kan jettu yoo ta'e, ***olaanaa ykn 3*** kenni.

Qormaanni ati qoramte bu'a qabeessa giddu-galeessaati kan jettu yoo ta'e, ***giddu-galeessa ykn 2*** kenni.

Qormaanni ati qoramte bu'a qabeessa gadi-aanaadha kan jettu yoo ta'e, ***gadi-aanaa ykn 1*** kenni.

Qormaanni ati qoramte bu'a qabeessa miti kan jettu yoo ta'e, tasa ***bu'aa hin qabu ykn 0*** kenni.

Mallattoo sororsuu [√] saanduqa keessatti guutuun filannoo kee agarsiisi

1. Qabiyyeen kitaaba barnoota ijoon haalan qormaaticha keessatti haammatameera.

   baayyee olaalaanaa (4) ☐ olaanaa(3) ☐ giddu-galeessa(2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

2. Qabiyyeen seer-lugaa kitaabichaa haalan qormaaticha keessatti haammatameera.

   baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

3. Qabiyyeendandeettiibarreessuubarnootichaahaalanqormaatichakeessattihaammatameera.

   baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

4. Qabiyyeen dubbisuu barnootichaa haalan qormaaticha keessatti haammatameera.

   baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

5. Qabiyyeen dubbachuu barnootichaa haalan qormaaticha keessatti haammatameera.

baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

6. Seer-lugni barnooticha keessatti barsiisamuhaalagaariinqormaatichattihaammatameera.

baayyeeolaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

7. Hiikaan jechootaa barnooticha keessatti barsiisamu haalan qormaaticha keessatti haammatameera.

baayyee olaalaanaa (4) ☐ olasanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

8. Shaakalli dhaggeeffachuu barnooticha keessatti irratti xiyyeeffatame haala gaariin qormaatichatti haammatameera.

baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

9. Gilgaaloonni barnooticha keessa jiranu haala gaariin qormaatichatti haammatameera.

baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

10. Qabiyyeen hiikaa jechootaa barnooticha keessa jiranu haala gaariin qormaatichatti haammatameera.

baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

11. Walumaagalatti qabiyyeen barnootichaa haala gaariin qormaatichatti haammatameera.

baayyee olaalaanaa (4) ☐ olaanaa (3) ☐ giddu-galeessa (2) ☐ gadi-aanaa (1) ☐ tasa (0) ☐

APPENDIX G: Interview Questions for Teachers

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE

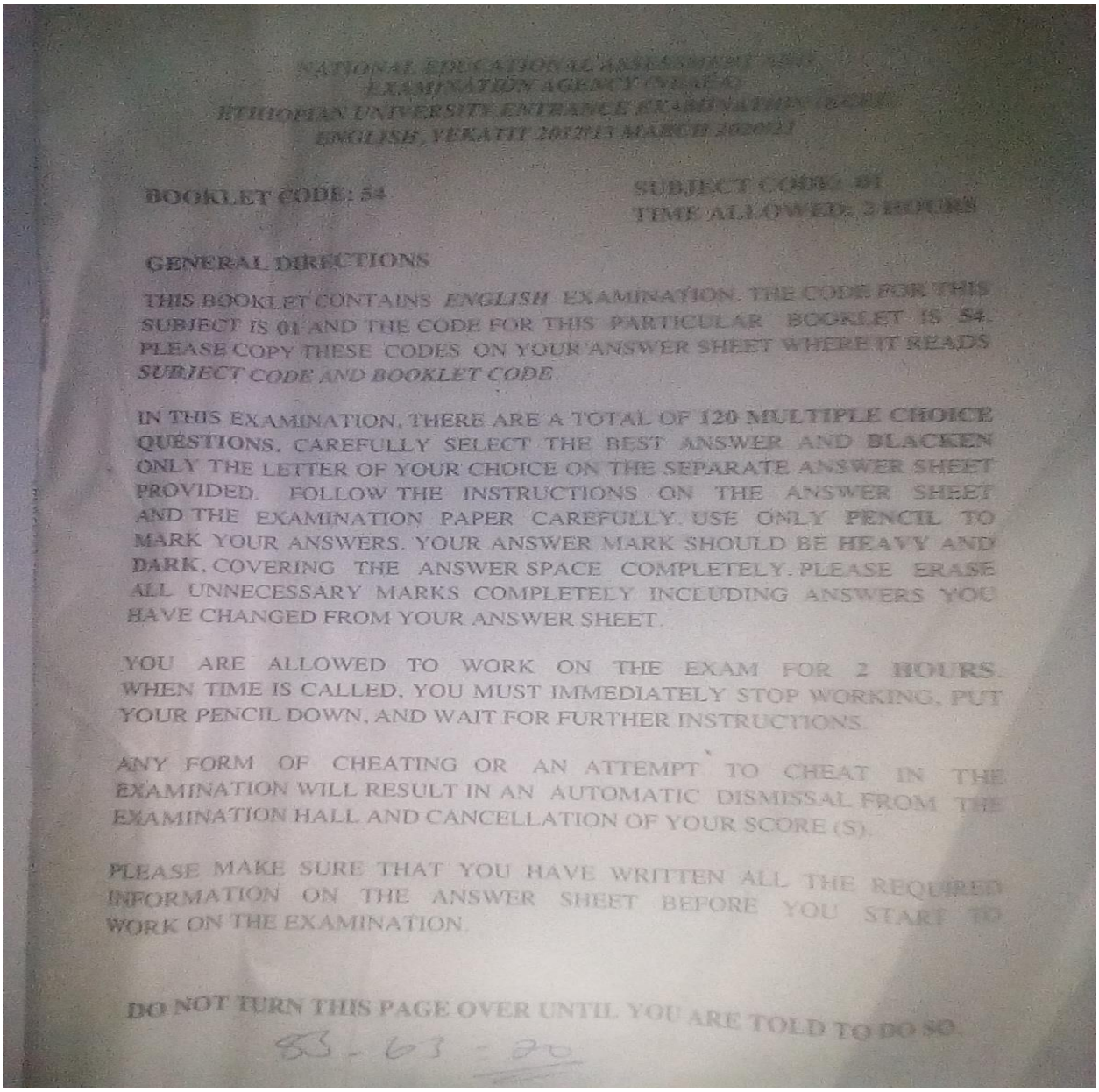
1. How far the content of the main course book were represented in the entrance exam, and what are the indicators for your assertion?
2. What is your attitude towards the inclusion of the content of the grammar contents that the students have learned in the classroom?
3. What you feel about the inclusion of writing courses in entrance exam?
4. How and in what form the reading/compression questions in entrance exam can impact the students learning?
5. What is the similarity and difference, and simplicity and difficulty relation between the vocabulary the students learned in the classroom and incorporated in the entrance exam?
6. On what types of language skills did the exam concentrates and what are the drawbacks?
7. How far the students result in entrance exam can represent the language skill of the learners?

APPENDIX H: A 2020 Grade 12 University Entrance Exam

JIMMA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF ENGLISH LANGUAGE AND LITERATURE



NATIONAL EDUCATIONAL ASSESSMENT AND
EXAMINATION AGENCY (NEAEA)
ETHIOPIAN UNIVERSITY ENTRANCE EXAMINATION (EUEE)
ENGLISH, YEKATIT 2012/13 MARCH 2020/21

BOOKLET CODE: 54

SUBJECT CODE: 01
TIME ALLOWED: 2 HOURS

GENERAL DIRECTIONS

THIS BOOKLET CONTAINS *ENGLISH* EXAMINATION. THE CODE FOR THIS SUBJECT IS 01 AND THE CODE FOR THIS PARTICULAR BOOKLET IS 54. PLEASE COPY THESE CODES ON YOUR ANSWER SHEET WHERE IT READS *SUBJECT CODE AND BOOKLET CODE*.

IN THIS EXAMINATION, THERE ARE A TOTAL OF **120 MULTIPLE CHOICE QUESTIONS**. CAREFULLY SELECT THE BEST ANSWER AND **BLACKEN** ONLY THE LETTER OF YOUR CHOICE ON THE SEPARATE ANSWER SHEET PROVIDED. FOLLOW THE INSTRUCTIONS ON THE ANSWER SHEET AND THE EXAMINATION PAPER CAREFULLY. USE ONLY **PENCIL** TO MARK YOUR ANSWERS. YOUR ANSWER MARK SHOULD BE **HEAVY** AND **DARK**, COVERING THE ANSWER SPACE COMPLETELY. PLEASE ERASE ALL UNNECESSARY MARKS COMPLETELY INCLUDING ANSWERS YOU HAVE CHANGED FROM YOUR ANSWER SHEET.

YOU ARE ALLOWED TO WORK ON THE EXAM FOR **2 HOURS**. WHEN TIME IS CALLED, YOU MUST IMMEDIATELY STOP WORKING, PUT YOUR PENCIL DOWN, AND WAIT FOR FURTHER INSTRUCTIONS.

ANY FORM OF CHEATING OR AN ATTEMPT TO CHEAT IN THE EXAMINATION WILL RESULT IN AN AUTOMATIC DISMISSAL FROM THE EXAMINATION HALL AND CANCELLATION OF YOUR SCORE (S).

PLEASE MAKE SURE THAT YOU HAVE WRITTEN ALL THE REQUIRED INFORMATION ON THE ANSWER SHEET BEFORE YOU START TO WORK ON THE EXAMINATION.

DO NOT TURN THIS PAGE OVER UNTIL YOU ARE TOLD TO DO SO.

83-63-20

List of the students and their scores in English Language Classroom Tests and in University Entrance Exam by the year 2020

No.	Name	Scores on UEE out of 100%	Scores on CT Out of 100%
1	EZB	40	99
2	YGM	35	84.5
3	TAM	35	81.5
4	KEG	34	70.5
5	CIM	33	74.5
6	KKA	33	65.5
7	EWF	32	59.5
8	LKL	31	51.5
9	MTG	31	62
10	MSG	30	56.5
11	BAL	29	67
12	TSM	29	77
13	BAG	29	52
14	HTG	28	60
15	KYJ	27	89.5
16	BCB	27	58
17	TAM	27	54
18	UTD	26	67
19	DTD	25	56
20	TKG	24	62
21	DTM	23	55.5
22	TYG	23	59
23	YTA	23	61
24	LNM	23	55.5
25	SAD	22	58.5
26	TTD	21	50.5
27	YKG	20	61.5
28	BBL	20	76.5
29	SBK	20	61
30	GHM	20	59
31	GAI	19	68
32	NMH	15	58.5
33	NKM	15	55.5

This hypothetical table was taken from Kubiszyn &Borich, 2003:301