Communicative Language Testing
Its Validity and Reliability at the Addis Ababa University

A Thesis Presented to the
School of Graduate Studies
of Addis Ababa University

In Partial fulfilment of the Requirements
of the Degree of master of Arts in Teaching English as a Foreign
Language (TEFL)

By
Seifu Bekuretsion
May, 1997

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COMMUNICATIVE LANGUAGE TESTING
ITS VALIDITY AND RELIABILITY AT THE ADDIS ABABA UNIVERSITY

BY

SEIFU BEKURETSION

APPROVED BY _____

Teshome Demisse _____

Advisor

John Norris _____

Examiner

Dejene Lela _____

Examiner

_____ _____

Examiner

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First of all, my heartfelt and respectful gratitude goes to my advisor, Dr. Teshome Demissie, whose invaluable, unreserved and constructive comments and advise have always been at my disposal whenever I needed them. Without his guide, much of the work would have been very difficult to deal with. I also would like to thank Dr. Dejenie Leta, who has given me a lot of constructive insights at different stages of the work. I am also highly indebted to the students who took part in the research sacrificing their time and energy. The contribution of the instructors who responded to the instructor's questionnaire has also been great. The help I received from my good friends, Mulugeta Teka and Sisay Haile, especially regarding the statistical part, is also highly appreciated. All my friends and relatives who helped me morally and financially also deserve thanks. Thank you all very much indeed.

Abstract

The aim of this study was to try and see if the English language can be tested communicatively at the university and if the test produced has the required types of validity and reliability.

So a test was prepared and administered for 120 students in the freshman programme. And after multiple marking was applied, the papers of 80 students (2/3 of the total number of subjects) were included in the item analysis list. Results of the item-analysis showed the test had items of varied and acceptable facility values (average= 54.8) and indexes of discrimination.(average =0.36)

The reliability of the test was also checked with the help of two methods: the Spearman-Brown split-half which gave r=0.79, and the Kuder - Richardson 20(KR 21) which gave r=0.8. KR 21 was used because it also takes into consideration the interitem consistency in the test. The results obtained in both showed the test could be accepted as a reliable instrument.

The results of the subjects were also correlated with their results in College English - I to check the test's concurrent validity. The two tests showed a good correlation at r=0.75.

Two questionnaires were also prepared: one for the students who took the test and one for instructors. The results of the analysis of the responses of the respondents showed the test had the different types of validity.

So it was concluded that communicative language testing can be applicable in the tertiary level, and recommendations where made for test writers.

CHAPTER ONE

INTRODUCTION

1.1. Statement of the Problem

In Ethiopia English is given as a school subject starting from grade one. Moreover it is the medium of instruction starting from grade 7 up to colleges and universities. A consideration of these facts only shows the key role English has to play in the Ethiopian educational system.

After considering the English language as a subject and talking about the tertiary level we can say that most of the students joining the Addis Ababa University are students exposed to a predominantly grammar - based language teaching throughout their school life. This language teaching in Ethiopian schools is a typical example of what Carroll (1980) describes as pedagogies that emphasise the teaching of language 'usage'. He says such pedagogies are characterized by "the selections and sequencing of learning units expressed in formal linguistic terms" (p.7). Formal correctness of linguistic forms is given prominence. The belief behind such pedagogies is that systematic mastery of formal patterns will necessarily lead to the effective' use of the target language which, in the Ethiopian case, is English.

But when we compare the teaching of English at schools with that at colleges and universities, we easily notice a very significant difference. And when we particularly focus on first year students, this difference can be particularly great.

Education at the tertiary level is very much different from (high) schools in that it is demanding. It is demanding in the sense that it requires students to use all the four skills of communication; and, what is more, they are required to use them at an advanced level. So the students joining higher education, in this sense, enter a new world that is challenging, for the focus regarding the English language is not just on linguistic forms.

Moreover, as Hughes (1989:141) says, "there is more to a skill than the sum of its parts." Wesche (1983) also says that language competence is not "additive" or the sum of its parts. According to her, teaching language as discrete points does not lead to overall mastery of the language. This view is shared by many authors (e.g. Oller, 1979; and Clark, 1978). Clark mentions Cooper, 1970; Jacobovits, 1969; 1970; Paulston, 1974; Spolsky, 1968; Upshur, 1972. To emphasise this point, Oller talks about "discrete elements of discrete aspects of discrete components of discrete skills..." (P.172). Consequently, in the Ethiopian case when students join the university, they are like individuals that Johnson (1981), quoted in Swan (1985) refers to as "structurally competent and communicatively incompetent."

So there has been a clear need for change so that students joining the university could cope with the demands of the new situation. The Department of Foreign Languages and Literature has clearly realized this felt-need for a change and has responded by the preparing a new teaching material which is communicative in nature.

Along with the teaching, it is clear that the testing should also change, i.e. it should be communicative as well. This is because testing is a "partner" of teaching, and it cannot be separated from it (Hughes, 1989, Heaton, 1988). Thus, in our case there again is a clear need to get valid and reliable communicative test which together with the teaching help to effect the whole process of teaching English at the University.

1.2. Objectives of the Study

The main objective of the study is to try to prepare a valid and reliable communicative test for first year students at the Addis Ababa University. The test focuses mainly on the skills of reading and writing. It also has passage (context) vocabulary and contextualized grammar items. Secondly, since the Department of Foreign Languages and Literature is making the Freshman English course more and more communicative in nature, this study aims at giving the testing body of the course some useful points. And, in a sense, we can say the study aims at identifying some practical problems that stand on the way of making the testing system communicative.

1.3. Significance of the Study

The results of the study are hoped to shed more light on the road of change the Department of Foreign Languages and Literature has embarked on; i.e. the results are hoped to be applicable in the testing system of the department.

Secondly, whatever problems are faced during the study can be taken into consideration so that the department's testing body can take appropriate measures to avoid similar occurrences of such problems during the actual test preparation.

It is also hoped that higher institutions that will be using the new teaching material prepared by the department or, in general, institutions that plan to prepare valid and reliable tests of the two skills will find the results of the study applicable.

More over such testing along with the communicative teaching is hoped to, as Weir (1990:14) says, enhance "the match between teaching, testing and reality."

And as the paper incorporates, as much as possible the most recent ideas in the field of language testing, instructors involved in preparation of tests, or in writing test items, at the tertiary level are hoped to gain some valuable points to apply.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

2.1. The Relationship between Testing and Teaching

How are teaching and testing related? Is it in the way Davies (1968:5) as quoted in Hughes (1989:2) says; that "the good test is an obedient servant since it follows and apes the teaching"? Or is there any other relationship between the two?

The ideas we find in the testing literature regard the relationship between teaching and testing differently from Davies.

For Hughes (1989) the two, i.e. teaching and testing, have a relationship of "partnership". He says there may be cases where the teaching is good and the testing bad, in which case there will result a harmful backwash. If the case is the other way round; i.e. if the teaching is bad and the testing good, then the testing is able to exert a beneficial influence on the teaching.

Similar ideas are discussed by Heaton (1988). He says it is a fact that teaching and testing are "so closely related to each other that it is virtually impossible to work in either field without being concerned constantly with the other" (p.5).

Many scholars emphasise the importance of backwash (washback) (for example Weir, 1990; Hughes, 1989; Heaton, 1988). Backwash is the influence of testing on teaching.

Though there are some scholars who seem to be sceptical about the very existence of washback itself (Alderson and Wall (1993), for example, say "Does washback exist?") many believe that washback really exists and that it can be both harmful and beneficial in the influence it exerts on teaching.

Talking particularly about language tests in the classroom, Oller (1987) says they serve a number of purposes including instructional, managerial (providing feedback), motivational (as rewarding or prodding instruments), diagnostic, and curricular (helping to define the curriculum as a whole) (p.44).

Similarly, Richards (1985) says knowing what the students have achieved from the teaching can be seen as a very important part in the evaluation of the objectives of language programmes. He says "in addition to what different types of tests we have and their purposes, we can see testing as a vital component of curriculum development and evaluation" (p. 15)

So we can conclude that teaching and testing are inseparable and that the agreement of the two is decisive to the effective achievement of objectives of the classroom teaching and overall language programmes.

2.2. Types of Tests: An Overview

There are different types of tests that are designed for different purposes and that go under many names. But there appears to be some confusion regarding the terminology used to denote the different types of tests in use" (Heaton 1988:171). In this section, the following terminologies are used to refer to the different types of tests: Achievement, proficiency, diagnostic, and placement. They are discussed in this same order.

2.2.1. Achievement Tests

Achievement tests, also called attainment tests (Heaton, 1988) refer to a broader aspect of language testing and are better understood by dividing them into two: progress achievement and final achievement.

2.2.1.1. Progress achievement Tests

These tests are given "at various stages throughout a language course to see what the students have learnt" (Alderson et al. 1995:11). They are designed in such a way that they "measure the extent to which the students have mastered the material taught in the classroom" (Heaton, 1988: 171). Heaton also says each of these several tests given at different points is unique and each depends on the individual teacher's aims and goals as well as the teacher's knowledge of his or her students and the language programme that the students have been following. He also says the progress could be looked at in two ways: the students' progress

in the language programme and the teacher's progress in achieving aims and objectives.

2.2.1.2. Final Achievement Tests

These tests, though similar to progress tests, tend to be given at the end of the course (Alderson et. Al. 1995). These tests, says Heaton (1988), are similar to progress tests in a number of ways but are "far more formal tests and are intended to measure achievement of a larger scale." (P.172).

Examples of such tests given by Heaton include: most annual school examination and all public tests which are intended to show mastery of a particular syllabus. Such tests take into consideration what the students are supposed to have learned, and not just what they have actually learned (Alderson et al., 1995; Hughes, 1989; Heaton, 1988).

2.2.2. Proficiency Tests

Proficiency tests, unlike achievement tests, are not based on any particular language programmes (Alderson et al. 1995), and whereas achievement tests "look back on what should have been learnt" proficiency tests "look forward, defining a student's language proficiency with reference to a particular task which she or he will be required to perform" (Heaton, 1989:172). Heaton's idea agrees with that of Alderson et al. In that proficiency tests are in no way related to any particular language teaching syllabus because, as mentioned above, they don't intend to measure mastery or otherwise of any syllabus. They are rather

interested in the general proficiency level of the test takers so they are concerned "Simply with measuring a student's control of the language in the light of what he or she will be expected to do with it in the future performance of a particular task" (Heaton, 1988: 172-73). So, according to him, these tests can be given to students from different schools, countries, and even language backgrounds.

Alderson et al. (1955) say if we are thinking of Specific Purpose (SP) tests, the content of the test will depend on needs analysis.

2.2.3. Diagnostic Tests

Diagnostic tests aim at identifying students' weaknesses and strengths. Hughes (1989) says they are intended "Primarily to ascertain what further teaching is necessary" (p.13). Alderson et al. (1995) also say they tell us in which areas a student needs further help. So our primary concern in giving these tests is to know in which areas, for example the major skills - listening, speaking, reading and writing- or the subskills - grammar and vocabulary that the student is weak so that appropriate remedial teaching is given.

Hughes (1989) says we can even go further analysing a student's performance in writing and speaking in order to create profiles of the student's ability with respect to such categories as 'grammatical accuracy' or 'linguistic appropriacy'(p.13).

Hughes (1989) refers to the lack of good diagnostic tests as "unfortunate" while Alderson et al. (1995) mention the difficulty of preparing good diagnostic tests because of the complex nature of language ability.

Heaton (1988) says such tests are frequently given to groups instead of to individuals. In such cases if just one or two students make a certain error the teacher may not give it much attention; but if, instead, several students make the error, then the teacher plans appropriate teaching to deal with the problem.

2.2.4. Placement Tests

Placement tests are designed, as the name suggests, to place students in different levels. Or, according to Harrison (1983), placement tests are designed "to sort new students into teaching groups, so that they can start a course at approximately the same level as other students in the class" (p.4).

Similarly, Hughes (1989) says these tests "provide information which will help to place students at the stage (or in the part) of the teaching programme most appropriate to their abilities" (p.14). The placement, according to Alderson et al. (1995), could be of a mixed type. For example, a student, after taking the test, could be placed "in the top reading class and bottom writing class, or some other combination" (p.11).

2.3. Communicative Language Testing

How can we define communicative Language Tests? Davies (1988:5) says communicative language testing, just like communicative language teaching, has " no clear cut definition, and it often means different things to different people. Referring back to his own earlier idea (Davies, 1982), he says it may be useful to talk in terms of a "continuum". The continuum has ideas of what communicative language tests are likely to be "more of" and "less of". He says communicative language tests are likely to be "more integrative and less discrete point; more direct and less indirect; more criterion referenced and less norm-referenced" (p.6).

Talking about the concerns of communicative language tests, Heaton (1988:19) simply says they are "concerned primarily (if not totally) with how language is used in communication". So communicative language tests go beyond checking mastery of formal linguistic patterns. As Wesche (1983:47) says, we want "to tap communicative and not only grammatical competence". Accordingly, success in such tests is judged "in terms of the effectiveness of the communication which takes place rather than formal linguistic accuracy" (Heaton, 1988:19).

In order to meet the objective of checking overall language control, Davies(1988) says, we need "a variety of situations and array of tasks," and according to Wesche (1983:42), "we need to go beyond the sentence level."

Wesche (1983: 47-48) presents several characteristics of communicative language tests. She says they must activate internal rule systems by which discourse is meaningfully processed, including those by which sociolinguistic variables influence language behaviour. Secondly, they should be as direct as possible so that we know whether the testees can actually "do" something in the target language with an acceptable degree of acceptance. They should also be criterion - referenced. To improve reliability, she goes on, we must carefully experiment with and analyse the results of our new testing approaches (with considerable expense - to train markers, for example). On the contrary, Wesche concludes since we also want our tests to be feasible, it is very unlikely that they meet all these criteria all the time. "The best we can do is try hard to take into consideration these criteria and strive to make our tests better measures of communicative performance" (Wesche, 1983:48).

So communicative language testing is clearly interested in language use. But, Heaton says, "use" should not be emphasised to the exclusion of "usage". In practice, he says, some communicative tests include some items to test " ability to handle the formal patterns of the target language: (p.19). This point would be justifiable because one is not likely to make use of the language if one does not have adequate control over linguistic patterns. On the other hand, communicative language tests don't usually test skills separately because

> The assessment of language skills in isolation may have only a very limited relevance to real life. For example, reading would rarely be taken solely for its own sake in academic study but rather for subsequent transfer of the information obtained to writing or speaking.
>
> (Heaton, 1988:20)

In a summarized comment regarding the criteria of a language test in the communicative paradigm, Weir (1990:36) says such a test might be expected to exhibit the following:

> ...it should be interactive; direct in nature with tasks reflecting realistic discourse processing activities; texts and tasks should be relevant to the intended situation of the target population; ability should be sampled with meaningful and developing contexts and the test should be based on an explicit apriori specification.

Talking about the characteristics that are considered important in the design of communicative tests, Weir (1990:38) presents the following list which is derived from a questionnaire administered to language school teachers (Weir, 1983) and from the work of Roger Hawkey (1982) and Keith Morrow (1977, 1979).

The characteristics that might be expected are:

- Realistic context - the task should be regarded as appropriate to the candidate's situation
- Relevant Information gap - candidates should have to process new information as they might in the real-life situation.
- Intersubjectivity - the tasks should involve candidates both as language receivers and language producers. In addition like language produced by the candidate, should be modified in accordance with what their expectations of the addressee are perceived to be

- Scope for development of activity by the candidates - the tasks should give candidates the chance to assert their communicative independence and allowance should also be made for the creative unpredictability of communication in the tasks set and the marking schemes that are applied.
- Allowance for self monitoring by candidates - the tasks should help candidates to use their discourse processing strategies to evaluate their communicative effectiveness and make any necessary adjustments in the course of an event.
- Processing of appropriately sized input - the size and scope of the activities should be such that they are processing the kind of input they would normally be expected to .
- Normal time constraint operative - the tasks should be accomplished under normal time constraints.

2.3.1. Communicative Testing of Reading

There are different points to consider when one plans to test reading communicatively. Broadly speaking, these points include what to test exactly and how to go about it.

In general, when we want to test reading, we should first specify what the candidate should be able to do, says Hughes (1989). In other words, we should be clear about the "population of abilities" we want to test.

We should specify content (like operation, types of text, addressees, and topics) and criterial levels of performance under operations, Hughes discusses the following macro-skills

- Scanning text to locate specific information
- Skimming text to obtain the gist
- Identifying examples presented in support of an argument

Underlying these are 'micro-skill' such as:

- Identifying references of pronouns etc.
- Using context to guess meaning of unfamiliar words
- Using relations between parts of a text by recognising indicators in discourse , especially for the introduction, development, transition and conclusion of ideas.

The other things we should take into consideration according to Hughes are types of text, addressee and topics.

Types of text might include "textbook, novel, magazine, newspaper (tabloid or quality), academic journal, letter, time-table, poem, etc. They might be further specified, for example newspaper report, newspaper advertisement, newspaper editorial (p.118).

Here the problem of backwash should be borne in mind. Hughes mentions this because if we stick to a limited range of text types students may be encouraged to also stick to that limited range of text types.

We should also consider authenticity of the texts we choose. Authentic texts (meant for native speakers) should be used. With appropriate items, such texts can be used for students at any level (including very low levels).

In selecting texts we should also take into consideration the addressee, i.e. what kind of audience the text was originally meant for.

Finally, Hughes says we should indicate topics, for which it may be enough to use only very general terms.

Now let's look at methods.

There are different testing methods that can be used to test reading communicatively. Weir (1990:43-51) discusses the following.

Multiple choice Questions (McQs). On this method the candidate is presented with a number of option, only one of which is correct and he or she is required to select the correct option. Marking here is entirely objective.

Short answer Questions- In this method the candidate is required to write down answers for specific questions .

Cloze. In this method words are deleted from a text after allowing a few words of introduction. The deletion is mechanical, usually between every fifth and eleventh word. Weir says Alderson in a research (1978) found that semantically

acceptable scoring procedure to be superior to any other. This method is well acclaimed by authorities in the field of language testing. Weir quotes Alderson (1978: 2) for the conclusion he arrived at in his research:

> The last decade, in particular, has seen a growing use of the cloze procedure with non-native speakers of English to measure not only their reading comprehension abilities but also their general linguistic proficiency in English as a Foreign Language.

In the same work Alderson has added that "As a measure of the comprehension of a text, cloze has been shown to correlate well with other types of test on the same text and also with standardized testing of reading comprehension" (p.39).

Selective deletion gap filling - In this method the selection of items to be deleted is not mechanical as in cloze procedure; rather it is based upon "what is known about language, about difficulty in text and about the way language works in communication" (p.48).

C- Tests - This method is a relatively new method. It is an adaptation of the cloze method and has been developed in Germany by Klein - Braley (1981, 1985; Klein-Braley and Raatz, 1984). In this method a text is selected and every second word is partially deleted. Students are given the first half of the deleted word (to ensure solution). The test taker has to give the whole word in the question paper.

Cloze elide - In this method words which do not belong are inserted into a reading passage and candidates have to indicate where these insertions have been made. Earlier this method was known as the "intrusive word method".

Information transfer - In this method candidates are required to transfer verbally presented information to a non verbal form e.g. by labelling a diagram, completing a chart or numbering a sequence of events.

In his extended discussion of these methods of testing reading comprehension, Weir(1990) has given the advantages and disadvantages of all. He concludes by saying:

> If we are to develop the communicative nature of our tests it is perhaps important to focus on performance tasks in reading tests, and the use of information transfer and other restricted response formats is advocated (p.51).

2.3.2. Communicative Testing of Writing

Hughes (1989) says the best way to test writing is to get the test writers write. And Weir (190) says "two different approaches to assessing writing ability can be adopted" (p.58). The first one is dividing writing into discrete levels, e.g. grammar, vocabulary spelling and punctuation, which can be tested by means of objective tests. The second one is by constructing "direct extended writing tasks of various types."

When we plan to test writing ability directly

1. We have to select writing tasks that are properly representative of the population of tasks that we should expect the students to be able to perform.
2. The tasks should elicit samples of writing which truly represent the students' ability.
3. It is essential that the samples of writing can and will be scored reliably (Hughes, 1989:75).

In setting the task, we should first specify all appropriate tasks and select a sample. We should have a clear idea of what tasks among the list are those we think the students should be able to perform. Here all the discussion about operations, types of text, addressees, and topics, discussed under the sub topic" communication Testing of Reading", (2:3.1) above, is relevant.

Weir (1990) says writing tests can be used as indirect ways of assessing linguistic competence. An example given for this is editing task.

In the editing tasks candidates are required to rewrite a text with errors of grammar, spelling and punctuation making all the necessary correction. Usually the errors are common errors noted to exist with the target group.

For the direct testing of writing, a more integrative approach could be considered. Hence we can incorporate items which can test a candidates "ability

to perform certain of the functional tasks required in the performance of duties in the target situation" (Weir,1990:60).

Weir also discusses the following methods:

Essay Tests - this method requires candidates to produce a sample of connected writing. Essay tests usually have a stimulus which can vary in length" from a limited number of words to several sentences" (p.60). As the topics are of a general type, candidates are usually expected to depend on their own for ideas. These tests are also free in nature. I.e. the candidates receive no guide as to how they have to go about doing the writing.

Controlled Writing Tasks - In this method candidates are given a "stimulus that can be written, spoken, or must effectively non-verbal, e.g., a graph, plan or drawing which the student is asked to interpret in writing.: (p.61)

Summary - this method requires candidates to identify"relevant facts from a mass of data and to combine them in an acceptable form" (p.61). It also requires "the ability to write a controlled composition containing the essential ideas of a piece of writing and omitting non-essentials" (p.61).

Here again, just like the methods for testing reading comprehension, Weir discusses the advantages and disadvantages of these methods of testing writing. Weir's recommendation is that "the writing component of any test should concentrate on controlled writing tasks where features of audience, medium, setting and purpose can be more clearly specified" (p.73).

2.4. Validity

Defined broadly, validity refers to whether a given test measures exactly what it is supposed to measure (Heaton 1988).

The question whether a given test is valid or not is a central concern to all testers. This is because, according to Alderson et al: (1995:170), "if a test is not valid for the purpose for which it was designed, then the scores do not mean what they are believed to mean" A very precise definition of validity is given by Henning (1987), which is quoted in Alderson et al (1995:170):

> Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is supposed to measure. It follows that the term valid when used to describe a test should always be accompanied by the preposition for. Any test then may be valid for some purposes, but not for others.
>
> (Henning; 1987:89)

Alderson et al also go on to say that we don't rule out the possibility of a test being valid for more than one purpose. But

> If a test is to be used for any purpose, the validity of use for that purpose needs to be established and demonstrated. It is not enough to assert "This test is valid, unless one can answer the following question 'How do you know?" and "For what is it valid?"

2.4.1. Types of validity

The testing literature discusses the different types of validity that go under a lot of different and, according to Alderson et al (1995:71), 'a confusing array" of names. This is to say that there usually is difference in the use of names to refer to different types of validity, although the content of the discussion is basically similar. So in this part of the paper the discussion of the different types of validity these references are used: construct, content face, and criterion related.

2.4.1.1. Construct Validity

According to Weir (1990:22). "The most helpful exegesis regards construct validity as a superordinate concept embracing all other forms of validity." He also quotes Chromback (1971:463) who commented that "Every time an educator asks 'but what does the instrument really measure?" he is calling for information on construct validity." Furthermore, he presents Anastasi's definition of construct validity:

> The extent to which the test may be said to measure a theoretical construct or trait. Each construct is developed to explain and organize observed response consistencies...Focusing on a broader more enduring and most abstract kind of behavioural description...construct validation requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestation are grist for this validity mill.

To make the meaning of the term "construct", Alderson et al (1995:183) quote Ebel and Frisbie's (1981:108) explanation of the term:

> The term construct refers to a psychological construct, a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly. Examples of construct are intelligence, achievement, motivation, anxiety, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the markers intend to measure. The goal is to determine the meaning of the scores from the test, to assume that the scores mean what we expect them to mean.

Gronlund (1985:58) is the other scholar quoted by Alderson et al (1955) for his "shorter explanation" of construct validity. According to him construct validation measures "how well test performance can be interpreted as a meaningful measure of some characteristic or quality."

In their detailed discussion of this validity type, Alderson and his colleagues present how it can be seen in light of comparison with theory and considering internal correlations.

They say (p.183) according to some test theorists this validity is primarily concerned with knowing to what extent the test has been effective in reflecting a certain underlying theory. The procedure followed to check this is: Selecting test experts, giving them the test and the definition of the "underlying theory", and asking them to inspect both and give judgement.

The consideration of internal correlation follows this procedure: component parts of the test are correlated and a fairly low correlation in the possible order of +.3 - +.5 is expected. Very high correlation (e.g. +.9) would indicate that the two

components are essentially testing the same thing and we might consider dropping one.

The other type of correlation considered is that between "each subtest and the whole test." This correlation is "expected at least according to classical test theory, to be higher possibly around +.7 or more - since the overall score is taken to be a more general measure of language ability than each individual component score". (p.184).

2.4.1.2. Content Validity

A test is said to have content validity "if its content constitutes a representative sample of the language skills, structures, etc, with which it is meant to be concerned" (Hughes, 1989:22) Kerlinger (1973:458) quoted in Alderson et al. (1995:173) also says "content validity is the representativeness or the sampling adequacy of the content - the substance, the matter, the topics - of a measuring instrument." Similarly, Morrow (1979:147) citing Davies (1968) says content validity refers to whether "the test accurately reflects the syllabus on which it is based." Weir also comments on a similar line: "The more a test simulates the dimension of observable performance, the more likely it is to have content...validity" (Weir,1990:24).

Hughes (1989) says to ensure content validity we need to plan proper sampling, which, in turn, depends on the purpose of the test. The other requirement he mentions is a specification of the skills, structures, etc meant to be covered by the test, the basis of judgement of the test's content validity is a

comparison of the specification and the test content. This involves, according to Alderson et al (1995) "gathering the judgement of 'experts': people whose judgement one is prepared to trust, even if it agrees with one's own." (P.173). They say a common way for the experts to do this is "to analyse the content of a test and to compare it with a statement of what the content ought to be." They also say such a statement may be "the test's specification it may be a formal teaching syllabus or curriculum, or it may be a domain specification: (p.173).

To ensure content validity, Weir (1990:24) says, during test construction "particular attention must be paid to content validity in an attempt to ensure that the sample of activities to be included in a test is as representative of the target domain as possible.

2.4.1.3. Face Validity

Morrow (1979) citing Davies (1968) defines face validity to refer to whether "the test looks like a good one: (p.147) Ingram (1977:18) says face validity "has to do with the surface credibility and public acceptability of a test." Weir (1990:26) quotes Anastasi (1982:136) who pointed out that face validity "is not validity in the technical sense; it refers not what the test actually measures but to what it appears specifically to measure."

But is face validity important? This question is approached in a number of ways in the testing literature.

Alderson et al. (1995) say face validity "is frequently dismissed by testers as being unscientific and irrelevant" (p.172). Weir also mentions people who have discounted face validity: Lado (1961), Davies (1965), Ingram (1977), Palmer (1981), and Bachman and Palmer (1981), and quotes Bachman and Palmer's (1981:55) argument against face validity:

> Since there is no generally accepted procedure for determining whether or not a test demonstrates this characteristic, and since it is not an acceptable basis for interpretive inferences from test scores, we feel it has no place in the discussion of test validity.

Bachman (1990) also writes a "post mortem" to face validity. Ingram (1977:18) says face validity "while it is sometimes important, it can be regarded as a public relationship problem rather than a technical one."

On the contrary, discussing the problems of a test that has no face validity, Weir (199-:26) says:

> If a test does not have face validity...it may not be acceptable to the students taking it, or the teachers and receiving institutions who may make use of it. If the students do not accept it as valid their adverse reaction to it may mean that they do not perform in a way which truly reflects their ability.

Anastasi is also quoted in Weir as saying that whatever its actual validity, if a test's content appears to be "irrelevant, inappropriate silly or childish" it will result in poor cooperation from the students. So in addition to being "objectively

valid" a test also needs face validity "to function effectively in practical situations" (Anastasis 1982:136).

After discussing the reactions of testee towards the concept of face validity, Alderson et al. (1995:173) present their own position regarding face validity which falls in line with Weir's and Anastasi's. They say

> ...face validity is important in testing. For one thing, tests that do not appear to be valid to users may not be taken seriously for their given purpose. For another, if test takers consider a test to be valid, we believe they are more likely to perform to the best of their ability on that test and respond appropriately to items.

The conclusion reached by Weir (1990) is that though face validity is important, it should not be taken as a substitute for "objectively determined validity" and that "the validity of the test in its final form should always be directly checked" (p.26).

2.4.1.4. Criterion -Related Validity

Criterion - related validity refers to the correlation of the test against "some independent and highly dependable assessment of the candidate's ability" (Hughes, 1989:23). This validity type is "a predominantly quantitative and a posteriori concept concerned with the extent to which test scores correlate with a suitable external criterion of performance" (Weir, 1990: 27). Ingram (1977:18) calls this validity type "pragmatic validity".

Criterion related validity is usually divided into two types: concurrent and predictive validity.

To determine concurrent validity, the test scores are correlated with another measurement of performance, usually an older established test, taken at roughly the same time as the test (Alderson et al. 1995:177; and Kelly, 1978; and Davies, 1983; both quoted in Weir, 1990). According Alderson et al. (1995: 177-78), these other measures could also be "scores from parallel versions of the test or from some other test." Or they could also be "the candidates' self-assessment of their language abilities, or ratings of the candidate on relevant dimension by teachers, subject specialists or other informants." Predictive validity differs from concurrent validity in that, as the name suggests the external data to correlate test results with will be collected some time often the test has been given, or with some future criterion of performance (Alderson et al. 1995; Bachman and Palmer, 1981 quoted in Weir, 1990).

The problem with criterion related validity is, according to Weir (1990), that it gives no information regarding the test's construct validity. In other words, we can have a test with a perfect criterion -related validity without knowing what the test is measuring.

The other problem with this type of validity is that one might be forced to put his "faith in a criterion which may itself not be a valid measure of the construct in question." This means one cannot say a test has criterion -related validity just

because it correlated highly with another test if the other test doesn't measure the construct in question" (Weir, 1990:28).

2.5. Reliability

Reliability can broadly be defined as a test 's quality to produce similar results consistently (Harrison, 1983) or simply as "consistency of measurement" (Bachman and Palmer, 1996:19).

Reliability is a fundamental criterion against which any language test has to be judged (Weir, 1990:31). If consistent results are produced, that means we can depend on them.

A reliable test score, according to Bachman and Palmer (1996:19-20) will be "consistent across different characteristics of the testing situation. This reliability can be considered to be a function of the consistency of scores from one set of tests and test tasks to the other."

For Bachman (1990), reliability is decisive, for

> a fundamental concern in the development and use of language tests is to identify potential sources of error in a given measure of language ability and to minimize the effect of these factors on that measure (p.60).

Three types of checking reliability are discussed in Weir (1990:31-32). These are: inter-and intra-marker reliability, internal consistency of sub-tests, and parallel forms reliability.

Inter-marker reliability can be checked by getting two markers mark the same test and correlating the results. Each marker's consistency can be checked by getting him mark the same test or component at a later date. Agreed upon marking guidelines are needed and training markers may be required to enhance the agreement between the different markers.

The second type of reliability checks to what extent sub- test are internally consistent. This is usually checked by using special formulae like the Kuder-Richardson formula which is most commonly used.

The third type of reliability is considered the most difficult to achieve. It requires the preparation of two exactly identical tests which are in effect "clones of each other," as Weir (1990:32) calls them. The problems are both theoretical and practical in this third type of reliability, the reliability of the test versions is directly proportional to the similarity of the results. And the two tests have to be administered to the same test population. The other type of reliability which is less frequent is estimated by administering a single test to the same test population twice. This is called the "test-retest method."

Hughes (1989 36-41) gives useful suggestions on how to make tests more reliable.

- take enough samples of behaviour
- do not allow candidates too much freedom
- write unambiguous items.

- provide clear and explicit instruction
- ensure that tests are well laid out and perfectly legible
- candidates should be familiar with format and testing techniques
- provide uniform and non-distracting conditions of administration
- use items that permit scoring which is as objective as possible
- make comparisons between candidates as direct as possible
- provide a detailed scoring key
- train scorers
- agree on acceptable responses and appropriate scores at outset of scoring
- identify candidates by number, not by name.
- employ multiple, independent scoring

2.6. Relationship between Validity and Reliability

How are validity and reliability related?

Hughes (1989) says if a test is to be valid, it has to produce consistently accurate measurements, or, in other words, it has to be reliable. But "a reliable test," he says, "may not be valid at all" (p.42).

So when people aim at increasing the reliability of their test, they may in reality be reducing its validity. This is why scholars talk about a reliability - validity "tension" (Guilford, 1965; and Davies, 1978; both quoted in Weir,1990).

Weir comments: "The problem is that while one can have a test reliability without test validity a test can only be valid if it is also reliable" (p.33). Hughes (1989) also warns that "in our efforts to make tests reliable, we must be wary of reducing their validity" (p.42).

So this "tension" between validity and reliability may require testers to sacrifice " a degree of reliability" to enhance validity. However, Weir warns, "if validity is lost to increase reliability we may finish up with a test which is a reliable measure of something other than what we want to measure." (P.33).

Discussing Rea's (1978) ideas, Weir (1990) makes the tension between validity and reliability mentioned above easier to understand. He says if we consider discrete types and communicative tests, the former are likely to have high reliability with a suspect validity while the latter though may claim high validity, the reliability may not be expected to be as high as the former. He also goes on to emphasise that just because communicative tests do not claim automatic reliability like discrete tests, this should not be taken on a justification to totally defend on discrete type tests.

On the other hand, some scholars suggest how we might consider the relationship between reliability and validity, especially when dealing with language tests. Weir (1990:33) quotes Moller (1981:67):

> While it is understood that a valid test must be reliable, it would seem that in such highly complex and personal behaviour as using a language other

> than one's mother tongue, validity could be claimed for measures that might have a lower than normally acceptable level of reliability.

Bachman appears to take a different line of looking at validity and reliability. He doesn't seem to like the existing discussion which considers them as two separate and distinct concepts. Rather, he thinks, a better understanding of the two can be achieved through considering them as "complementary aspects of a common concern in measurement..." (P.160). Accordingly validity and reliability ate "two complementary objectives in designing and developing tests"; reliability aims at minimizing "the effects of measurement error", and validity aims at the maximizing "the effects of language abilities we want to measure "

But through we can look at the concepts of validity and reliability as complementary concepts, there seems to be a "tension" between the two - the tension that one is affected when the other is increased. This means that the two concepts are as Weir (1990:33) says, in certain circumstances "mutually exclusive". But, Weir concludes, "if a choice has to be made validity, after all, is all the more important."

CHAPTER THREE

METHODOLOGY

3.1. Subjects

The subjects used for the study were first year students of the Addis Ababa University. As the study focused on first year students as a whole and not particular groups, the method of selecting the subjects was random sampling. The study was conducted by taking as subjects students of three sections. Out of these, two sections were from the social sciences campus and one section was from the natural sciences campus. The total number of students taken as subjects was about 120.

3.2. Instruments

Three instruments were used in the study. The first and the most important instrument used to collect data for the study was the test that was prepared and administered for the three sections mentioned above.

Secondly, a questionnaire was prepared for the students who took the test. The questionnaire was to be completed by the students right after they had taken the test. The purpose of the student questionnaire was primarily to determine the face validity of the tests. About 10 statements that were to be completed on five scale basis and two open ended questions were included in the questionnaire.

The third instrument was a questionnaire prepared for instructors. The instructors were selected because they are experts in the field of language testing

or have experience in relation to language teaching and /or test preparation. The instructor's questionnaire aimed to help determine the test's content, and face validities.

3.3. Procedure

3.3.1. Test Preparation

The test used in the study was prepared after reviewing relevant and available materials on the area of (language) testing, particularly on the areas of communicative language testing.

The texts included in the test were selected from different books. The following things were done during the preparation and after the test was pretested on a section.

- A very long passage was shortened to a relatively more manageable size without distorting the sense of the original passage as much as possible.
- Items for a passage were constructed by making reference to sample papers selected by authorities in the field of language testing (especially Weir, 1990).
- A complete chart was described in detail and points were deleted from both the chart and the detailed description to produce an information transfer activity.
- A passage was modified so that it became a contextualized test of the use of prepositions and tenses.
- A diagram was modified to make it more meaningful and easier to understand.
- A text was included in the test with no modification.

After the test was prepared it was administered for about 120 students in three sections in the social and natural sciences campuses of the University.

Following the checklist given by Alderson et al. (1995:38) we can present the following paragraph as a summarized specification of the test.

Generally speaking the test is a proficiency test designed in the communicative paradigm. The test has 8 subsections under 3 major sections. There are four written texts in the test. The language skills tested are reading and writing and the subskills vocabulary and grammar. The tasks in the test are integrated types rather than discrete ones. The testing methods used are open ended (short responses) matching (listing), information-transfer, blank-filling, summarizing, description, and deciding on forms given in brackets.

3.3.2. Questionnaires Preparation

The items in the two questionnaires were included based on ideas from different books on testing. The items were worded so that they can as much as possible help to find out whether the test has the different types of validity.

3.3.3. Data Analysis

Once the data was gathered with the help of these instruments, it was organized and analysed as follows.

Firstly, the test papers were corrected by three markers. The markers were asked to give points to the students' responses not by writing in the test papers but by using separate papers. Next the papers on which the markers wrote the points for the items in the test were collected and the decision was made on the basis that a student would be given a point for an item if at least two of the three markers marked it right. For the connected writing parts, the decision was made that a student's points would be the average of the points given by the three markers.

The next thing done was to conduct item analysis for those items marked relatively objectively. To do this the total scores the students got were listed from top to bottom. Then the middle 1/3 were discarded. This process reduced the number of papers to be included in the item analysis to 80.

Next all the individual items were tallied on the basis of 1 and 0 for the right and wrong responses respectively. Once the tallying was done the difficulty level or facility value and the discrimination index for each item were calculated. The reliability of the test was also calculated using the split-half method. The Kuder-Richardson formula 21 (KR 21) was used to also check the interitem consistency in the test.

The responses of the students to the student questionnaire were tallied and the average or the mean for each item was calculated. The responses of the instructors to the instructor questionnaire were also described.

CHAPTER FOUR

RESULTS AND DISCUSSION

As mentioned in chapter three a test that is communicative in nature was prepared and administered for three freshman sections in the university. Questionnaires were also prepared both for students and instructors. The results from the different instruments are presented and discussed below.

4.1. Data from the Student's Questionnaire

Right after the test was administered, the student questionnaire was administered to find out to what extent the test had been viewed and accepted positively (or negatively) by the students who took it.

Out of the 120 students who took the test and completed the questionnaire the papers of those students who were selected according to their performance for the item analyses were picked out, and their responses were coded and tallied. Next to this the responses given to each item were added together and the mean was calculated for each item. The summary of the students' responses to the items in the Student's questionnaire is found in Table 1.

Item No	$\sum X$	$\bar{X}$
1	330	4.125
2	270	3.375
3	328	4.1
4	312	3.9
5	370	4.625
6	331	4.13
7	378	4.125
8	328	4.1
9	342	4.275
10	350	4.375

Table 1. *Summary of Students' responses to the items in the student questionnaire*

The procedure followed to arrive at the average points of the responses is this. The positively worded statements were coded starting from 5 and going down to 1, while the negatively worded statements were coded starting from 1 and going up to 5. This means that when the average response to each item was calculated it would show the general opinions of the students regarding the item. The result would thus be interpreted so that 5 would mean 100% positive opinion and 1 would mean 100% negative opinion.

As shown in the table, the points range between 3.37 and 4.725. Eight items had their average above 4. Only two items (Nos 2 and 4 had below 4, 3.375 and 3.9 respectively). The average of the mean scores was found to be 4.17.

So in general we can say that the test was positively accepted by the students who participated in the research and who were selected as representatives.

The purpose of the questionnaire was to check the face validity of the test. As mentioned in Chapter two, if a test appears good to students and if it is acceptable to the students who take it, then they would cooperate satisfactorily and they would perform to the best of their abilities. This in turn would mean that the scores we get from the test would mean what we believe them to mean.

Consequently, we can deduce from the table that the test was accepted by the students fairly positively. This helps us to accept the results the students got in the test as dependable data. Talking about the items in the questionnaire in more specific terms, we can say this. The students have found the appearance of the test acceptable. From items number 2 and 3 we can say that even though the students have felt it was a long test, they have found the variety of tests in the test motivating and enjoyable. They also have agreed that 3 hours was enough to complete the test. They have found no problem in understanding the instructions, and the challenging level of the test was also reasonable in the eyes of the test-takers. They have also felt the test was a good instrument of testing reading and writing.

The responses the students gave to the two open ended questions at the end of the questionnaire were essentially similar. While nobody in particular mentioned anything in particular as a problem with the test, in their responses to the second question almost all the students talked about themselves. In short their responses say that they had not had enough experience of taking tests such as this one and they had been accustomed to multiple choice formatted tests predominantly. Some even felt the need to recommend that such tests start being

given at least at the highschool level. Most also commented that it really was an interesting experience taking part in the research.

In general, it was surprising to find that almost all the students preferred to talk about themselves instead of about the test, with the exception of two students who said sitting for three hours to take a test was a little bit tiresome.

4.2. Data from the Instructor's Questionnaire

As mentioned earlier, instructors were also asked to comment on the test. These instructors were given the questionnaire along with a copy of the test so that they could have a good look at the test and comment on it.

The responses of the instructors can be summarized like this. In general they agreed to accept the test as a reasonable communicative language test. The test can also test the skills of reading and writing adequately at the freshman level. Reasonably enough amount of tasks of the major and subskills are included in the test. All instructors disagreed with the item that says "unnecessarily too much weight is given to reading" (item no.6). According to the respondent instructors if the test is given at the end of the semester if could not be difficult. The time and effort spent to mark such tests was not viewed as "unnecessary" by the respondents. For item number 10, a comment was given that the test reflects the requirements of university education but specifically that of the freshman level, otherwise it might be "claiming too much". The sampling was accepted as adequate but still it could be improved. The purpose of the test is clear.

Cheating, especially where continuous writing is involved, is minimized. The testing of grammar points by means of contextualized passage was appreciated. The time (3 hours) was enough. Processing information presented both implicitly and explicitly was involved but still it could be improved, according to the respondents. The reading passage, according to them, could not be described as "unnecessarily too difficult", though it is challenging. The test was too long. There is likely to be subjectivity in marking "unless appropriate methods are followed." The test is not too easy and the writing part is not complex as clear instructions and first sentences were given. We cannot also say it is too easy (i.e. the writing). Giving judgement regarding a student's level of performance based on the result he or she gets in the test can agreeably be acceptable but only if the level is specified and the decision is restricted to the skills involved.

Returning to relatively more negative comments we get the following. One of the respondent instructors said the test lacks "thematic approach" which resulted in the students having to spend a long time reading a number of texts. In other words, the test could in general have been about one central point (e.g. about comets). One instructor described this "theme" as a "hook" that could have served to carry all the items. There was also a comment questioning the basis of weighting.

Although all the responses and comments given by the respondent instructors were helpful, meaningful and interesting, we could bring into this discussion the following points of justification.

The first one comes from the response obtained from the students regarding questionnaire item no.3 which says "the variety of texts in the test make the test motivating/enjoyable". The average point obtained was 4.1. It could arguably be said that motivation and enjoyment contribute a good deal to the face validity of a test. On a similar line, we can also mention Hughes (1989:119) who recommends the inclusion of as many texts as possible so that the students get "a good number of fresh starts". According to him having many texts in a test also helps to obtain "acceptable reliability" (p.119).

4.3. Data from the Item Analysis

After the test was given the 120 papers were marked by three markers. For the items that could be marked relatively more objectively the decision was made on the basis that in order for a student to get a full point on a given item he or she had to be marked correct by at least two of the three markers. If two or three of the markers decided a certain answer was wrong then the student would be given no point for that item. For the connected writing exercises the points were taken to be the averages of the points given by the three markers.

Next the marks were added and listed top to bottom. The next thing done was to decide on the top 1/3 and the bottom 1/3, which led to the discarding of about 40 papers. The remaining 80 papers (40 high and 40 low scorers) were included in the items analysis list.

In the item analysis list, only those items that could be relatively more objectively marked were included. First the items were coded as shown in Table 2.

Item Description	Code	No.of items
▸ Passage Open ended questions (understanding reference, interpretation etc.)	PO	10
▸ Paragraph topic sentences (Summaries)	PS	8
▸ Passage vocabulary (understanding contextual meanings of words)	PV	15
▸ Reading for information transfer	IT	20
▸ Writing stages of a process (writing brief summaries)	WSP	5
▸ Using prepositions appropriately	GRP	7
▸ Using tenses appropriately	GRT	15

Table 2. *Codes used in the item analysis*

The total number of items in the item analysis was 80. For all the items the responses of the students included in the item analysis were tallied on the basis of 1 and 0 for right and wrong responses respectively.

Next the following were calculated:

- the number of correct responses
- the number of wrong responses
- the number of students in the upper group who got the item right (RU)
- the number of students in the lower group who got the item right (RL)
- the difficulty level or facility value of the item (F.V.)
- the discriminating power or index of the item (D.I.)

Table 3. Presents the results obtained.

Item Code	No. Of Rights	No. Of Wrongs	RU	RL	F.V.(%)	D.I.
PO1	47	33	31	16	58.75	0.375
PO2	39	41	29	10	48./75	0.475
PO3	27	53	18	9	33.75	0.225
PO4	56	24	35	21	70.00	0.35
PO5	46	34	29	17	57.50	0.3
PO6	45	35	24	21	56.25	0.075
PO7	48	32	31	17	60.00	0.35
PO8	35	45	24	11	43.75	0.325
PO9	28	52	15	13	35.00	0.05
PO10	48	32	28	20	60.00	0.2
				Average	**53.36**	**0.27**
PS1	59	21	39	20	73.75	0.475
PS2	51	29	32	19	63.75	0.325
PS3	50	30	28	22	62.50	0.15
PS4	54	26	35	19	67.50	0.4
PS5	58	22	36	22	72.50	0.35
PS6	49	31	34	15	61.25	0.475
PS7	45	35	32	13	56.25	0.475
PS8	51	29	32	19	63.75	0.325
				Average	**65.22**	**0.37**
PV1	49	31	28	21	61.25	0.175
PV2	33	47	24	9	41.25	0.375
PV3	32	48	23	9	40.00	0.35
PV4	51	29	35	16	63.75	0.425
PV5	37	43	26	11	46.25	0.375
PV6	38	42	21	17	47.50	0.1
PV7	42	38	26	16	52.50	0.25
PV8	31	49	16	15	38.50	0.025
PV9	48	32	31	17	60.00	0.35
PV10	49	31	33	16	61.25	61.25
PV11	53	27	36	17	66.25	66.25
PV12	52	28	33	19	65.00	65
PV13	54	26	37	17	67.50	67.5
PV14	41	39	28	13	51.25	51.25
PV15	30	50	22	8	37.50	37.5
				Average	**53.33**	**0.33**

Item Code	No. Of Rights	No. Of Wrongs	RU	RL	F.V.(%)	D.I.
IT1	56	24	36	20	70	0.4
IT2	64	16	39	25	80	0.35
IT3	52	28	36	16	65	0.5
IT4	52	28	37	15	65	0.55
IT5	54	26	38	16	67.5	0.55
IT6	55	25	38	17	68.75	0.525
IT7	57	23	38	19	71.25	0.475
IT8	50	30	35	15	62.5	0.5
IT9	44	36	34	10	55	0.35
IT10	41	39	31	10	51.25	0.5
IT11	53	27	38	15	66.25	0.575
IT12	48	32	35	13	60	0.55
IT13	56	24	39	17	70	0.55
IT14	51	29	37	14	63.75	0.575
IT15	48	32	36	12	60	0.6
IT16	50	30	36	14	62.5	0.55
IT17	51	29	36	15	63.75	0.525
IT18	53	27	36	17	66.25	0.475
IT19	45	35	34	11	56.25	0.425
IT20	45	35	32	13	56.25	0.475
				Average	**64.06**	**0.5**
WSP1	60	20	34	26	75	0.3
WSP2	50	30	34	16	62.5	0.45
WSP3	37	23	35	22	71.25	0.325
WSP4	61	19	37	24	76.25	0.325
WSP5	70	10	39	31	87.5	0.11
				Average	**74.5**	**0.3**
GRP1	34	46	27	7	42.5	0.5
GRP2	43	37	31	12	53.75	0.475
GRP3	44	36	30	14	55	0.5
GRP4	35	45	25	10	43.75	0.375
GRP5	39	41	27	12	48.75	0.375
GRP6	37	43	27	10	46.25	0.425
GRP7	25	55	17	8	31.25	0.225
				Average	**45.89**	**0.41**
GRT1	22	58	15	7	27.5	0.2
GRT2	17	63	14	3	21.25	0.225
GRT3	28	52	15	13	35	0.05
GRT4	37	43	24	13	46.25	0.275
GRT5	36	44	25	11	45	0.35
GRT6	24	56	14	10	30	0.1
GRT7	19	61	14	5	23.75	0.225
GRT8	52	28	33	19	65	0.35
GRT9	45	35	29	16	56.25	0.325
GRT10	20	60	12	8	25	0.15
GRT11	24	56	17	7	30	0.25
GRT12	28	52	20	8	35	0.3
GRT13	30	50	23	7	37.5	0.4
GRT14	41	39	25	16	51.25	0.225
GRT15	30	50	19	11	37.5	0.2
				Average	**37.75**	**0.25**
				Overall Average: 54.82		**0.36**

Tabel 3 *Results obtained in the item analysis*

The difficulty level of the items was calculated using the formula

$$F.V. = \frac{R}{T} \times 100$$

Where R = the number of students who got the item right and

T = the number of students who tried the item.

The average facility value (F.V.) For an item is 50% so values near 50% are expected. The greater the value from 50 the easier the item is, and vice versa; i.e., the lower the value from 50 the more difficult the item is. 100% facility value would mean that the item is so easy that all the students have got it right and 0% facility value would mean that the item is so difficult that there was not a single student who got it right. Both extremes suggest problem with the item.

In the test used for this study, the easiest item was WS P5 with an F.V. of 87.5% and the most difficult item was GRT2 with an F.V. of 21.25% WSP5 is part of the writing brief summaries activity and GRT2 is part of the grammar activity that required students to use tenses appropriately.

Out of the 80 items in the list, 53 had their F.V. above 50 and 27 had their F.V. below 50.31 items had their F.V. between 40% and 60% and 68 items had their F.V. between 30% and 70% . 8 items had their F.V. above 70% and 4 items had their F.V. below 30%. The overall average F.V. was 54.82%.

From the different values obtained regarding the facility values of the individual items we can say that the items to a larger part have performed

acceptably. With 68 of the 80 items (85%) having facility values between 30 and 70 we can conclude that the test has items with fair facility values.

The next thing done in the process of item analysis was to check the discrimination index (D.I.) Of each item. The formula used for this was

$$D.I. = \frac{RU - RL}{\frac{1}{2} T}$$

Where RU = the number of students in the upper group who got the item right

RL = the number of students in the lower group who got the item right

T = the total number of students included in the item analysis list.

The discrimination discussed here is the discrimination between good and poor students. An item is said to be good when it properly discriminates between poor and good students. If the D.I. Of an item is + 1 then it would mean that the item perfectly discriminates. If it is -1 then it would mean that the item has done no discrimination at all. The average D.I. expected is .5. An item has a positive discriminating power when more students in the upper group than the lower group get it right. If the reverse happens, i.e., if more students in the lower group than the upper group get it right, then that item is said to have a negative discriminating power.

In the test used for this study the items with low D.I. were PV6 and GRT6; the D.I. of both was 0.1. This was because almost equal number of students from the two groups got them right. Items IT5, IT8, IT10, and GRP3 had a very good D.I. of 0.5.

The items with the least D.I. were PV8, P09, GRT3 and P06 with D.I. of 0.025, 0.05, 0.05, and 0.075 respectively. The average D.I. was 0.36.

Very low D.I. means there were close number of students from the two groups who answered the items right. In this case PV8 was answered by 16 and 15 students from the two groups, P09 by 15 and 13, GRT3 by 15 and 13 and P06 by 24 and 21. These figures show that these items have not been so successful in their discrimination. There were also other low D.I. values such as 0.1 (PV6,GRT6), 0.11 (WSP5), 0.15 (PS3,GRT10) and 0.175 (PV1).

The reasons for these items to have such low D.I. values is their relatively similar easiness and difficulty for both groups. For example PV6 required students to pick a word which means "essential" from paragraph 3; GRT6 required students to supply the preposition "to". WSP5 is the last stage in the process of making bread and most students wrote simply "baking" and they were right. PS3 is the summary of the third paragraph in the reading passage. GRT 10 is a tense question that expected students to use the past perfect tense of the verb "pass". This item was answered by only 12 students in the upper and 8 students in the lower group. PV1 expected the test takers to pick a word which means "cloudy"

from the first paragraph. This was answered by 28 students in the upper and 21 students in the lower group.

Regarding the rest of the items, whose D.I. values range between 0.2 and 0.6, we can say that their discriminating power is reasonably acceptable because the need to discriminate between students is relatively low at the tertiary level.

4.4. Reliability of the Test

The reliability of the test was checked using two methods. These are the Spearman -Brown Split-half method and the Kuder-Richardson formula 21 (KR-21). The Spearman -Brown Split-half method requires the test split into two reasonably equal halves. The two halves are treated as two tests. In two columns the results students got are listed and the correlation between the two halves is worked out to be followed by the checking of the reliability. Appendix A presents the results of the research subjects in the two halves.

First the correlation of the results in the two halves was checked using the formula

$$r_{xy} = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\ [n\Sigma y^2 - (\Sigma y)^2]}}$$

Which gave

$$= \frac{380772}{581679}$$

$$= \mathbf{0.65}$$

To find out the reliability coefficient of the whole test the following formula was used.

$$r_{xx'} = \frac{2r_{hh'}}{1+r_{hh'}}$$

Which gave

$$= \frac{2 \text{ X } 0.65}{1 + 0.65} = \frac{1.3}{1.65}$$

$$= \mathbf{0.79}$$

Secondly the interitem consistency of the test was also checked using the Kuder-Richardson formula 21 (KR 21). This method helps to know to what extent the items in the test have been internally consistent and homogeneous.

The formula (KR 21) is

$$r_{tt} = \frac{nv - M(n-M)}{(n-1)v}$$

Where r_{tt} = the reliability coefficient

n = the number of items in the test

M = the Mean Score

V = the variance of the test scores

When this formula was used the following result was obtained.

$$= \frac{9085.79}{10537.49} = 0.8$$

According to the literature (e.g. Alderson et. Al. 1995) the results we get when we apply split - half and KR 21 are likely to be close to each other, which is also the case here.

So from the two results obtained, we can conclude that the test has an acceptable reliability with items functioning internally consistently.

4.5. Concurrent Validity of the Test

To check the concurrent validity of the test, the results of the students were correlated with the raw scores they got in College English test given at the end of the first semester of the 1996/97 A.Y. First the results of the students in the two tests were listed as shown in Appendix B.

The following table gives some information about the subjects performance the two tests.

	The Research test	College English- I Test
Average	46	57
S.D	17.11	11.61
Variance	292.66	134.71
Range	14-88	28-78

Table 4. *Summary of the subjects' performance in the two tests*

The pass mark for the research was decided to be 40 while the pass mark for College English was 45. So the average points of both the research and College English test fall in the pass range.

The formula used to find out the correlation between the two was again.

$$r_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\ [n\Sigma y^2 - (\Sigma y)^2]}}$$

Which gave

$$= \frac{1196523}{1899832}$$

= 0.75

This coefficient shows that the research test and the College English -I test, which was accepted as the reasonable representation of the students' overall proficiency, correlated at a good level.

Here a question could be raised as to whether the College English test itself has any properly worked out reliability. So far we have no information regarding College. But it was selected for two reasons. The first one is that since the course is a result of over two years of pilot study and it tries to involve all the four skills of communication, its test could reasonably represent the subjects' overall proficiency in the English language better than any other test. The second one is that exams like the Ethiopian School Leaving Certificate Examination (ESLCE) have a 100% Multiple-choice format and their validity is questionable.

So it could be argued reasonably that it is acceptable to correlate the tests results with those from the College English test.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

In general the whole purpose of this research has been to see if we can really apply communicative language testing at the tertiary level and to see if the test prepared is valid and reliable. The results obtained with the help of the two questionnaires and the test are reasonable enough to conclude that we can really test the language of students at the tertiary level using valid test and obtaining reliable results. And as the items in the test were constructed in line with ideas in the language testing literature, and based on recommendations given, again it would be reasonable to conclude that the test has shown that we can make use of the different ideas and recommendations in the literature and test the language of our students communicatively.

In addition, what is new about this research is the actual checking of the different types of validity of the test by means of questionnaires designed for the students who took the test and language teachers and testing experts and the involvement of adequate amount of statistics that helped to give the test a reasonable ground as a test.

The responses of the students and the instructors to the respective questionnaires helped in determining the test's face, content and construct validity. To check the test's concurrent validity the results the students had obtained in the test had to be correlated with the results they had got in another

measurement. There was at this point a problem of getting a test that was itself properly validated to correlate the results in the test with. But the problem was sensibly solved by selecting the English language test given in the Addis Ababa University freshman level which is a result of a long time survey and which tried to take into consideration all the skills of communication. The test was accepted as valid. Interestingly the results in the two tests showed a good correlation. The Ethiopian School Leaving Certificate Examination (ESLCE) was also considered but it was thought to be inconvenient because, as mentioned earlier, it has a multiple-choice format throughout.

Talking about the test itself its reliability and the items internal consistency were checked using two methods. To check the reliability of the test the Spearman Brown split - half method was used and to also check the interitem consistency of the test the Kuder - Richardson formula 21 (KR 21) was used. Both these methods showed acceptable reliability of the test. So we can conclude here that the test has been a reliable instrument with the items functioning internally consistently.

In general the aim of the test was to help the students make use of their own thinking to process information presented both explicitly and implicitly and produce their own responses. And this aim was acceptably achieved. And though this is not saying the test was a flawless instrument, one can reasonably conclude that it has been effective.

5.2. Recommendations

One can recommend a number of things based on the process followed and the results obtained in this study. Firstly, in general, we can recommend that in order to prepare tests, especially communicative language tests, test writers have to consider a lot of points. As tests are prepared at some institutions, random inclusion of items in tests (as the writer of this paper himself had observed and done) could actually be no testing at all. Items should be thought of properly before they are included in a test. It should be known that each item in the test has its own difficulty level or facility value and power of discriminating between poor and good students. Test writers should be aware of all these.

It should also be known that items, if not properly planned, may not give the required type of information regarding the test takers. Items should not be so difficult that nobody could answer them and so easy that all the students could answer them. It is also recommended to include in a test items with varied levels of difficulty because this will be, as Heaton (1989) says a motivating factor for students, who are at different levels of performance. In other words if the items in a test are all difficult students who are relatively poor will be very likely to be frustrated and demotivated and, on the other hand, if the items are all simple good students will lose interest and will be likely to be demotivated.

Another useful point to recommend is that once tests are constructed, no matter how effectively they may appear to the test writer, as Alderson et. al. (1995) recommend, they have to be pretested on some people to see if they really

are effective as they appear before they are put into actual use. Related to this it would be sensible to recommend that tests should also look like good ones. If students do not take a test seriously to be a good test then their cooperation (probably without their conscious thought) could be minimized. This could have a lot of far-reaching consequences as it is very likely the decisions would be made based on the results students get in the test.

It is clear that a long-time is needed to determine the different types of validity and the reliability of a given test along with the energy spent (not to mention financial and other administrative factors, which are beyond the scope of this paper). But one can reasonably argue that the time and energy spent on such a process results in something worthwhile and should be acceptable as well-spent.

BIBLIOGRAPHY

Alderson, J.C. et.al. 1995. **Language Test Construction and Evaluation**. Cambridge: Cambridge University Press.

Alderson, J.C. and D.Wall. 1993. "Does Washback Exist?" **Applied Linguistics** Vol. 14. No. 2.

Anastasi, A. 1961. **Psychological Testing**. 2nd ed. New York: The Macmillan Publishers Company.

Bachman, L.F. 1990. **Fundamental Considerations in Language Testing**. Oxford: Oxford University Press.

Bachman, L.F. and A.S. Palmer. 1996. **Language Testing in Practice**. Oxford University Press.

Buckingham, T. And R. Yorkey. 1984. **Cloze Encounters**: **ESL Exercises in a Cultural Context.** Englewood Cliffs, N.J. Prentice - Hall, Inc.

Carrol, B.J. 1980. Testing **Communicative Performance: An interim Study**. Oxford: Pergamon Institute of English.

Clark, J.D. "Psychometric Considerations in Language Testing." in B. Spolsky (Ed.). 1978. **Advances in Language** Testing. Series 2 Arlington, Virginia: Center for Applied Linguistics. Pp. 15-30.

Davies, A. 1988. "Communicative Language Testing". In A. Hughes (Ed.). **Testing English for University Study.** Modern English Publications & The British Council.

Glendinning, E. And Mantell, H. 1983. **Write Ideas: An Intermediate course in Writing Skills**. London: Longman.

Gronulund, N.E. and Linn, R.L. 1990. **Measurement and Evaluation in Teaching**. 6th ed. New York: The Macmillan Publishers Company.

Harrison, A. 1983. **A Language Testing Handbook**. London: Longman.

Heaton, J.B. 1989. **Writing English Language Tests**. London: Longman.

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

Ingram, E. 1977. " Basic Concepts in Testing". In J.P.B. Allen and A. Davies (Eds.) **Testing and Experimental Methods**. London: Oxford University Press.

Johnson, K. **Communicate in Writing: A Functional Approach to writing through Reading Comprehension** London: Longman.

Morrow, K.E. 1983. "Communicative Language Testing: Evaluation or revolution?" in K. Johnson and C. Brumfit. (Eds.). **The Communicative Approach to Language Teaching**. Oxford: Oxford University Press.

Oller, J. 1979. **Language Tests at School: A Pragmatic Approach**. London : Longman.

_____. 1987. "Practical Ideas for Language Teachers from a Quarter Century of Language Testing" **Forum** Vol 25 No. 4. Pp 42-46.

Richards, J.C. 1985. **The Context of Language Teaching**. Cambridge: Cambridge University Press.

Swan, M. 1985. "A Critical Look at the Communicative Approach" (1) **ELT Journal.** Vol. 39.No.1 pp 2-12.

Weir. C.J. 1990 **Communicative Language Testing**. New York: Prentice Hall.

Wesche, M.B. 1983."Communicative Testing in a Second Language." **The Modern Language Journal** Vol. 67. NO.1. pp 41-45.

APPENDIX-A *Scores of the subjects in the two halves of the test.*

Subject code	Score in the first half(X)	Scores in the Second half (y)	XY
1	37	34	1258
2	38	31	1178
3	39	28	1092
4	34	31	1054
5	35	30	1050
6	32	32	1024
7	34	30	1020
8	33	31	1023
9	32	29	957
10	32	30	960
11	32	30	960
12	31	30	930
13	30	31	930
14	32	28	896
15	29	31	899
16	29	30	870
17	31	28	868
18	32	27	864
19	32	27	864
20	29	30	870
21	33	24	792
22	33	22	792
23	26	31	806
24	31	26	806
25	28	29	812
26	30	25	750
27	28	27	756
28	30	28	750
29	27	28	756
30	28	27	756
31	28	26	728
32	24	30	720
33	27	27	729
34	29	24	696
35	28	25	700
36	28	25	700
37	44	26	624
38	27	23	621
39	29	21	609
40	21	29	609
81	17	20	640
82	20	17	340
83	19	17	323
84	18	18	324
85	17	18	306
86	21	14	294
87	14	21	294
88	15	18	270
89	17	16	272
90	14	19	266
91	18	15	270
92	16	17	272
93	19	14	266

Subject code	Score in the first half(X)	Scores in the Second half (y)	XY
94	18	14	252
95	19	13	247
96	20	11	220
97	17	14	238
98	16	15	240
99	16	15	240
100	14	17	238
101	14	17	238
102	18	12	216
103	16	14	224
104	20	10	200
105	11	19	209
106	14	16	224
107	14	14	196
108	15	12	180
109	14	13	182
110	14	12	168
111	12	14	168
112	10	16	160
113	9	16	144
114	11	13	143
115	12	12	144
116	14	9	126
117	12	10	120
118	13	8	104
119	13	7	91
120	7	7	49

APPENDIX -B. *The Subjects' Results in the research test and in College English -I test*

Subject Code	Result in the test (x)	Result in College English - I test (y)	xy
1	88	78	6864
2	81	85	6885
3	74	79	5846
4	73	78	5694
5	70	85	5950
6	70	76	5320
7	70	68	4760
8	66	75	4950
9	66	70	4620
10	64	73	2672
11	64	70	4480
12	63	70	4410
13	63	64	4032
14	62	75	4650
15	62	67	4154
16	62	61	3782
17	62	67	4154
18	60	61	3660
19	60	64	3840
20	59	70	4130
21	59	62	3658
22	59	62	3658
23	59	62	3658
24	59	62	3658
25	58	62	3596
26	58	61	3538
27	57	65	3705
28	57	67	3819
29	57	61	3477
30	56	61	3416
31	55	61	3355
32	55	60	3300
33	54	60	3240
34	53	60	3180
35	53	60	3180
36	53	60	3180
37	50	59	2950
38	50	58	2900
39	49	52	2548
40	49	57	2793

Subject Code	Result in the test (x)	Result in College English - I test (y)	xy
81	43	56	2408
82	41	55	2255
83	39	55	2145
84	37	54	1998
85	35	54	1890
86	35	54	1890
87	33	53	1749
88	33	53	1749
89	33	53	1749
90	33	53	1749
91	33	58	1914
92	33	52	1716
93	32	48	1536
94	32	52	1664
95	32	51	1632
96	32	51	1632
97	31	51	1581
98	31	49	1519
99	31	49	1519
100	31	48	1488
101	31	48	1488
102	31	48	1488
103	30	48	1440
104	30	52	1560
105	30	47	1410
106	29	46	1334
107	29	45	1305
108	28	45	1260
109	27	45	1215
110	27	44	1188
111	26	44	1144
112	26	43	1118
113	26	43	1118
114	24	43	1032
115	24	42	1008
116	24	41	984
117	22	41	902
118	22	40	880
119	20	32	640
120	14	28	392

APPENDIX C - STUDENT QUESTIONNAIRE

Addis Ababa University
School of Graduate Studies
Institute of Language Studies
Department of Foreign languages and Literature

Dear, Respondent,

This questionnaire is designed to obtain your opinions regarding the test you just took. Your opinions are very much helpful for the successful completion of a thesis research being conducted to effect the betterment of the testing system of the university at the first year level. The information you provide will be used purely for academic purposes. So please look at the items before responding; and respond to all the items.

Thank you very much for participating in the research.

I.D. No:_____

SECTION:_____

Part I - *Please respond by putting a cross (X) in the most appropriate box.*

No.	Item	Response Categories				
		Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
1.	The test looks like a good one.					
2.	The test was too long.					
3.	The variety of texts in the test make it motivating/enjoyable					
4.	The time allowed for the test was enough					
5.	The test lacks clarity of instruction					
6.	The test was reasonably challenging					
7.	The test was too easy					
8.	The test helps a student to best perform his/her writing skills					
9.	The test helps a student to best perform his/her reading skills					
10.	The test was too difficult					

PART II - *Please answer the following questions*

2.1. Was anything wrong with the test? If so, please mention it.

2.2. Do you have any comments regarding the test?

D

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

INSTITUTE OF LANGUAGE STUDIES

DEPARTMENT OF FOREIGN LANGUAGES AND LITERATURE

Dear Respondent,

This questionnaire is designed to obtain your professional opinions regarding the test attached herewith.

The questionnaire is part of a thesis research being conducted with the aim of preparing a valid and reliable communicative test of reading and writing for the freshman level of the university.

Your opinions and comments will surely render an invaluable help for the successful attainment of the research's objectives.

So please look at the test and complete the questionnaire.

Thank you very much for your time and help.

PART I - Please put a cross (X) in the box that corresponds your response to each item.

No	Item	Response Categories				
		Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
1	Generally speaking the test is a satisfactory communicative test of reading and writing					
2	The test measures reading skills adequately at freshman level.					
3	The test measures writing adequately at freshman level.					
4	The score a student gets in the test can represent his/her ability of the skills of reading and writing.					
5	In terms of tasks the two skills (reading and writing) are given appropriate weight.					
6	Unnecessarily too much weight (mark) is given to reading					
7	The writing tasks amount below what could be regarded as adequate					

8	The test is too difficult for the level of proficiency of freshman students even if it is given at the end of the semester.					
9	The test requires unnecessarily too much marking time and effort					
10	The test adequately reflects the requirements of university education					

No	Item	Response Categories				
		Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
11	The test lacks proper sampling					
12	The purpose of the test is clear					
13	For larger part of the test the chance of students' cheating is highly minimized.					
14	The grammar part is well contextualized					
15	Time allocated for the test is not enough					
16	The test rightly requires students to process information presented both explicitly and implicitly					
17	The reading passage is unnecessarily too difficult					
18	The test is to long					
19	There is likely to be too much subjectivity in marking					
20	The test in general is too easy					
21	The writing part is too complex					
22	The writing part is too easy					

23	Giving judgement about a student level of performance based on the result she/he gets in this test is acceptable/dependable.					

PART II - Please answer the following questions

2.1. Is anything wrong with the test? If there is, please mention it.

_____
_____
_____
_____
_____
_____

2.2. Do you have any comments regarding the test?

_____
_____
_____
_____
_____
_____

APPENDIX - E THE TEST

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
INSTITUTE OF LANGUAGE STUDIES
DEPARTMENT OF FOREIGN LANGUAGES AND LITERATURE

Time Allowed 3 hrs.

ID.No. _____

SECTION _____

`PART I- READING

READ THE FOLLOWING PASSAGE AND ANSWER THE QUESTION SET ON IT (10 marks)

COMETS

1. Man has always regarded comets with awe, fear and superstition because they appear so suddenly and dramatically in the sky. Called "kometes aster" or "hairy star" by the ancient Greeks, comets are nebulous, celestial bodies composed of swarms of metallic particles orbiting in elliptical, parabolic or hyperbolic paths around the sun. Comets consist of a head or coma which is the brightest part, a nucleus, which is the only solid part with a diameter ranging from 1.6 k.m. to 48 k.m. and sometimes one or even two tails - very bright and up to millions of kilometers long. The space behind the nucleus and inside the tail is known as the hollow cone. The tails appear to be straight but they are in fact curved and are only seen when the comet is closest to the sun at perihelion. The tail always points away from the sun so that it always precedes the comet as it recedes into space.

2. A theory that explained comets' tails was devised in 1936 by the German astronomer, Basel. Basel said that a repulsive force acted on the comet particles and varied their gravitational attraction.

3. Although only minor constituents of the solar system, comets have a complex history and are undoubtedly the oldest and best preserved materials in it. Probably critical in its early development. There are three main area in the study of comets: the discovery of new

comets and the recovery of periodic comets. The measurement of their orbits and position is spaces; and a study of their physical aspects. Although there is a lot we do not yet know, there is too much we have learned about comets and their behaviour. We know there are two types of comet, classified according to their orbital period. Comets less that 200 years old are called short period comets, those over 200 years old are called long period comets.

4. A very famous astronomer called Halley showed that some comets were periodic, returning after a number of years. He determined the orbit of the first periodic comet which bears his name. He deduced that five comets which had appeared from 1378 to 1682 were successive returns of the same comet. Halley's comet, first seen in 1682, has an orbit of 76.09 years. The date of April 13th to 1759 was given as its perihelion passage, with a month either side for error. It returned in December 1758 and reached perihelion on 12th March 1759; in 1835 it returned again, within three days of the date predicted, and appeared yet again in 1910. After Halley's death in 1742, it was shown that the comet had made twenty-nine appearances since 239 BC and that it was probably not only the comet depicted on the Bateaux tapestry in connection with the Norman conquest of England in 1066 but also the comet that terrified the Christians three years after the Turks captured Constantinople in 1453. Halley's comet, the brightest of all the comets, reappeared in 1986.

5. The shortest orbit of any comet is that of Encke's comet, first seen in 1786- it has an orbit of just 3.30

years. Encke was a pupil of the German mathematician and astronomer Gauss, who, in 1809, constructed a method for computing any type of orbit, improving on a rather limited method used for computing parabolic orbits, published twelve years earlier.

6. Short periods comets move mainly in direct orbits that lie close to the mean plane of the solar system and are often refereed to as belonging to the family of Jupiter, because they have aphelion distances close to the orbit of that planet. They rarely have conspicuous tails that can sometimes disintegrate, causing showers of meteoric particles. Comets that have a hyperbolic orbit are usually seen only once and do not reappear; others, like Halley's comet are elliptical in orbit and the date of their next appearance can be calculated.

7. Near the turn of the century it became possible to observe the constituents within a comet with the aid of the wide-field view of the objective prism. Our present knowledge of comets is derived from early pioneering observation, solely visual, until the beginning of the present century.

8. Comets are named after their discoverer with up to three names permitted; then the year of discovery is given, followed by a lower case letter giving the order of discovery (or recovery in the case of periodic comets). E.g. 1979 a. 1979 B. 1979
c. and so on. Permanent designations are eventually allotted in order of in order of perihelion passage; the first comets to pass in 1979 would be called 1979 I, the second 1979 II, and so on, using upper case

Roman numerals. Sometimes there may be a difference between the preliminary and the definitive designations as comets can be discovered long before or long after their perihelion passage. If there is a multiple observation of a comet at the same time, then an impersonal name, e.g. Northern comet, is given.

1. What does `they in' "They rarely ..." in paragraph 6 line 5 refer to?

_____

_____

_____

_____

2. What does "it" in paragraph 3 line 3 refer to?

_____

_____

_____

_____

3. When was the method used by Gauss for computing parabolic orbits published?

_____

_____

_____

_____

4. What two points were known about Halley's comet after Halley's death?

a. _____

_____

_____

_____

b. _____

_____

_____

_____

5. Approximately, when do you think Halley's comet will reappear again?

_____

_____

_____

_____

6. What is the difference between short-and long-period comets?

_____

_____

_____

_____

7. Write another word or phrase that could replace the word "impersonal" in the last line of the last paragraph.

_____

_____

_____

_____

8. Explain what you understand by the phrase "with a mon the either side for error .... " in paragraph 4, line

_____

_____

_____

_____

9. In one or two sentences, what is the purpose of the passage, do you think?

_____

_____

_____

_____

B. The following sentences summarize the contents of the 8 paragraphs in the passage. Match the sentences with the sentences with the paragraphs by writing the paragraph numbers next to the sentences. (8 marks)

1. The shortest orbit of a comet was determined after a method was properly improved.
   Paragraph _____

2. Astronomers follow an already established uniform pattern when naming and identifying the comets they discover. Paragraph _____

3. Gradually, different types of comets were discovered, their orbits were determined, and the times they had appeared earlier was known.
   Paragraph _____

4. Comets are special types of starts that appear suddenly in the space and their different parts are described and their contents are known.
   Paragraph _____

5. Short-period comets can sometimes exhibit odd behaviours. Paragraph _____

6. Observations of early astronauts contributed a lot to what we know today about comets.
   Paragraph _____

7. There is available a theory explaining comets' tails.
   Paragraph _____

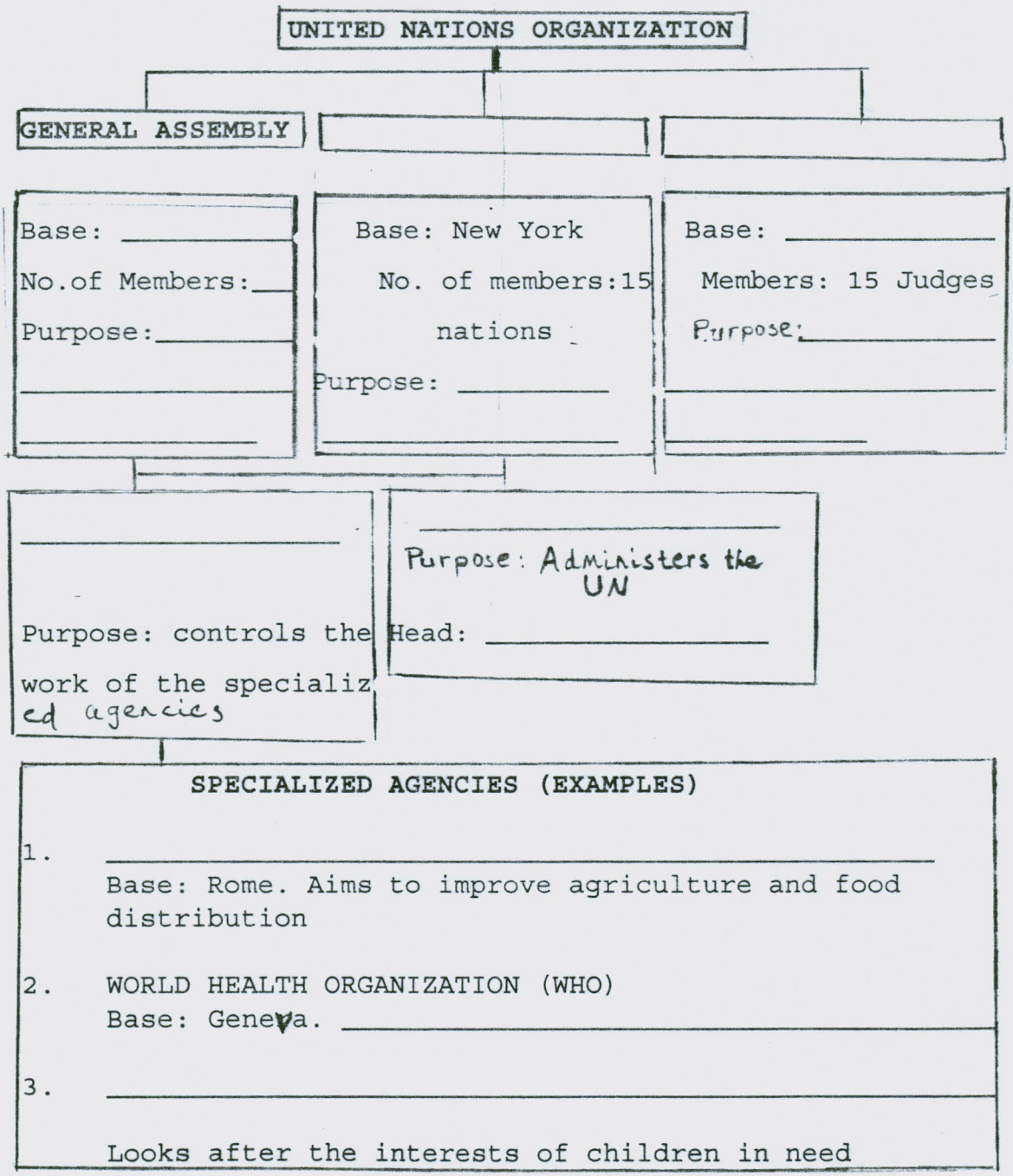
8. There are different areas in the study of comets and comets are now divided into two types.
   Paragraph _____

C. Below are meanings of some words used in the passage. Pick the words fun the passage and write them beside their meanings (as used in the passage) (15 marks)

1. Cloudy (para. 1.) _____
2. measuring (para. 1.) _____
3. goes ahead of (para.1.) _____
4. invented (para. 2.) _____
5. intricate (para 3.) _____
6. essential (para. 3) _____
7. forecast (para 4.) _____
8. described (para. 5) _____
9. narrow (Para.5.) _____
10. distinct, easily seen (para.6) _____
11. break up (para.6) _____
12. results from (Para.7) _____
13. allowed (para. 8) _____
14. given (para. 8) _____
15. Conclusive (para. 8) _____

C. INFORMATION TRANSFER

Below are an incomplete chart and an incomplete passage about the united nations organization. The information in the chart and in the passage can be used to complete each other. So read both carefully and fill in the missing information in both. (20 marks)



SPECIALIZED AGENCIES (EXAMPLES)

1. _____
   Base: Rome. Aims to improve agriculture and food distribution

2. WORLD HEALTH ORGANIZATION (WHO)
   Base: Geneva. _____

3. _____
   Looks after the interests of children in need

THE UNITED NATIONS ORGANIZATION

The united Nations Organization (UNO) is one very important world body involved in a number of ways in countries around the world. Its structure helps the organization to achieve its objectives. The structure of the organization shows how it is divided into major and sub bodies and what each of them serves the world. Here is a description:

Firstly, the organization is divided into three major bodies: the _____, the security council, and the international Court of Justice. Each of these has a separate head quarter located at different places in big cities. The General Assembly and the Security Council have their headquarters in New York, while the International Court of Justice has its head quarters in the Hague.

The General Assembly, which has over 130 member countries, has the purpose of discussing world problems and controlling UN finances. The Security Council. Which has _ _____ member countries, tries to keep world peace. Judgement and advice on international law is provided by the International Court of Justice, which has _____ judges.

The UN is administered by the Secretariat, whose head ins the Secretary General. The organization's Economic and Social Council,. which has 18 member nations, has the purpose of _____. The organization has a lot of these agencies that focus on particular aspects. For example, the Food and Agriculture Organization (FAO), which is based in Rome, aims to _____ _____ A second example is the

_____ which is based in Geneva. It aims to improve health and medical services. The interest of children in need is looked after by the United Nations International Children's Emergency Fund (UNICEF).

PART II. WRITING

A. Look at the following text about bread and complete the exercises set on it.

The discovery that the seeds (or `grains') of some plants can be eaten had an important effect on man's development. It made him realise that instead of spending all his time moving from place to place in search of animals to eat, he could actually stay in one place and grow some of his own food. It is no exaggeration to say that this discovery helped to turn man into an animal which settles and forms a permanent abode.

The grains of the wheat plant (in the form of a powder known as `flour') from the basic ingredient of one of the world's most common foods-bread. The other ingredients of bread are yeast, sugar, water, salt and fat.

Bread is usually made in five stages. The first is to make what is called `dough'. The yeast is mixed with sugar and water, and after about fifteen minutes is begins to `eat' the sugar. Flour, fat and salt are then put together and the yeast mixture is added. All these ingredients are then pressed (or `kneaded') with the hands for about ten minutes until they form a large ball of dough. After the dough has been made in this way, it is left to `rise'. As the yeast continues to eat the sugar it makes the dough increase in size, and this second stage of rising takes about two hours.

At the third stage the risen dough is needed again and pushed into the shape the bread is to be. The dough must then be allowed to rise again, this time for about one hour. It is then ready for the final stage of baking. Which takes about forty-five minutes in a hot oven.

In some countries the dough is not left to rise, the result being flat pieces of bread called ``unleavened (un risen) bread'.

1. There are five stages of making bread mentioned in the passage. Briefly write the stages below. (5 marks)

   How bread is made.

   Stage 1: _____

   Stage 2: _____

   Stage 3: _____

   Stage 4: _____

   Stage 5: _____

2. The passage describes how bread is made. Use this description to give a written instruction on how to make bread. The first two sentences have already been given (10 marks)

First mix the yeast with the sugar and water. After about fifteen minutes it will begin to `eat' the sugar. Then ___

_____

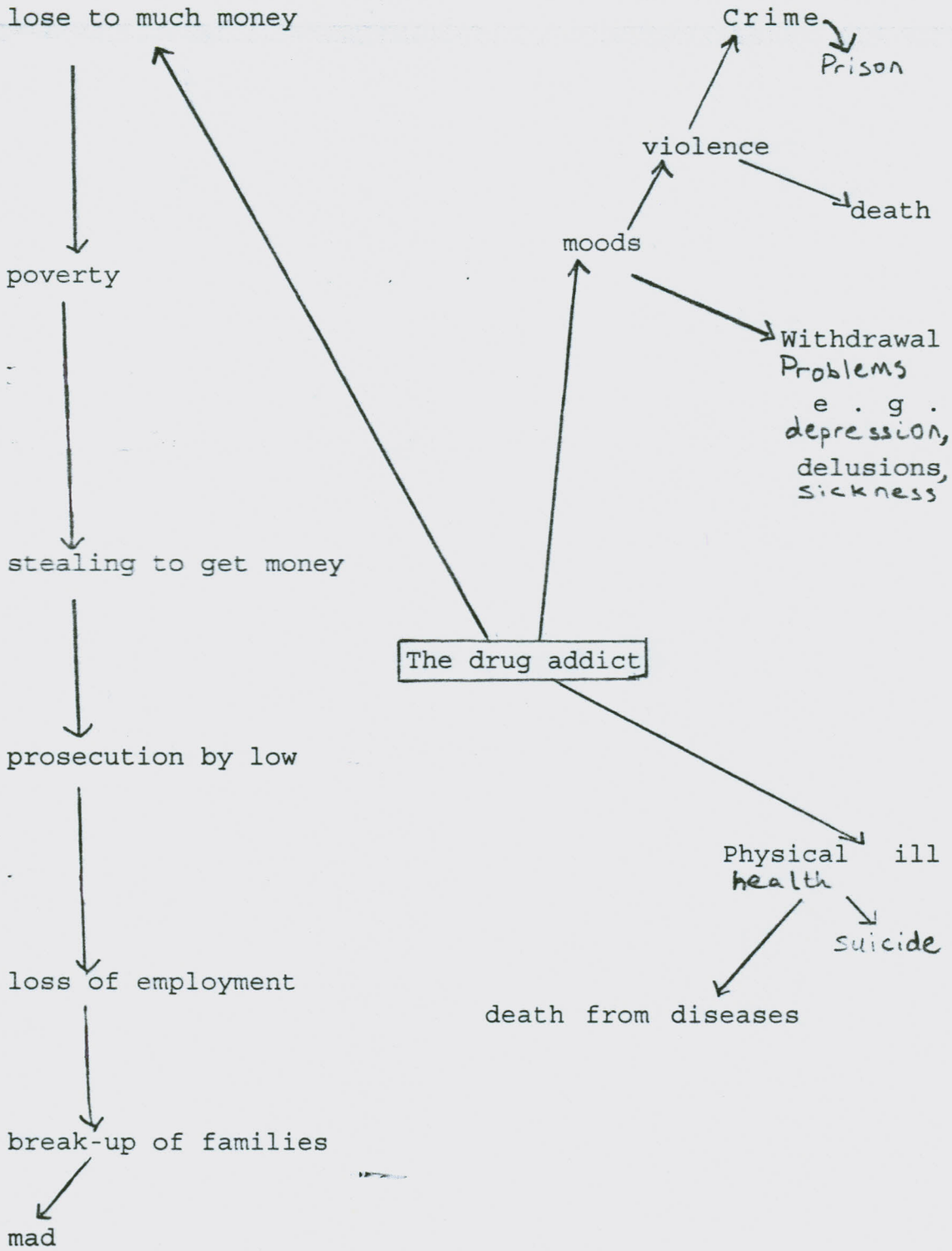
_____

_____

_____

3. Look at the following diagram carefully and write an essay of about three paragraphs about the problems of drug addicts. (N.B. You can also use any other logical connection than the one shown here).

The first sentence has already been given (10 marks)



There are a lot of problems that the drug addict might find himself or herself in. First of all, he or she needs too much money

PART III-GRAMMAR

A. In the following passage there are 15 verbs in brackets and 7 blank spaces. Give the appropriate tenses for the verbs in brackets and write the correct forms on the space provided below the passage. In the blank spaces put in appropriate prepositions (1/2 mark each)

The Rights of Women

Ever since 1776, Americans 1. (quote) a famous statement from their Declaration of Independence: "All men are created equal". _____more than 200 years, most Americans. 2. (Interpret) this phrase to include women, However, many women in the United States know better. Some of them argue that they 3. (not yet achieve) political or economic equality with men. Of course, if you asked the men, some might complain that women already have been given too much power.

Equal rights for women has been a live issue in America politics_____ at least 134 years, ever _____ the year 1848, when a few women 4. (hold) a Women's Rights convention _____ Seneca Falls, New York. However, not very much was accomplished there. American women 5. (have) to wait for 72 years before they were granted the right to vote, and many women 6. (feel) that full equality has not been attained yet.

People _____ the United States consider themselves modern and progressive but their country was many years behind some other nations in granting women the right to vote in national elections. For example, women in New Zealand 7. (vote) since 1899, and in Finland since 1906. Women in Norway had already been voting for seven years when American women were finally 8 (grant) the same right in 1920.

Ever since 1972, Americans have been very conscious of women's determination _____ achieve equality with men. It was in that year that congress 9. (propose) the ERA- the Equal Rights Amendment.

For five years each of the states debated the issue, and by 1977, thirty five states 10.(pass) the amendment. Then the drive for passage slowed down. Times 11. (gradually change) since 1972, and by the constitutional deadline of march 1979 no other state had passed the ERA since the term of President Ford. In fact, four states that 12. (already approve) the amendment later voted _____ change their vote! Although the time for passage 13. (be extend) three years, the ERA proposal was still three votes short to

the necessary thirty eight of the fifty states by the 1982 deadline; consequently the amendment 154. (fail) to become part of the constitution.

Since this defeat, many women have become discouraged. However, other women have worked for so many years for women's right that they cannot give up so easily. They have already said that they will reintroduce the ERA proposal _ _____ the next session of congress. The country 15 (not yet see) the end of the battle.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____
13. _____
14. _____
15. _____

APPENDIX F - *MARKING GUIDELINE*

Part I - *Reading*

Passage open ended questions (1 mark each)

1. 'They' refers to short period comets.
2. 'It' refers to the solar system
3. The method was published in 1797
4. A. That the comet had made 29 appearances since 239 B.C.
   B. That is was probably not only the comet depicted in the Bateaux tapestry but also the comet that terrified the Christians three years after the Turks captured Constantinople in 1453.
5. It will probably reappear in 2062.
6. The difference between the two is that short period comets are comets with orbital period of less than 200 years old while long period comets are comets with orbital period of over 200 years old.
7. That does not refer to any person
8. A month was left on both sides of the date calculated so that if the passage actually was on a day within one month before or after the date calculated, it would be acceptable.
9. Introducing and giving a detailed information about comets.

B- Paragraph Summaries (1 mark each)

1. Paragraph - 5
2. " 8
3. " 4
4. " 1
5. " 6
6. " 7
7. " 2
8. " 3

C. Passage Vocabulary (1 mark each)

1. Nebulous
2. Ranging
3. Precedes
4. Devised
5. Complex
6. Main
7. Predict
8. Depicted
9. Limited
10. Conspicuous
11. Disintegrate
12. Derived from
13. Permitted
14. Allotted
15. Definitive

D. Information Transfer

1. The security Council
2. The International court of Justice
3. New York
4. Over 130 member nations
5. Discussing world problems and controlling UN finances
6. Trying to keep world peace
7. The Hague
8. Giving judgement and advice on international law
9. Economic & social council
10. The Secretariat
11. The secretary General
12. The Food & Agriculture Organization (FAO)
13. Improves health & medical services
14. United Nations International Children Emergency Fund. (UNICEF)
15. The General Assembly
16. 15
17. 15
18. Controlling the work of the special agencies.
19. Improve agriculture& food distribution
20. World Health Organization (WHO)

PART II - Writing

A. Writing stages of making bread (1 mark each)

1. Making the dough
2. Pressing or 'kneading' the dough and leaving it to rise
3. Putting the dough into the shape needed
4. Allowing it to rise
5. Baking

Giving instructions. Check if the student has used imperative sentences and that he or she has made no awkward grammatical errors that distort the idea of his or her writing. To some extent consider also other points of accuracy like spelling, punctuation, etc. (10 marks)

B. The drug addict -check if the student has properly described the diagram (or his or her own preference) and that he or she has made no errors that distort the meaning of his writing. To some extent consider also accuracy. (10 Marks)

Part III - Grammar

Prepositions

1. For
2. for
3. Since
4. at
5. in
6. to
7. at

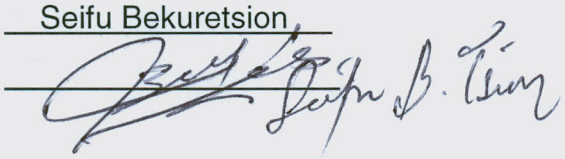
Tenses

1. have quoted
2. have interpreted
3. have not yet achieved
4. Held
5. had
6. feel
7. have been voting
8. granted
9. Proposed
10. had passed
11. had gradually changed
12. had already approved
13. was
14. failed
15. has not yet seen

DECLARATION

I, the undersigned, declare that this thesis is my work and that all sources of material used for this thesis have been duly acknowledged.

Name Seifu Bekuretsion

Signature 

Place **Institute of Language Studies,**
Addis Ababa University

Date of Submission: 23 May , 1997