Addis Ababa University

College of Humanities, Language Studies and Journalism and Communication Department of Foreign Languages and Literature

Item Analysis of English Language University Entrance Examination in Ethiopia: 2005 Exam in Focus

A Master's Thesis Submitted to the Department of Foreign Languages and Literature.

By:

Hadya Hassen

May 2014

ACKNOWLEDGEMENTS

First of all I would like to express my great gratitude to my families for their financial and moral support. I am grateful to my supervisor, Dr.Geremew Lemmu, for his continuous feedback for the success of my work. I would like to thank staff of Ethiopian National Educational Assessment and Examination Agency for their willingness of giving the data. Specially, I would like to extend my heartfelt thanks to Ato Yoseph for his assistance in doing the statistical data.

TABLE OF CONTENTS

Content **page**

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

CHAPTER FOUR

DATA ANALYZES AND DISCUSSION

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

LIST OF TABLES

List of Abbreviations and Acronyms

IF: Item Facility

DI: Discrimination Index

EUEE: Ethiopian University Entrance Examination

NEAEA: National Educational Assessment and Evaluation

MCQ: Multiple Choice questions

ELT: English language teaching

ABSTRACT

The main purpose of this study was to examine the effectiveness of the items of 2005 E.C university entrance examination of English language based on the parameters of facility value and discrimination index. The study has used descriptive research design concerned with the quantitative methods which reveals the statistical result of the exam. The data were collected from the data base of National Educational Assessment and Examination (NEAEA). The study covered all of the 120 items of the Ethiopian University Entrance Examination (EUEE). One fourth of the examinees' i.e. 40,400 examinees' result was taken for analysis. Secondary data was also used in the analysis process. The collected data were analyzed quantitatively. Within the two main parameters of item analysis, namely, item facility and discrimination index of the study, the items were categorized in to four ranges based on their effectiveness. Findings of the study revealed that 33 items were very difficult, and 27 items were not discriminating at all. The findings shows that high number of the EUEE examination items was not functioning as expected, which means there are many items that should be discarded or improved. This calls for implementation of post test item analysis regularly for improved exam preparation for the next exam periods.

Key words: Achievement test, multiple choice test items, Item analysis, Item facility, Discrimination index

Chapter One: Introduction

1.1. National Examination in Ethiopia

One of the most common uses of language tests is making a decision regarding selection in conjunction with measures of other abilities, such as grades from previous instruction, academic aptitude and achievement. In Ethiopia, university entrance examinations include testing English as a foreign language, in addition to tests of academic subjects like mathematics, science and history. In many programs, such as in primary schools, entrance is nearly automatic with age while other programs require a selection or entrance test for a certain field of study. The purpose of this entrance test is to determine whether or not students are ready for instruction and/or training.

In the history of Ethiopia, starting from the day English is being given as a subject, national examinations are started to be given in different grade levels and for different purposes. For example, in grade, 6 and 8[1] for the purpose of passing to the next grade, and at grade10 and 12 for the purpose of joining college or university, and for the selection of the next level study (social science or natural science) at preparatory levels.

High stake tests in Ethiopia are all designed in multiple choice questions and students take federally organized external examinations at the end of grade 10 and 12. These exams are conducted by national educational assessment and examination agency (NEAEA). The tests are all multiple choices and marked electronically. Currently, each examinee takes a test in a maximum of 10 subjects. English is common for both social science and natural science students. To this end, the research is concerned with item analysis of university entrance examination of the year 2005E.C through data collected from NEAEA.

[1] This had been used in the old curriculum, before the curriculum of the current government.

1.2. Statement of the Problem

The feasibility and validity as well as the effectiveness of a given test strongly depend on high quality questions. One of the tools to achieve this purpose is to conduct post exam item analysis continuously and use analysis results for improvement. Once tests are administered, items should be subjected to statistical analysis. In the process of teaching and learning it is not advised not to be aware of the effectiveness and quality of ELT material and administered examinations. It makes the teaching learning difficult if the exams are administered in the same way for a series of years without knowing an exam is valid and reliable enough. So, in order to eliminate such kinds of problems timely, post exam analysis is very essential.

The researcher observed that University Entrance Examination and similar national examinations are creating fear and anxiety to those who take them. For that matter, nowadays grade 8 (regional level),grade 10 and grade 12 students want and insist their English teachers manage the course in relation to national examinations and to focus on grammatical and vocabulary aspect of the course. Hence the form, content and style of the exams are similar in every year, teachers in these grade levels collect examinations of different years and discuss the answer of the exams during the time of the lesson.

Indeed, there are some researchers who conducted research on related aspects of this study. Among the researchers, Kassaw (2006), in his thesis, examined the content validity of university entrance examinations. According to his study, the multiple choice items do not function properly for all skills and sub skills. In addition, unfamiliar task design and test format leads candidates to anxiety. Consequently, even high achiever students may not perform as they are expected. As a result, detailed researches are needed to be conducted on the analysis of each item.

Other theses and dissertations are also there. For instance, Simachew (2012), studied the wash back effect of university entrance examination. And Kassaw(2006) and Kifle (1995) studied the content validity of the national examination.

However, as there is scant practice by NEAEA and an academic as well as applied research conducted on post test exam analysis, university entrance examinations are always similar in form and content though the curriculum and the teaching learning materials are changed or improved in the Ethiopian situation. Therefore, the researcher intended to assess if those bad consequences of the exam are caused by difficulty of the exam and as a result of lack of quality items.

The researcher of this study assumes that the feedbacks given by researched and published works on test item analysis are not enough. So, the researcher wanted to deal with these specific issues to fill the gap by giving attention on analyzing and improving test items.

1.3. Objective of the Research

1.3.1. General objective

The main objective of this study is to identify the effectiveness items of English language university entrance examination of the year 2005 E.C using difficulty level and discrimination index.

1.3.2. Specific objectives

The specific objectives of this study are:

- Assessing the discrimination index of test items
- Investigating the level of difficulty of test items.

1.4. Research questions

- Do the test items fall in the standard range of discrimination level?
- Do the test items fall in the standard range of facility value?

1.5. Significance of the study

Because much time and energy is devoted in preparing language tests the result of this study is important first for the institution to keep record of the evaluation result of the exam and allow them to think more for the future preparation, to conceive to what extent the professionals perform their job related duties with competence in preparing the exam. It also reminds stakeholders to assess the practical quality of every year examination.

Second, teachers who have prepared the current year candidates and who will be preparing other students to take the test in the future will be aware of how much the items are effective in accordance with their classroom tests. So that they can able to prepare effective test items during their lesson and to develop the habit of post test item analysis.

Next, it is relevant for other parties who may need the information about the effectiveness of the exam. For example it helps Ministry of Education which is responsible for designing curriculum and giving the mandate of exam preparation.

Finally, as local researchers do not specifically focus on item analysis; it can be the basis for other researchers to deal with the issue in detail.

1.6. Scope of the Study

Due to limited time and resources this research doesn't deal with other test qualities like reliability, validity and other practicality related concerns. Thus, item analysis is the only focus of this research. Moreover, it focuses only on one year's 2005 E.C exam items.

1.7. Limitations of the study

Even if the researcher tried to use all possible efforts, there are still some limitations in the study. Firstly, because of the researcher had limited time and budget, the study didn't include further observational evaluation of the exam items.

Secondly, as a result of the bureaucracy of the NEAEA the researcher couldn't find the data at the right time and encountered shortage of time for data analysis and interpretation.

1.8. Operational Definitions of Terms

*Achievement test***:** A summative tests given at the end of a particular academic year to assess the effectiveness of a particular course.

Multiple choice test items: the stem of a multiple choice question.

Item analysis: The process of collecting, summarizing and using information from students' responses to assess the quality of test items.

Item facility: The proportion of students selecting the correct answer

Item discrimination: Statistical index of item quality and reflects the degree to which the item was able to differentiate between the examinees who score high and low in the exam.

Chapter Two: Review of Related Literature

2.1. Test

Test is one of the instruments among variety of measurement and evaluation methods. Harold S. Madsen (1983) asserted that testing is an important part of every teaching and learning experience. In education it is used to assess and evaluate to what extent students learned, the teaching learning material is effective and the teacher is effective in achieving the goal of the teaching learning process. “A test in simple terms, is a method of measuring a person’s ability, knowledge or performance in a given domain” (H.D.Brown 2003) .As stated above there is a brief domain given in a teaching learning process and testing leads to measure how much a student performs well towards a given domain. A well-constructed test is an instrument that provides an accurate measure of the test taker's ability within a particular domain. The definition sounds fairly simple, but in fact, constructing a good test is a complex task involving both science and art Brown (2003).According to Brown, most language tests measure one's ability to perform language; to speak, write, read, or listen to a subset of language.

Carroll (1968) as cited in Bachman (1995) defined test as” A psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual.”(Carroll 1968: 46)

Test is a fundamental part of the teaching learning process used not only as a basis for ranking students at the end of the teaching learning process but to guide teaching and to aid the development of curriculum as well to assess needs learning difficulties ; level of mastery and difference among students. It is logical to assess the ability of students and the effectiveness of all the teaching learning process, material and the curriculum in education.

The quality of assessment methods and processes is as important as the quality of the teaching and learning process in any form of educational activity. Construction and selection of appropriate assessment items is an essential task for test makers. The critical issue throughout that process is how best to formulate items to optimally assess comprehension.

The fundamental use of testing in an educational program is to provide information for making decisions, that is, for evaluation. An educational program, in the broadest sense, is any situation in which one or more persons are engaged in teaching and learning. (Bachman, 1995)

It is not as such easy to design a question to assess student's language ability like vocabulary definition, reciting grammatical rules, reading comprehension, writing skill and answering to the questions after listening to a particular speech. According to Madsen (1983) if language test is designed properly it has different advantages such as it eliminates foreign language anxiety and it allows learners to think of they can accomplish any task in the target language.

There are different categories of language tests according to their purpose. Bailey (1998) noted eight kinds of language tests have different kinds of purposes of language tests, language aptitude tests, language dominance test, proficiency tests, admission tests, placement tests, diagnostic tests, progress tests and achievement tests. However, James Dean Brown (1996)categorized language tests in to two main categories; norm referenced and criterion referenced and grouped all the above kinds of tests in his category .Achievement test is grouped in criterion referenced language tests.

2.1.1 Achievement Test

An achievement test is related directly to classroom lessons, units, or even a total curriculum. Achievement tests are (or should be) limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objective in question.

Based on Brown (2003) achievement tests sometimes function as summative tests, because they can be given at the end of a particular course. They can also serve as diagnostic purpose when they are used to identify which kind of students needs are to be well addressed for the future and when they help to reform the curriculum. If achievement test are administered properly, they gives information about the effectiveness of the teacher and the attainability of the course and the producttivity of students.so it can serve as input to the feedbak of the exam.

"Achievement tests range from five or ten minute quizzes to three-four hour final examination, with an almost infinite variety of item types and formats" (H.D Brown 2003). The length of the test types of items and test formats depends up on the objective of the exam.

One of the achievement tests given in Ethiopia is university entrance examination which contains 120 multiple choice test items .All the language skills and sub skills are considered to be included. The time allotted to accomplish the exam is 1 hour and 30 minutes. This means one minute is given for each item. Harrison(1983)cited in Kassaw Baye (2006) explained that an achievement test looks back over a long period of time and is intended to show the standard which the students have now reached in relation to other students at the same stage. The test must also reflect the content of the whole content. Based on this idea, if the exam is intended to test the achievement of students, it must be evaluated to find out if the test is well functioning in determining every stage ability students performance and if the test takers result is based on their real ability and performance or based on cheating, guessing or else. Therefore the feedback about each item and the whole test is needed to be generalized

2.2. Characteristics of Tests

The procedure of test preparation seems tedious. However regardless of the complexity of the tasks in determining the reliability validity and practicality of the test, these parts are indispensable part of the test construction. In order to have an acceptable and defendable test, up on which reasonably sound decisions can be made, test developers should pass through different steps of test preparation such as planning, preparing, reviewing, pretesting. In addition the effectiveness of both item and test should be evaluated through its validity, reliability and practicality indexes.

Brown(2003) also stated some issues as he called five cardinal criteria “For testing a test”; which are test practicality, reliability, validity, authenticity, and wash back. Every individual criterion also includes different things to be assessed specifically.

2.2.1. Validity

Validity is certainly the most important single characteristic of a test. If not valid, even a reliable test does not worth much. The reason is that a reliable test may not be valid; however, a valid test is to some extent reliable as well. Furthermore, where reliability is an independent statistical concept and has nothing to do with the content of the test, validity is directly related to the content and form of the test. In fact, validity is defined as "the extent to which a test measures what it is supposed to measure". This means that if a test is designed to measure examinees’ language ability, it should measure their language ability and nothing else. Otherwise, it will not be a valid test for the purposes intended. Messick (1989) describes validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the Adequacy and appropriateness of inferences and actions based on test scores.”

2.2.2. Reliability

Reliability is one of the most important characteristics of all tests in general, and language tests in particular. In fact, an unreliable test is worth nothing In order to understand the concept of reliability; it is the consistency and dependability of the test scores when using different ways of examining reliability. A reliable test is consistent and dependable. If you give the same test to the same students or matched students on two different occasions, the test should yield similar results.(Brown 2003)

2.2.3. Practicality

The last characteristic of a good test is practicality. It refers to facilities available to test developers regarding both administration and scoring procedures of a test. As far as administration is concerned, test developers should be attentive to the possibilities of giving a test under reasonably acceptable conditions. For example, suppose a team of experts decide on giving a listening comprehension test to large groups of examinees. In this case, test developers should make sure that facilities such as audio equipments and/or suitable acoustic rooms are available. Otherwise, no matter how reliable and valid the test may be, it will not be practical.

2.3. Multiple Choice Test Item

Multiple-choice tests are of considerably widespread use as a means of objective measurement. The main reason behind such popularity is the many dominant advantages associated with multiple-choice tests. They can be used for diagnostic as well as formative purposes and can assess a broad range of knowledge. In addition, they are scored easily, quickly, and objectively either by human-beings or by scoring machines. These and many similar advantages make multiple-choice tests suitable for a wide range of purposes ranging from classroom achievement testing to large-scale standardized tests. Thus,

improving the quality of multiple-choice test items appears to be of a lot of importance, (Baghaei & Amrahi 2011)

Multiple choice items, which may appear to be the simplest kind of item to construct, are extremely difficult to design correctly. Hughes (2003) stated that designing a multiple choice test item needs an intellectual science and art. Because of various reasons for MCT not to be functional, one who prepares a multiple choice test item is expected to be wise and careful.

Properly constructed multiple choice questions assess higher-order cognitive processing of Bloom's taxonomy such as interpretation, synthesis and application of knowledge, instead of just testing recall of isolated facts. All this is possible if the examiner knows the correct method of formulating a question, commonly referred to as an item, consisting of a stem and several options.

Selection of appropriate language items is not enough by itself to ensure a good test. Each question needs to function properly; otherwise, it can weaken the exam. After test, an important and valuable endeavor for test makers is evaluating items to determine which ones were good and which ones were poor. Fortunately, there are some rather simple statistical ways of checking individual items.

The analysis tells us, basically, three things: how difficult each item is (item facility), whether or not each item discriminates between high and low students (item discrimination index), and which distracters are working as they should (distracter efficiency). These procedures are called item analysis. It is most often used with multiple choice questions. An analysis like this is used with any important exam for example, review tests and tests given at the end of a school term or course.

Since the item is the basic unit, or building block in testing, one way to improve a test is to examine the individual items and revise the test so that only those items that are performing well remain in the revised version of the

test. Teachers often look at the total scores of their students on a test, but careful examination of the individual items that contributed to the total scores can also prove which item should be eliminated. This process of carefully inspecting individual test items is called item analysis. More formally, item analysis is the systematic evaluation of the effectiveness of the individual items on a test.

2.4. Item Analysis

Item analysis is a valuable, yet relatively simple, procedure performed after an examination that provides information regarding the reliability and validity of a test item.

It is the process of collecting, summarizing and using information from students' responses to assess the quality of test items. It tells how difficult or easy the questions were,(the difficulty index) and whether the questions were able to discriminate between students who performed well on the test, from those who did not(the discrimination index). Another important technique is analysis of distractors that provides information regarding the individual distractors and the key of a test item.

"Item analysis indicates which item may be too easy or too difficult and which may fail for other reasons. This makes it transparent to discriminate clearly between the better and the poorer examinees" (Ebel 1972). Using these tools, the examiner is able to modify or remove specific items from subsequent exams. It further helps to detect specific technical flaws and thus provide further information for improving test items. Similarly, it helps in selecting the best items for the final test, reject poor items and modify some of the items.

"Item analysis is usually done for tow purposes; one for selecting the "best" items that will remain on a revised and improved version of the test. The other is simply to investigate how well the items on a test are working with a

particular group of students." (Mc Namara 2000). Brown (1971) mentioned that item analysis has two purposes: First it enables to identify defective items, to improve the test and evaluation procedures. Second, through indicating which items or material students have and have not mastered, onecan plan, revise, and improve the ins tructios. When standardized achievement tests once given, it can be done for the purpose determining which item was effective and which one was not. Most probably based on the uses of post test analysis it helps to allocate appropriate item types and formats for the future parallel exam preparation.

In addition to the general matters of test characteristics validity and reliability madsen,(1983) referred to be concerned on the effect of examinations by taking time to evaluate individual items by using the ways we can improve our tests. He stated facility value,descrimination index and distractor efficiency as parameters to evaluate. He also argues anxiety will be generated because of too difficult items or unfamiliar task types so that teachers or test designers need to be careful in designig test items inoreder to block this gap.

According to J, Alderson, (1995) post test analysis should be made on different aspects. Such as:

- Descriptive analysis of the whole test and each of its components: histogram, mean, mode, median, range and standard deviation.
- Item analysis for each objective item: facility value and discrimination index.
- Correlation between components and correlation between each component and the total minus the component.
- Reliability of each objective section.
- Reliability of marking for each subjective section.

The test should be ideal within its difficulty and expected to measure students' ability as expected. As items quality plays a great role for test quality the way we can check for test effectiveness is the individual item effectiveness. The two main psychometric parameters are difficulty value (item facility) and discrimination index, which are the main focus of this research. The appropriate selection and arrangement of suitable multiple choice items on a test can be accomplished by measuring items against three indices: item facility (or item difficulty), item discrimination (sometimes called item differentiation), and distractor analysis. (Brown, 2003).

2.4.1. Item Facility Analysis

Item difficulty is defined as the proportion of students selecting the correct answer. This is a common practice as tests are often rejected as reliable measures of examinee performance due to the misfit of item difficulty to the ability of the examinees, (Bachman, 1990). The most effective questions in terms of distinguishing between high and low scoring students will be answered correctly by about half of the students. In practical terms, questions in most classroom tests will have a range of difficulties from low or easy (.90) to high or very difficult (.40). Questions having difficulty estimates outside of these ranges may not contribute much to the effective evaluation of student performance. Bailey (1998) puts criteria of the range of facility value of an item.

1. Above 0.85 =very easy
2. Below 0.3 = very difficult
3. 0.3-0.39 –reasonably acceptable and
4. 0.4-0.85 –ideal items

Very easy questions may not sufficiently challenge the most able student. However, having a few relatively easy questions in a test may be important to verify the mastery of some course objectives. Very difficult questions, if they form most of a test, may produce frustration among students. Some very

difficult questions are needed to challenge the best students. When tests are too easy or too difficult, the scoring distribution will tend to unnaturally concentrate at one end of the continuum. As a result, it is difficult to distinguish candidates' ability at the concentrated end. This, inevitably, also results in loss of person separability or reliability (Henning, 1987).

Item facility (IF) (also called item difficulty or item easiness) is a statistical index used to examine the percentage of students who correctly answer a given item. The greater the difficulty of an item, the lower its index is (Wood, 1960). To calculate the difficulty of an item, the number of persons who answered it correctly is divided by the total number of the persons who answered it. Usually this proportion is indicated by the letter p, which indicates the difficulty of the item (Crocker and Algina, 1986).

Brown (2003) IF simply reflects the percentage of students answering the item correctly. The formula looks like this

$$IF = \frac{\#\text{of Ss answering the item correctly}}{\text{Total\# of Ss responding to that item}}$$

Very easy items and very difficult items don't do a good job of discriminating between students who know the content and those who do not. (The section on discrimination discusses this further.) However, there may have very good reason for putting either type of question on the exam. For example, some teachers deliberately start their exam with an easy question or two to settle down anxious test takers or to help students feel some early success with the exam.

2.4.2. Item Discrimination Analysis

The discrimination index is a statistical index of item quality and reflects the degree to which the item was able to differentiate between examinees who scored well and those who scored poorly in an assessment, and indicates the extent to which the item correlates with overall examinee performance on the examination .e.g. high (positive) discrimination index reveals that the item was correctly answered by those examinees who performed well on the overall assessment.

The discrimination index is a useful measure of item quality whenever the purpose of a test is to produce a spread of scores, reflecting differences in student achievement, so that distinctions may be made among the performances of respondents, (Hotiu 2006). Item discrimination (ID) indicates the degree to which an item separates the students who performed well from those who performed poorly. These two groups are sometimes referred to as the high and low scorers or upper and lower-proficiency students. The reason for identifying these two groups is that ID allows teachers to contrast the performance of the upper group students on the test with that of the lower-group students. The process begins by determining which students had scores in the top group on the whole test and which had scores in the bottom group. The formula for calculating ID is

$$ID = \frac{\text{high group \# correct-low group \# correct}}{1/2 \text{X total of your two comparison groups}}$$

The result of this calculation can help teachers to select that subset of CRT items that are most closely related to the instruction and learning in a course and/or that subset most closely related to the distinction between students who passed or failed the test. With sound CRTs in place, teachers can indeed judge the performance of their students. However, equally important, teachers can also examine the fit between what they think they are teaching and what the students are actually absorbing (Brown, H.D 1996)

The higher the discrimination index, the better the item can determine the difference between those with high test scores and those with low ones. One of the ways to obtain discrimination indices is by sample separation which essentially involves a computation procedure that separates the highest scoring group and the lowest scoring group from the entire sample on the basis of the total score in a test (Gronlund, 1976 and Henning, 1987). This will produce discrimination indices that range from zero to one. Ebel's (1972) set criterion for item revision and test evaluation the list of criteria is based on a range of DI that categorically defines the items in a test:

1. If DI ≥ 0.40, the item is functioning quite satisfactory.
2. If 0.30≤DI≤0.39, little or no revision is required.
3. If 0.20 ≤DI≤0.29, the item is marginal and needs revision.
4. If DI ≤ 0.19, the items should be eliminated or completely revised.

The possible range of the discrimination index is -1.0 to 1.0; however, if an item has discrimination below 0.0, it suggests a problem. When an item is discriminating negatively, overall the most knowledgeable examinees are getting the item wrong and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item is measuring something other than what the rest of the test is measuring. More often, it is a sign that the item has been mis-keyed.

2.4.3. Distractor Efficiency

Distracter evaluation is another useful step in reviewing the effectiveness of a test item. All of the incorrect options, or distractors, should actually be distracting. Preferably, each distracter should be selected by a greater proportion of the lower scorers than of the top group. In order for a distractor to be acceptable it should attract at least one candidate. If no one selects a distractor, it is important to revise the option and attempt to make the distractor a more plausible choice.

It is important to keep in mind that the statistical functioning of an item should not be the sole basis for deleting or retaining an item. The most important quality of a classroom test is its validity, the extent to which items measure relevant tasks. Items that perform poorly statistically might be retained (and perhaps revised) if they correspond to specific instructional objectives in the course. Items that perform well statistically but are not related to specific instructional objectives should be reviewed carefully before being reused.

2.5. Empirical Studies

Angelica Hotiu (2006) investigated the relationship between difficulty value and discrimination index of one physical science course. The study is on the analysis of a multiple choice examination, and it investigates that there is a proportional relationship between difficulty value and discrimination index, in one item which means when the difficulty value of an item is acceptable the power to discriminate between the low achiever and high achiever students also become good, which is not common and unexpected. Usually if the item is very low or too difficult the discrimination power will have a vise versa effect. However, there is a maximum degree of difficulty beyond which the discrimination index starts to fall off. At that point, the test items become too difficult for both the high scorers and the low scorers to answer, so that they no longer discriminate effectively. There was a critical level of difficulty beyond

which the discrimination decreased. Therefore the difficulty level of the item has a great impact on the discrimination index of an item.

Madziah, Zubairi and Noor Lide (2006) demonstrated item quality by classifying in to two groups; classical theory and Rasch analysis, and they find out items which are too easy and in other group items which are too difficult, and in the third group with items in average difficulty level. According to classical test item analysis, from 35 reading comprehension questions, they found 11 items too easy, 11 items too difficult and 13 items average difficult items. They also generalized one third of reading questions were weak in discriminating between high and low achiever students.in addition to that the research also provided 24 misfit reading items based on Rasch analysis.

Mozaffer and Farhan (2012) in their analysis of one best answer multiple choice test item they analyzed the item using the three item analysis parameters ;difficulty value, discrimination index and distracter efficiency. They investigated that there were 100% efficient items. They demonstrated that the number of non functioning distracters affect the discrimination of the item. They also revealed reducing the number of distracters from four to three decreases the difficulty level of the item but increases the DI and reliability. The research suggested that writing items with four distracter is difficult because test makers are not wise in making the forth distracter plausible enough. The items which have almost all plausible distracters have good difficulty level and discrimination index. Therefore the effectiveness of distracters affects the effectiveness of the item.

Boopathiraj.C and Dr.K (2013) conducted the researcher made test item analysis investigated 21%of the items un acceptable which must be rejected, 58%acceptable and 11%acceptable with necessary revision. The research revealed that the length of the test, range of difficulty indices and purposes for which the test has been designed have a role on the acceptability of the size of

the item. The research recommends that student teachers who are going to be teachers must be careful in selecting test items.

Yu- mein shih(2010) Analyzed the items of one English achievement test items in the parameters of item facility and discrimination index using language skills(listening, reading ,writing, vocabulary and grammar) as a different variables to categorize the result. The finding of the research reveals that vocabulary tests were too difficult and powerful in discriminating between high and low able students. From all categories reading test items are the easiest. Based on the result most of the listening, reading, writing and grammar test items were not discriminatory enough. Suggestions were given in the research that the test needs to be revised in order to assess the ability and performance of students effectively.

Omirin M.S (2007) compared the difficulty and discrimination indices of three multiple choice tests using the confidence scoring procedure (CSP). The study was also set to determine whether or not the difficulty and discrimination indices would be improved, if the tests were scored by the confidence scoring procedure. The three multiple-choice tests were mixed and the examinees did not know that they were answering different forms of the same test. The test scripts were scored using the confidence scoring method. The result revealed that the contribution of blind guessing to examinees was not directly related to the discrimination and difficulty indices of the three multiple choice tests used. The confidence scoring procedure improved significantly the difficulty index of multiple-choice tests but did not significantly improve the discrimination index of three-index test used. The result showed that confidence scoring procedure rewards partial knowledge of examinees on the multiple-choice tests. It was recommended that confidence-scoring procedure should be encouraged for scoring multiple choice tests hence it discourages guessing.

Fahmi Ishaq and El--Uri Naser Malas(2013) conducted an item analysis of one medical examination and investigated that there is a good difficulty level and

insured of content coverage. Both difficulty and discrimination indices in the exam were achieved well. Only one item was identified with low discrimination index. No question was identified as unacceptable. Most of the items were founded with very good quality.

Though many researchers carried out item analysis in different countries and in different contexts, it is not well experienced in our country. The intension in this research is to conduct the analysis of each item individually by using the two common item analysis parameters (Item Facility and Discrimination index).

In relation to local researches the researcher tried to see the study by Madsen entitled "Item Analysis of College Entrance Examination in Ethiopia (1967)." But every situation was very different from the current practice, and the researcher not found things that could be basis for this study.

In addition to that even if the researcher had informed that there is a study entitled "Item Analysis of University Entrance Examination in Ethiopia" conducted by Desalegn Chanckisa, et al (2001),she couldn't find the research available in the referred place: in Addis Ababa University IER library. As a result all the above reviewed researches are from other different countries.

Chapter Three: Research Design and Methodology

3.1. Research Design

This study is descriptive in its design as it focuses on the analysis of test items. Hence it deals with statistical analysis of the items and reveals numerical results the study is a quantitative.

3.2. Research Sample and Source of Data

A multiple choice test with 120 items was used for data collection. The exam was prepared for the purpose of grade 12 university entrance of the 2005 Ethiopian academic year. This university entrance examination was set by subject area experts. It was a paper-pencil test administered by Ethiopian Educational Assessment and Examination agency. Moreover, this exam consists of seven sections and these are word order, paragraph coherence, vocabulary, grammar, reading, communicative activities and writing. All items of all the afromentioned sections were analyzed accordingly.

By the 2005 Ethiopian academic year some 170,000 examinees were registered in the Ethiopian Educational Assessment and Examination agency to take the entrance examination. The results of 40,400 students (almost one fourth of the total examinees) who have taken this exam were selected using random sampling technique. The way it was randomized is taking code from the four booklet code of the exam. Based on this assumption, results of the examinees of code 14 were selected for the analysis purposes.

3.3. Data Collection and organization

The researcher collected the data from the record result of the Ethiopian Educational Assessment and Examination Agency. The collected data had gone through SPSS and excel spread sheet and then item facility was done using all individual examinees result. To do this, the proportion of students who answered the questions correctly were taken as a formula (p=C/TN).Here, the

level of difficulty was classified into four groups based on the criteria of Bailey (1998).

1. above 0.85 =very easy
2. below 0.3 =very difficult
3. 0.3-0.39 =reasonably acceptable and
4. 0.4-0.85 =ideal items

Although, there are various similar ways of calculating the discrimination power, the researcher has used the simplified technique; taking the upper 27% (10,908) and lower 27%(10,908)of the examinees results. In this work the researcher used the following formula to calculate the discrimination index of the items:

$$\frac{\text{Upper group of students-lower group of students}}{\text{Total number of students}} = \frac{(Us-Ls)}{Ts}$$

The items were categorized into four groups by the level of their discriminating power based on the range given by Ebels (1992).

1. If DI ≥ 0.40, the item is functioning quite satisfactorily.
2. If 0.30≤DI≤0.39, little or no revision is required.
3. If 0.20 ≤DI≤0.29, the item is marginal and needs revision.
4. If DI ≤ 0.19, the items should be eliminated or completely revised.

Chapter Four: Data Analysis and Interpretation

4.1 Statistical Analysis and Results

In this part 120 multiple choice items were analyzed based on their order in the examination. Therefore, the general display of the analysis of the item is followed by the items facility value and discrimination index. Finally, the facility result of items in their category is discussed. The results are displayed below.

Table 1:The general display of the analysis result of all item

ITEM NO	IF	DI	ITEM NO	IF	DI
1 word order	.58	-.1	61	.27	.53
2	.78	.22	62	.38	.21
3	.59	.20	63	.42	-.1
4	.38	-.5	64	.38	.35
5	.83	.21	65	.39	.06
6 paragraph coherence	.26	.24	66	.29	.21
7	.81	.23	67	.46	.40
8	.68	.31	68	.55	.71
9	.39	.21	69	.23	.48
10	.42	.27	70	.36	.21
11	.51	.40	71	.35	.21

12	.30	.50	72	.42	.27
13 Reading	.41	.33	73	.24	-.3
14	.47	.21	74	.16	.33
15	.54	.43	75	.41	,04
16	.17	.36	76	.14	.02
17	.50	.45	77	.30	.05
18	.50	.07	78	.56	.29
19	.26	.20	79	.44	-.3
20	.22	-.3	80	.37	-.1
21	.20	-.1	81	.18	.0
22	.37	.42	82	.56	.23
23	.16	.01	83	.40	.24
24	.34	.21	84	.33	.27
25	.27	-.2	85	.52	.51
26	.60	.21	86	.42	.61
27 vocabulary	.22	.21	87	.40	.16
28	.36	.45	88	.47	.60
29	.42	.41	89	.55	.52
30	.03	.21	90	.53	.41

31	.28	.61	91	.31	.15
32	.16	.43	92	.57	.40
33	.51	.40	93	.17	.22
34	.28	.41	94	.75	.0
35	.55	.36	95	.42	.32
36	.08	.90	96	.43	.0
37	.15	.04	97	.63	.16
38	.19	.36	98	.47	.72
39	.47	.25	99	.50	.41
40	.36	.22	100	.73	.31
41	.50	.08	101	.53	.23
42.	.20	.06	102	.77	.30
43	.24	.04	103	.61	.29
44	.32	34	104	.45	.27
45	.39	.21	105	.26	-.3
46	.53	.42	106 writing	.53	.34
47	.40	.23	107	.45	.30
48	.45	.27	108	.51	-.1

49	.42	.21	109	.46	.11
50	.39	.01	110	.64	.20
51	.32	.33	111	.26	.35
52	.29	.06	112	.45	.44
53	.51	22	113	.13	.21
54	.37	.57	114	.32	.20
55	.48	.69	115	.24	.35
56	.46	.21	116	.20	.26
57	.47	.6	117	.27	.21
58	.52	.50	118	.33	.35
59	.53	.34	119	.31	.44
60	.39	.21	120	.28	-.1

IF=item facility

DI=discrimination index

The above table presents the general analysis figure of all items. It is also possible to observe the specific number of items in the examination and their facility level and discrimination index figures next to each item.

Table2-The list, number and percent of all items in terms of facility value

	Item facility		
	Item no	**Total**	**%**
Very easy	no very easy item		
Ideal	1,2,3,5,7,8,10,11,,13,14,15,17,18, 26,28 ,29 ,33,35, 39, 40, 41, 46, 47, 48, 49, 53, 55, 56, 57,58,59,63,67, 68, 72,75,78,79,82,83,85,86,87,88,89,90,92,94,95,96,97 ,98,99,100,101,102,103,104,106,107,108,109,110,1 12	64	53.3
Reasonably acceptable	4,9,12,22,24,44,45,50,51,54,60,62,64,65,70,71,77,8 0,84,91,114,118,119	23	19.16
Very difficult	6,16,19,20,21,23,25,27,30,31,32,34,36,37,38,42,43, 52,61,66,69,73,74,76,81,93,105,111,113,115,116,11 7,120	33	27.5

The above table illustrates that in the present study 64(53.3%) of the items are items with excellent facility value (0.4-0.85), 23(19.6%) items are in range of reasonably acceptable (0.3-0.39), 33(27.5%) items are in the range of below 0.3.

Table 3-the list, number and percent of total items in terms of discrimination index

	Discrimination index		
	Item no	**Total**	**%**
Quite satisfactory	11,12,15,17,22,28,29,3132,,33,34,46,54,55,57,58, 61,67,68,69,85,86,88,89,90,92,98,99,112	29	24.16
No or little revision required	2,6,7,8,10,13,14,16,30,35,38,39,40,44,47,48,51,53 ,59,62,64,66,72,74,78,82,83,84,93,95,100,101,102 ,103,104,106,107,110,111113,115,116,118	43	35.3
Marginal(needs revision)	3,,5,9,24,26,2,45,49,56,60,65,70,71,73,87,91,97,,1 09,114,117,119,	21	17.5
Should be eliminated	1,4,18,19,20,21,23,25,36,37,41,42,43,50,52,63,75, 76,77,79,80,81,94 96,105,108,120	27	22.5

From the above table, one can understand that from all 120 items 29(24.1%) items have a good power to discriminate high achievers from low achievers (ranges $\geq$0.40). On the other hand, 43(35.83%) of the items have acceptable index of discriminating power while (0.2-0.29).21(17.5%) items need some revision whereas, a significant number of items 27(22.5%) exhibit negative power of discrimination ($\leq$ 0.9).

Table 4 –item facility result of items in their category

Total items	**Very easy**	**Ideal**	**Reasonably acceptable**	**Very difficult**
Word order(5)	-	4	1	-
Paragraph coherence (7)	-	4	2	1
Reading (14)	-	6	2	6
Vocabulary(31)	-	15	5	11
Grammar(28)	-	12	9	7
Communicative activities(20)	-	17	1	2
Writing(15)	-	6	3	6

As can be seen from the above table, there are five word order items. Among them, four of the items are ideal in terms of difficulty level while one is reasonably acceptable and there is no very difficult item.

Among the 7 paragraph coherence items under Table 4, only one item is very difficult whereas four of the items are ideal and two are reasonably acceptable.

In the above table, there are 14 reading questions. According to the analysis 6 of the items are found to be very difficult while 2 of the items are acceptable. The other six items are found to be ideal.

Table 4 also presents the vocabulary part of the examination. Among the 31 vocabulary items, 11 of them were found to be very difficult while about half (15) of the items were ideal. On the other hand, five of the items were reasonably acceptable.

In table 4, the item facility of grammatical items is also presented. From 28 grammar items, 12 items are found to be ideal. On the other hand, 9 items which range between 0.4-0.85 are medium and 7 items which score 0.3-0.39 are very difficult. There is no a very easy grammatical item in the exam.

As can be seen from table 4, there are 20 dialogue filling questions for communication purpose. Among them only two items are found to be very difficult while the remaining items are very good (17) and in the range of medium (1) facility value.

The last section of the exam is writing which contains 15 items. Of which, 6 items are found to be very difficult, and 6 of the items have very good facility value. Whereas, the other three items of the writing section are not found to be very easy or very difficult.

Table 5- discrimination index result of items in their category

Total items	**Should be eliminated**	**Marginal (needs revision)**	**No or little revision required**	**Quite satisfactory**
Word order (5)	2	2	1	
Paragraph coherence(7)		1	4	2
Reading(14)	6	2	3	3
Vocabulary(31)	7	4	10	10
Grammar(28)	7	5	10	6
Communicative activities(20)	3	3	7	7
Writing(15)	2	4	8	1

According to the above table, two word order items should be eliminated and the other two items need some improvement. Only one item is reasonably acceptable which ranges 0.30-0.39.

Paragraph coherence questions are good in discriminating: only one item needs some improvement, two of them discriminates quite satisfactorily and the rest 4 items are reasonably acceptable.

Among 14 reading items in table 5,6 of items don't discriminate at all (should be eliminated). On the other hand, two of items need some improvement, and 3

of them are reasonably acceptable. The other three reading items remain quite satisfactory.

From Table 5 one can easily understand that most of the vocabulary items in the examination found to be good in discriminating high and low achievers of the 31 vocabulary items, 10 of the items are quite satisfactory, 10of them are reasonably acceptable, and 4 items are marginal which needs some revision. However, the rest 7 items should be illuminated.

According to Table 5, the discrimination power of grammar items shows that 12 of the items are unable to discriminate students at all. Whereas, 6of the items have good discrimination index and the rest 10 items have reasonably acceptable result of discrimination index.

When we see the discrimination power of communicative activities in the exam, 6 of the items don't discriminate between high and low group students. The other 7 items are in the range of reasonably acceptable. The rest 7 of the items are good in discriminating high and low achievers.

In general, there are 15 writing items included in the examination. However, only 1 of the writing items does discriminate well. Six of the items in this part have a poor power of discrimination while the rest 8 of the items are reasonably acceptable.

Table 6- list of problematic items in their category

Category	**Problematic items in terms of**			
	Item facility	**Total**	**Discrimination index**	**Total**
Word order			1,4	2
Paragraph coherence	6	1		
Reading	16,19,20,21,23,25,	6	18,19,20,21,23,25	6
Vocabulary	27,30,31,32,34,36,37,38,42, 43,52	11	36,37,41,42,43,50, 52	7
Grammar	61,66,69,73,74,76,81	7	63,75,76,77,79,80, 81,	7
Communicative activities	93,105	2	94,96,105	3
Writing	111,113,115,116,117,120	6	108,120	2

The correlation of facility value and discrimination index of the items was found to be negative. In other words, items with a very good facility value do not necessarily have a very good power of discrimination. Proportionally very difficult items are found with good index of discrimination.

4.2. Discussion

Administration of an objective test and use of item analysis at the end of the period of instruction, sometimes even as small as a single lecture, has great advantages for the teacher. It enables him/her to get an active feedback from the students and determine areas which require emphasis, reinforcement or a revision in teaching methodology, perhaps using other learning aids. In the ranking situation, usually items which have a good (positive) discrimination and ideal difficulty are chosen. In fact, teachers must aim at getting high facility values and low discrimination indices, as the aim of classroom teaching is not to distinguish between good and poor students, but to ensure that all students have learnt the lesson correctly.

One Best multiple choice question (MCQ), if properly written and well constructed is one of the strategies of the assessment tool that quickly assesses any level of cognition according to Bloom's taxonomy. Therefore, it is important for teachers to evaluate their MCQ items to see how effective they are in assessing the knowledge of their students, and in predicting their total test scores. Difficulty and discrimination indices are fundamental tools to check whether the MCQs are well constructed or not.

High difficulty index should be reviewed by the respective content experts. This study serves as an effective feedback to the respective experts about the quality control of every year’s exam. When the difficulty index is very small, indicating difficult question, it may be that the test item is not taught well or is difficult for students to grasp. It also may create anxiety within the examinees and may be inappropriate at that level for the student. The wide scatter of item discrimination values for questions with a similar level of difficulty may reflect that some extent of guessing practices still occurred.

Developing the perfect test is an unattainable goal for anyone in an evaluative position. Even when guidelines for constructing fair and systematic tests are applied, excess factors may enter into a student's perception of the test items. Looking at an item's difficulty and discrimination will assist the test developer in determining what is wrong with individual items. Item and test analysis provide empirical data about how individual items and whole tests are performing in real test situations.

Regarding facility value, in this study, many items 33 (27.5%) are very difficult. As English is a common course and the exam has to be given for both social and natural science students, we need to take into account that there are different levels of students from different social group. But based on the result of this study it is difficult to say such things are considered.

When we see the proportion of difficulty level of items from the seven sections of language parts in the exam, vocabulary items are the most difficult items; 11 (9.2%)of the items range below 3.0(extremely difficult). However, grammar items are better than vocabulary items 7 (5.8%) of the items are difficult and need to be improved or discarded. Both reading and writing items are proportionally easier than vocabulary and grammar. 6(5%) items from writing items and 6(5%) others from reading are found need to be improved or discarded since they are very difficult questions. However, there is no very difficult item in the word order, but 1 item from paragraph coherence and 2 items from communicative activities are found to be very difficult items. Most of the items are found to be the range of ideal and reasonably acceptable facility value.

The average result of difficulty level of the items is 0.3 which reveals that the exam is not as such extremely difficult in terms of facility value. Rather, it is in the in reasonably acceptable range which needs some revision. In addition to

the average result of facility value, 53 %(more than half) of the items are not very difficult.

Discriminating power is one powerful indicator of item effectiveness. If the test and an item measure the same ability or competence specially when the exam is a group of high stake tests, we would expect that those items having a high overall test scorer would have a high probability of being able to answer the item. We would also expect the opposite, which is to say that those having low test scores would have a low probability of answering the item correctly. Thus, a good item should discriminate between those who score high on the test and those who score low.

Regarding discrimination index, in the present study 27 (22.5%) of the total items have an index of discrimination< 0.1 (should be illuminated). This shows that no focus is given to the effectiveness of individual item, Experience of time wise test item analysis helps to avoid this kind of high number of not functioning test items.

Proportionally from the seven parts of the examination, most of the test items in the vocabulary section have a good discrimination power. Among 31 of vocabulary items, 10 are in the range of satisfactorily discriminating and another 10 items needs no or little revision. The grammar test items have also a good power of discriminating. From a total of 28 items, 6 items are quite satisfactory and 10 items need no or some revision. Except one of the items in word order, all of them do not discriminate at all. The other parts of test items are found to be not discriminating, and they need to be discarded or improved in order to increase the effectiveness of the exam.

The average discrimination index of the whole test is 0.14 which means most of the items need revision. This indicates that most of the items in the exam do not have a good range of discrimination index.

The correlation of items facility value and discrimination index is negative, which means that items with good facility value are not likely to be discriminating, and items which have satisfactory index of discrimination were found to be very difficult. Only 5 reading, 3 vocabulary, 2 dialogue, and 1 writing items are problematic both in difficulty and facility value.

The quality of test items may be further improved based on an action that can be taken in reviewing the distractors by the item writer based on the calculated discrimination and difficulty index values. Items showing poor discrimination should be referred back to the content experts for revision to improve the standard of these test items. It is important to evaluate the test items to see how effective they are in assessing the knowledge of the students based on the difficulty and discrimination indices of the test items.

Difficulty of the instruction can be a cause for ineffectiveness of the items. In addition to the numerical result of the items, in this study inconsistent language and unclear instruction are observed in the exam. For example the instruction for the grammar section lacks consistency.

It reads: *Questions 58-85 are incomplete sentences. There are four alternative words or phrases, A-D, Given below each question. Choose the word that best completes the sentence and blacken the letter of your choice on the separate answer sheet provided.*

There are items which do not meet their objective. The mismatch between the instruction and the intended objective to meet within the item is one factor which makes the item ineffective. To make the item effective there should be only one objective consistent with the instruction given. But when we see some items in the examination there are items which don't examine the intended ability of students. For instance, in the last section of the exam the writing skill of students is the intended skill to assess. But most of the items under this category are items which miss their objective. For example:

Item number 110 has written as below

- *Not suitable for children under 10! This is most likely a warning:*

A. in a movie *C. in x ray room*

B. on a box of whisky *D. on a pack of cigarettes*

The above question is prepared to assess students' writing skill, but it is not likely assessing the intended skill.

Distractor efficiency is one basic thing in the effectiveness of the item. All distracters must be functional. If distracters are not plausible enough, the effectiveness of the item could be damaged. As a result, it affects the quality of the exam. In this exam, there are almost not functioning distracters which were chosen by less than 5% of the examinees.

Chapter Five: Conclusions and Recommendations

5.1. Conclusions

In this study the researcher has analyzed all the items of the national university entrance examination of 2005 E.C ,which has seven sections; word order, vocabulary, grammar, reading, paragraph coherence, communicative activities and writing. As a result, based on the above discussion, the following conclusions have been drawn.

According to the result of the study, 1 item from paragraph coherence, 6 items from reading, 11 items from vocabulary, 7 items from grammar, 2 items from communicative activities and 6 items from writing section of the exam were found to be very difficult. Totally 33 (27.5 %) of the items in the exam are very difficult based on the range of facility value. Therefore, from the above data one can conclude that not functioning items are too much according to the standardization level of the exam. These items do not fall in the standard range of facility value, which leads to incorrect generalization about students' results.

According to the discrimination index, 2 word order items, 6 reading items, 7 vocabulary items, 7 grammar items, 3 communicative activity items and 2 writing items were exhibited not to be discriminating (should be eliminated). Totally, 27 (22,5%) of the exam items are under the standard range based on discrimination index. So, based on this result, it is possible to say that the quality of many items in the exam is under the standard range of discrimination index.

As it can be seen from Tables 4 and 5, most difficult items are from the vocabulary part of the exam. The grammar part of the exam also has a high number of difficult items. The reading items are also found to be poor in

difficulty level and discrimination power. 6 items were found to be very difficult. Unlike the above four sections, the other three sections: word order, paragraph coherence and communicative activities are not difficult and are discriminating.

In this study the correlation of facility value and discrimination index of items were negative, which means that all of the difficult items were not similarly bad in discriminating between high and low achiever students. And items which have bad discrimination index are not mostly difficult. From problematic items only 10 are both difficult and they are better in discriminating in comparison with others.

5.2. Recommendations

High stake tests have a great role of identifying students learning background, deciding their future educational career and have a great role of creating either anxiety or positive feeling towards a subject. So test makers need to be careful in preparing this kind of tests.

Therefore, based on the conclusions, the following recommendations have been forwarded.

- Once the test is administered, its effectiveness has to be assessed based on different criteria, like facility value, discrimination index and distractor efficiency in order to decide about the score of students. Therefore, the researcher would like to stress that the examination agency needs to develop the habit of doing item analysis in order to assess the practical quality of every year's examination before scoring the exam.
- Very difficult and very easy items need to be properly reconstructed and revalidated.

- As not functioning items affect the quality of the exam, it leads students, teachers and other stakeholders to generalize incorrectly about their students; therefore it is better if not functioning items are improved or discarded before scoring the university entrance examination.
- This high stake test has great relevance for curriculum designing and materials preparation. It is better if the ministry of education develops a habit of having detailed report of the exam rather than generalized, pass /fail report.
- Based on the result of the study, vocabulary items were found to be the most difficult part of the exam. If the vocabulary is somewhat difficult, the item will likely measure reading ability in addition to the achievement of the objective for which the item was written. Therefore, in preparing the exam the researcher recommends that it is better to use difficult and technical vocabulary only when essential for measuring the objective.
- As distractors have a great role in insuring the quality of the item, making the distractors plausible enough and functional so that they will be chosen by students who score higher in the whole test result is essential. As a result, they will have positive discrimination power.
- Difficulty of instruction affects the effectiveness of the items. In the examination instructions with not clear and inconsistent language are observed. So, instructions must be set in a clear, short and precise language as much as possible.
- After the exam is administered, it needs to be checked to see if the item is perfect in assessing only one objective or not since an item is expected to assess only one objective. In doing this the exam would be better if the irrelevant redundancy is removed from it. For instance, the last section of the exam (writing part) contains different items with different objectives rather than assessing writing skill .So it needs to be improved and reconstructed to directly assess the intended student's ability. Rather, it leads student in to confusion over what he or she is being asked to do.

References

Alderson, J. C, Clapham, C and Wall, D, (1995), *Language Test construction And Evaluation*.Cambridge: Cambridge University Press.

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

Bachman, L.F (1996), *Language Testing in Practice*. Oxford, Oxford University Press.

Baily, Katheleen,M,(1998).*Learning About Language Assessment*, Heinle& Heinle Publishers.

Bloom, B.S (1956*) Taxonomy o f Educational Objectives*. Handbook 1 : The cognitive Domain New York.

Boophathiraj.C(2013) *Analysis of test items on difficulty level and discrimination index in the test for research in education*. International journal of social science &interdisciplinary research. Vol 2(2) online available at Indian research journal. com. Retrived in March 20/2014.

Brown,H.D, (1996), *Testing in Language Programs*. Prentice Hall Regents.

Crocker. L. &Algina, J, (1986) Introduction to Classical And Modern Test Theory. New York: Holt, Rinehart And Winston.

CyrilJ.Weir (2005), *Language Testing and Validation*. For Research in Testing Evaluation and Curriculum. Roehapton University Press.

Davidson,F.and Brian K.Lynch (2002) Testcraft: *A Teachers Guide To Writing And Using Language Test Specification.* Yale University Press

Ebel RL.(1972) *Essentials of Educational Measurement* (1st Ed) New Jersey: Prentice Hall.

Ebel.L.E & Frisisble, D.A.(1991) *Essentials Of Educational Measurement*. Prentice Hall.

El---Uri FJ, Malas N, *Analysis of use of a single best answer format in an undergraduate medical examination .Qatar Medical Journa l*2013:**1**http:// dx.doi.org/10.5339/ qmj.2013.1. Retrived in march14/ 2014

HarroldS.Madsen.(1983) *Techniques in testing*.Oxford University Press.

Henning (1987) A Guide To Language Testing Development, Evaluation, Research. London:Newbury House Publisher.

Hotiu ,A,(2006),*the relationship between item difficulty and discrimination indices in multiple choice tests in a a physical science course*. M.A thesis,florida Atlantic University.

Kassaw baye (2006) *An exploration of content validity of 1997 E.C EHEFC English Language Examination*. M.A Thesis .un published.

Madsen, S, (1967) *English Language Testing in Ethiopia*: The ESLCE Examination, Ethiopian Journal of Education.

Madziah and Lide ,(2006),classical and rasch analysis of dictomously scored reading comprehension test items. Malysian journal of ELT research.vol.2, International Islamic university Malaysia.

Mein Shih,Y(2010) *An item analysis of test items on difficulty level and discrimination index in the test for research in education.*

McNamara , T,(2000),*Language Testing*, Oxford: Oxford University Press.

Messick, S.(1989),Validity: *Educational Measurement* (3rded) New York, American Council & Macmillan.

Omirin(2007) *Difficulty and discrimination indices of three-multiple choice tests using the confidence scoring procedure*. Educational research and reviewVol.1.Accademicjournal.Onlineathttp://www. Accademic journals. org/ERR.Retrived in February 21 2014.

Rahim and Jaleel,(2002),*Anylisis of one best MSQ :the difficulty index,discrimination index and Distractor efficiency*:journal of Pakistan medical association.

Simachew Gashaye(2011) wash back of the University Entrance English Examination(UEEE) on teachers and students practices: the case of preparatory school in Amhara National Region.

Wood, D.A,(1960),*Test Construction Development and Interpretation of Achievement* Test, Colombus, OH. Merill Books

Appendixes

INSTITUTE OF EDUCATIONAL RESEARCH (IER)

ADDIS ABABA UNIVERSITY (AAU)

ETHIPIAN UNIVERSITY ENTRANCE EXAMINATION (EUUEE)

ENGLISH, GINBOT 2005/JUNE2013

BOOKLET CODE: 14 **SUBJECT CODE: 01**

TIME ALLAOWED: 2 HOURS

GENERAL DIRECTIONS

THIS BOOKLET CONTAINS ***ENGLISH*** EXAMINATION. THE CODE FOR THIS EXAMINATION IS 01 AND THE CODE FOR THIS PARTICULAR BOOKLET IS **14**.PLEASE COPY THESE CODES ON YOUR ANSWER SHEET WHERE READS ***BOOKLET CODE AND SUBJECT CODE*** .AND BLACKEN THESE CORRESPONDING BOXES IN THE COLUMNS BELOW EACH NUMBER.

IN THIS EXAMINATION,THERE ARE A TOTAL OF **120 QUESTIONS** DIVIDED IN TO **SEVEN** SECTIONS .EACH SECTION CONTAINS **MULTIPLE CHOICE QUESTIONS** CONSISTING OF FOUR POSSIBLE ANSWETRS.CAREFULLY SELECT THE BEST ANSWER AND BLACKEN ONLY THE LETTER OF YOUR CHOICE ON THE SEPARATE ANSWER SHEET PROVIDED.FOLLOW THE INSTRUCTIONS ON THE ANSWER SHEET AND THE EXAMINATION PAPER CAREFULLY.USE ONLY **PENCIL** TO MARK YOUR ANSWERS YOUR ANSWER MARK SHOULD BE HEAVY AND DARK,COVERING THE ANSWER SPACE COMPLETELY.PLEASE ERASE ALL UNNECESSARY MARKS COMPLETELY FROM YOUR ANSWER SHEET.

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

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

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

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