Proceedings of the Sixth National Conference on Private Higher Education Institutions (PHEIs) In Ethiopia

Major Theme: Nurturing the Teaching Research Nexus in Private Higher Education Institutions (HEIs)

Organized & Sponsored
By
St. Mary's University College

August, 2008
UN Conference Center
Addis Ababa, Ethiopia

Problems of Teacher-Made Tests: A Prime Concern for Quality of Education

Yalew Endawoke (PhD)

Abstract

The quality of education could be attributed to a number of factors which include assessment. The type of assessment techniques employed by teachers affect the decisions they make about students and their instructional management. The validity and reliability of the decisions made depend on the quality of tests. In our educational system, the promotion and repetition of students in their schooling are determined by the achievements of students measured by classroom or teacher made tests. The quality of the tests is ensured by applying the principles and suggestions of item writing. Consequently, in this study 9 tests of various subjects prepared by teachers from higher learning institutions (Private Colleges), a TVET college, and Higher Education Preparatory Secondary Schools were analyzed for their quality. The results indicated that 100% of the tests had problems related to language use, consideration of item writing principles and suggestions, content validity, and other technical issues. The implication of the study is that the quality of education could be severely jeopardized by the nature of the tests developed and used by teachers. So assessments used by teachers should be the concern of all stakeholders. Ignoring this crucial element of the educational process is tantamount to paying no heed to the quality of education which leaves the objectives of the national policy unachievable.

Introduction

These days, Ethiopia is expanding education not only to meet the Millennium Development Goals set out by the UN but also to fill the labor market demand of the country in various fields of specializations. The human capital requirement of the country is much higher than the market demands, specially, in the areas of Engineering, Science and Teaching professions, to mention a few. To respond to the ever increasing demand of professionals and skilled human power, the Ethiopian Government aggressively invests a huge amount of fund to the establishment and expansion of higher education. Student population which is booming at primary and secondary education, on the one hand, and the ever increasing demand of human capital, on the other, serve as a push factor to the government to expand higher education institutions. Such efforts can only be materialized if and only if those rudimentary elements of educational quality are well defined and met.

The teaching-learning process is organized, arranged and planned in a systematic way so as to effectively and efficiently achieve those educational goals set out by the government. More specifically, the Education and Training Policy (1994) expounded that the purpose of formal education is to promote the physical, intellectual, and social development of children so that they become useful citizens of the country, as education is the process of increasing the knowledge, skill, and understanding ability of a person that makes one's life more interesting and enjoyable (World Book Encyclopedia, 1985). But how do we know whether students develop the required level of skills, understanding, and ability to carryout the expected activities or not?

Ethiopia is a developing nation that is striving hard to become self-reliant, and endeavors to eradicate poverty in the coming few years, and attain the developmental level of middle income countries. Such efforts are not, however, without challenges and problems. The road to development is painstaking and rough as well as looks like a mirage that seems visible but difficult to reach and touch. There are considerable challenges in recent decades that encounter the country. One such challenge includes quality of education. It is a fact that there are undeniable developmental indictors that evidenced the country's move towards a positive direction. The efforts so far made

should be complemented with other consequential inputs to development, one of which is education. It is believed that the more educated the society becomes, the more economically and socially advantageous it becomes. In other words, the future economic and social well-fare of the society relies heavily on the level of education everyone attains. It has been argued that education positively and strongly correlates with the economic and social development of a society. That is why countries' all over the world invest large sums of budget to the education sector.

Expansion is one of commendable and major steps the Ethiopian Government is taking. Yet, one grave concern for the nation that attracts the attention not only of those that are directly involved in the process but of the government and the society at large is the quest for the quality of education.

The quality of education in Ethiopia is becoming one of the major concerns of various stakeholders. Quality of education, which can be defined in a number of ways, becomes an everyday topic for teachers, educators, educational bureaus, the government and parents. The major purpose of education is to enable everyone "to learn, realize their full potential, and participate meaningfully in society" (UNICEF, n.d.).

Enrollment rates, equity, participation rates, and other educational factors are given due attention by both the regional and federal governments and the outcomes are immense and encouraging. In spite of such increments in enrolment rates, and the attention given to equity and equality, the UNICEF (n.d.) report indicated that too many children are learning far less than what they are taught about or what they ought to learn in school.

The results of the three national learning assessments conducted in country clearly crystallized such claim. For instance, in the third national learning assessment over 10,000 grade 8 students have participated, and were tested on Mathematics, English, Physics, Chemistry, and Biology. The results showed that no one region scored a minimum pass point, which is 50%. The overall average (the national composite score) was a little more than 35%.

At this point a number of questions may be raised. Is it due to the low quality of education provided to the students? Is it due to the inefficiency of teachers in providing students with the necessary knowledge and skills? Is it due to factors related to the students themselves? Who is more accountable to this low achievement of students?

This could imply the very critical issue of the quality of education. This low learning achievement could be most likely due to a combination of factors that include low morale of teacher, inadequate professional training of teachers, inadequate learning environments, inappropriate teaching methods and frequently unmotivated teachers, and the nature of assessment techniques adopted by teachers.

Studies on the quality of education took into account a number of factors as indicators or standards of education. According to National Center for Educational Statistics [NCES] (2000), a school that is characterized by high quality is a major cause for students' success. It is unquestionable that multitudes of factors account for students' academic successes and failures. NCES gave much emphasis to school quality as the main cause of student learning and performance. It argues that defining school quality is the first step toward measuring and monitoring it. It acknowledges that both social and academic dimensions are important ingredients of student learning. The social component, which includes attitudes, ambitions, and mental well-being of students, resulted from interactions students have with their parents, teachers, friends, and significant others; and the academic dimension pertains to student learning which is affected by in-and out of school contexts. NCES (2005: vii) stated the following on the issue.

> Many factors are associated with school success, persistence, and progress toward high school graduation or a college degree. These include students' early school experiences, motivation and effort, and courses taken and other learning experiences, as well as various student characteristics, such as gender, race/ ethnicity, parents' educational attainment, and family income. Monitoring these factors in relation to the progress of different groups of students through the educational system and tracking students' attainment are important for knowing how well we are doing as a nation in education.

Though the factors are many and vital, the Center responds to the timely concern of the nation which is the academic quality "by focusing solely on the school characteristics that have been shown to improve student learning" (NCES, 2000: 2) and ultimately their performance. Accordingly, as "student learning is, in part, a function of various characteristics of the schools and the process of schooling, examining the characteristics of schools that are related to learning illuminates some of the reasons why students are, or are not, learning at optimum levels" (NCES, 2000: 1).

The Center identified 13 indicators of school quality on the basis of recent research suggestions related to student learning. These indicators were grouped into three major categories classified as the characteristics of teachers, the characteristics of classrooms, and the characteristics of schools as organizations. As represented in Figure 1, those school quality factors can affect student learning both directly and indirectly (NCES, 2000:4).

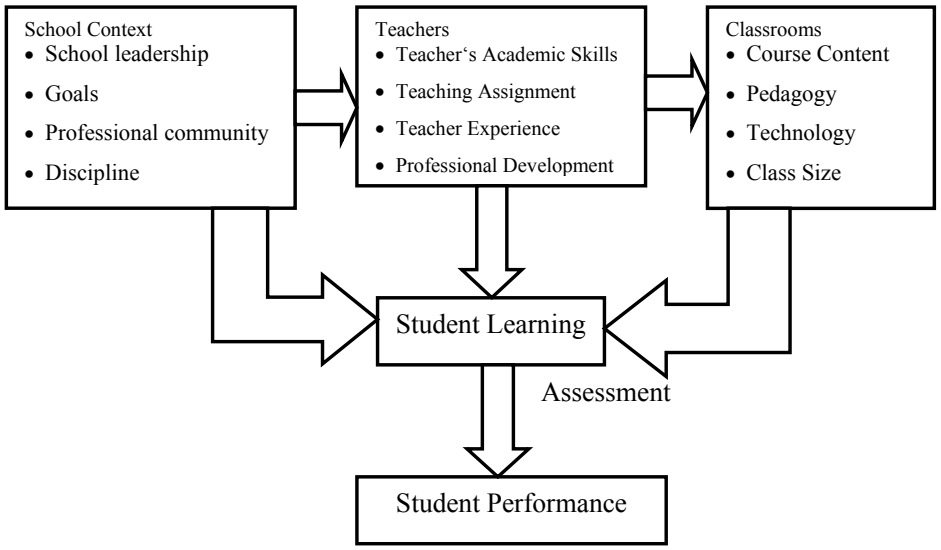


***Figure1**: School Quality Indicators and their Relationship to Student Learning and Performance (adapted with modifications from Monitoring School Quality: An Indicators Report, NCES, 2000)*

Research repeatedly showed that teachers are crucial elements in the learning of students. For example, Hanushek (1992), as quoted by NCES (2000: 5), 'The estimated difference in annual achievement growth between having a good and having a bad teacher can be more than one grade-level equivalent in test performance.' Moreover, NCES (2000: 5) revealed that teacher quality is the most important determinant of school quality by quoting the result of a study by group of researchers:

This analysis identifies large differences in the quality of schools in a way that rules out the possibility that they are driven by non school factors ... we conclude that the most significant [source of achievement variation] is ... teacher quality ... (Rivkin, Hanushek, and Kain, 1998: 32)

Of the various criteria that could be used to define teacher quality, the most salient one is teachers' ability and skill to construct and use quality assessment techniques that best measures their students' mastery or achievement of the learning objectives.

According to Ferrara (2007:18), student outcomes serve as indicators to weakness of a system. She, thus, stated:

> Deficiencies in school functioning or in student learning or performance are seldom merely the result of a single weakness in the organization of the school or in poor instructional programming in a particular area of learning. Rather, deficiencies in school or student performance or in school and student outcomes often serve as indicators of systemic weakness--horizontally, vertically, and interactively—in leadership, in characteristics of the culture, in programming, in the choices of instructional methods and approaches across the learning environment, and in decision making at all levels of the school and the instructional program.

In any educational institution and curriculum provision, evaluation of the attainment of the objectives set forth by teachers, schools, education bureaus and/or Ministry of Education is of a prime concern. Measurement experts argue that evaluation procedures,

course objectives, instructional provisions, and course contents are interrelated wherein one cannot go without the other. Evaluation plays a central role in ensuring whether the objectives have been achieved using the course contents or not. Without evaluation it is impossible to determine the extent to which students have mastered the courses they have been taught, the degree to which teachers were effective in their lesson presentations and other instructional processes, and to what degree the objectives have been achieved. The interplay among instructional objectives, instructional processes, course contents, and evaluation procedures can be represented diagrammatically as in Figure 2. Our goals and objectives are the destinations we wish to reach, the instructional processes are the interactions that involve students and teachers based on the course contents, and the course contents serve as the vehicles that take us to the desired outcomes we wish to effect in our students, and the evaluation procedures are the compass that indicate whether we are on the right direction or not. In this case, any failure in any one of these educational process elements could lead to a failure in the other three. Specially, devastating would be the problem of having valid and reliable assessment and Specific, Measurable, Attainable, Realistic and Time-bound (SMART) instructional objectives.

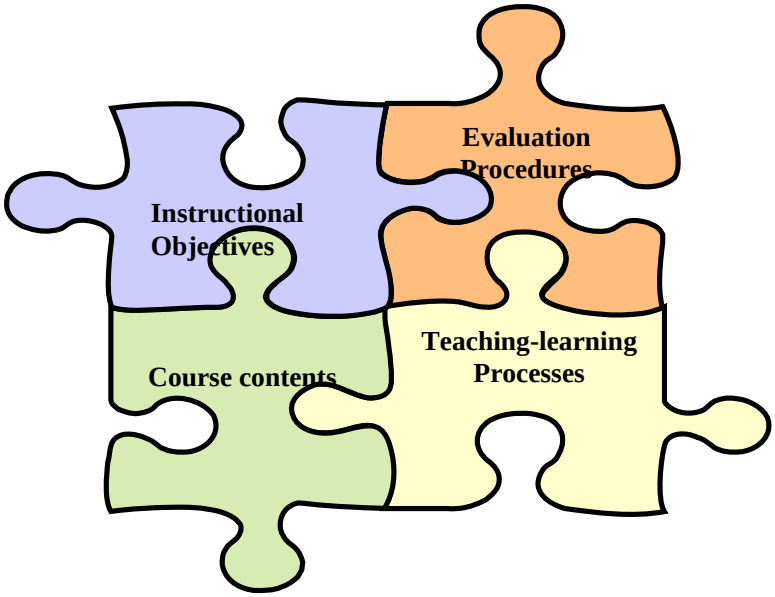


***Figure 2**: The Interplay among the four basic Elements of Educational Process.*

In general, evaluation serves a number of purposes not only at classroom levels but also at local and national levels. It is used to determine whether the education system lives up to the objectives set or not. It also helps to identify the weaknesses and strengths of the education system, and provides formative information for system managers, directors, teachers and a wider society. It serves as a ground to open up dialogue and provides the basis for development planning and school improvement. Evaluation helps to focus on processes intended to achieve those learning outcomes.

Measurement experts argue that in the absence of good measures of performance, the quality reform process could not be well-guided. This is tantamount to saying without measuring the effectiveness and efficiency of performances, it is practically difficult to determine whether it is up to the standard or not. Either quantitatively or qualitatively, the human mind "measures" the status of the performance achieved by individuals who performed certain tasks.

In every walk of life, we talk about decisions, decisions, decisions…. But decisions do not come out of vacuum. They are made based on valid and reliable data. The data that are used to make decisions are gathered through evaluation techniques which include tests, systematic observation, anecdotal records, interviews, questionnaires, assignments, projects, etc. depending on the nature of the behavior being measured and the type of decision to be made.

The quality of decision we make and the relevance as well as significance of actions we take depends almost entirely on the nature of evaluation techniques we employ to gather data. This can be best illustrated by Figure 3.

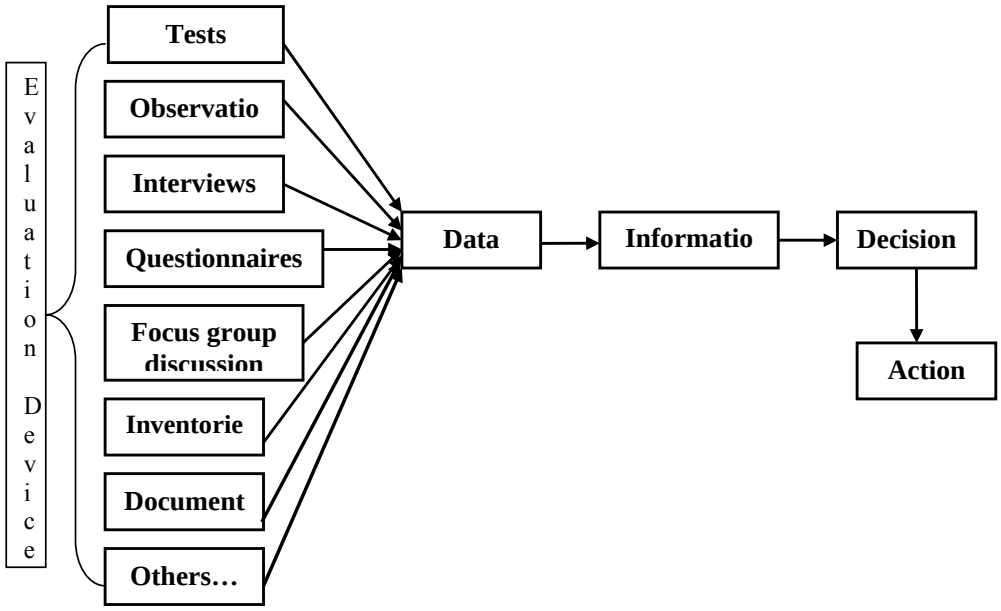


Figure 3*: The Link among Evaluation Devices, Data Collection, Decision, and Action.*

As can be easily understood from Figure 3, the reliability and validity of the data depend on the nature and type of evaluation devices used to gather the data that are used as inputs to making decisions and taking actions.

To ascertain whether these objectives are met or not, and maintain the quality of the teaching-learning process, evaluation of students' achievement play a predominant role. Supporting this, Cole and Chan (1994), Stapleton (2001), Elliot, Kratochwill, Cook and Travers (2000) argued that teachers constantly collect information through different techniques such as through asking questions, performance monitoring in class activities, written assignments and tests to determine students' level of achievement, to evaluate the effectiveness of instruction, to identify topics that require additional teaching and to plan new instruction that are considered important aspects of the teachers' roles. In addition, Ebel and Frisbie (1991:30) said, "to teach without evaluating the extent of learning would be foolish." This is due to the fact that it is through evaluation that one can understand whether the process is going well or not. Thus, without evaluation instruction becomes

meaningless. In line with this idea, the DES Task Group on Assessment and Testing, as cited in Stapleton (2001:3), stated that "promoting …learning is a principal aim of schools. Assessment lies at the heart."

In schools and higher learning institutions, it is tests that are most frequently used for making decisions about student progress, teaching effectiveness, and the achievement of the schools or institutions in general. Specially, in the Ethiopian education system, tests are the sole evaluation techniques used to make decision about students and other conditions related to the teaching-learning processes. Tests have the effect of motivating students to work hard, of helping educators make decisions regarding placement, selection, classification, of causing program improvement, of assessing student' progress; and of assisting employers to evaluate the competence of applicants and to recruit the prospective employees that fit vacancies. As a result, student assessment is pervasive in schools. Teachers construct daily, weekly, and term assessments for their classrooms.

Though measurement experts posited that effective assessments are used to tap and reflect students' abilities, achievement, skills and potential, and to make predictions about future behavior, but, unfortunately, the role of the assessment of students learning is not well understood and remained the under-researched aspect of higher education.

The study of assessment of students' achievement is important for a variety of reasons. In the first case, it is through assessment that we get any information about students, second it enables us to make decisions of any sort, and it helps us to improve the quality of the whole educational process. However, in order to do all this, it should be accurate and meaningful, as indicated in the previous section.

Therefore, the major part of the quality of education rests on the nature and quality of tests. Tests that fail to provide valid and reliable data are subject to erroneous decisions, which ultimately lead to flawed actions that jeopardize the social and economic health of the nation, in general, and students in particular. The testing is considered by NCES (2005: 1) as foundation for development by stating that "the nation's economic and social health depends on the quality of its schools. If students are not taught the values and

social skills necessary to become good citizens and do not learn the academic skills necessary to be economically productive, then the schools have not succeeded in their mission," which is ascertained through rigorous use of evaluation techniques including tests.

The assumption held in this study is tests are the primary concern of education quality. This assumption emanated from two serious problems the researcher observed from the tests developed by teachers at various levels. The first problem is the technical or mechanical aspect of the tests, and the second is the quality of the tests.

The course *Measurement and Evaluation* at a graduate level, requires students to do projects on tests developed by teachers in preparatory schools and higher learning institutions. The researcher, thus, realized that a broader study has to be carried out to unveil the problems of teacher-made tests to the public so that necessary measures shall be taken to curve out the problems.

To sum up, the purposes of this study are to answer the following questions:

1. What are the major problems of teacher-made tests?
2. Do teachers apply the principles and suggestions for writing test items?
3. How do test results affect the quality of education?
4. What are the major problems of teachers in the development of tests?

Methods

Participants and Sampling

The researcher used 9 test papers developed by preparatory school teachers, higher learning institutions, and a TVET college. The papers were selected using incidental sampling. Graduate students were sent to preparatory schools, the TVET College, and private colleges. The students collected the test papers from those teachers who were willing to provide them and others were obtained from the students who have taken them.

Data Analysis

Once the test papers were obtained, they were analyzed in terms of the general principles of item writing, their content validity, suggestions for item writing, the type of objective taxonomy and domain type they covered, and language use. In this case, checklists used by various measurement experts (e.g., Nitok, 1996; Gronlund, 1988; Mehrens & Lehman, 1984) were used in determining the quality of the tests. Students' textbooks and some course outlines were used to ensure the content validity of the tests. The data were analyzed using qualitative data analysis technique.

Results

In this study, nine papers collected from two private colleges, a government TVET college, and two higher education preparation secondary schools (HEPSSs) have been analyzed. Subject-wise, there were two Introduction to Psychology Final Tests (from a Private medical college), History Final Test (from a Private Teachers College), English Tests (from HEPSSs), one Introduction to Sociology Final Test (from a Private Medical College), one Social Studies final test (from a Private Teachers College), one Civics and Ethical Education final test (from a TVET College), and two Geography tests (one mid and one final tests from HEPSS). This pool of tests was randomly selected from the schools and colleges to be analyzed and evaluated. The results obtained are presented in the following section.

Test 1: Introduction to Psychology

This is a final test developed by a teacher in a private medical college. The test has four parts consisting of 25 multiple choice items with four options, 15 true-false items, 10 matching items, and 1 short item. For the multiple choice items, the students were instructed to "choose the best answer." In terms of taxonomy of objectives, 100% of the items measure cognitive domain which focused mainly on simple learning outcomes or knowledge level (86.3% or 44 items), 5.9% (3 multiple choice items) dealt with comprehension, and 7.8% (4 multiple choice items, viz., 5, 10, 12, and 13) somehow assess the application levels. Though this weakness (the emphasis given to simple

learning outcomes), which is a major characteristic of teacher-made tests, is not peculiar to this test, the grave problem is that it is not professionally prepared and useful in meeting the desired objectives. It seriously jeopardizes the performance of students and contributes to the deterioration of educational quality. One obvious problem of the test which is seen at the surface is language usage. It is imaginable that the teacher has a serious language problem. But on top of that, the teacher should be so careless in the preparation and editing of the test items. He/she might either be the only staff in the college to prepare the test or not willing to allow colleagues to revise or edit the test, or he/she might have no time to go through the test before it was ready for administration. Or the college may not have a section that checks the quality of the items, or experts who have professional competence to do the editing. In either way, the problem is so worse that it could affect not only the performance of the learners but also their language competence. As teachers are models for their students, there is a high degree of likelihood to copy what teachers do and say. In this case teacher' language problems could easily be 'transplanted' onto students, which is the unfortunate downsize of the quality of education.

The teacher in the test instructed the students to "choose the best answer" for the multiple choice items. Unfortunately the instruction given to the students is erroneous as all the items are correct type. The items deal with definitions, facts, names, and to some extent labels. In other words, the items focus on who, what, when, and where types of questions which have one correct answer that require the students to rote memorize names of persons, things, years, fact, definitions, places, etc. In such conditions the use of "the best answer type" is not appropriate. The best answer type is used when the items deal with questions of how, why, in what manner and other similar nature. In these types of questions there may not be one specific correct answer. All the options could be potential answers to the question, but comparatively one is the best answer from the given list of alternatives that best satisfies the item.

Items 1, 2, 3, 4, 5, 7, 9, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 24, and 25 had serious language and/or conceptual problems and the majority of them are meaningless and difficult to understand. For example, take item number 1. It says, "Which one of the

following stage; that children are sexually dormant and active in playing & learning basic skills." It is grammatically erroneous, and conceptually confusing to the students. Take item 2. It runs as follows: Who is the physician that classified personality with related bodily fluid? The reader can easily understand how difficult it would be to the students to conceptualize the idea presented in the item. Or the students may answer it simply being led by the clues included in the item, such as "physician", and "bodily fluid". See also items 15 and 21. Both items start with blank spaces. Item 15 says: "_____ a process of attaining adult form is:-" This item is so vague. It is also copied directly from source material wrongly. Item 21 has a clue that could help uninformed students to get the correct answer. The item runs as follows:

"_____ is the scientific study of how people think, feel, and behave is [sic] social situation.

a) Developmental psychology b) Social psychology

c) Industrial Psychology d) Health psychology"

Here, the probability of getting the correct answer is high because the word "social" in the stem gives clue to the students to associate it with the word "social" in alternative "b", which is the correct answer.

In the test, the option "None" or "None of the above" was used where it should not be and more than suggested by measurement experts.

Coming to the true-false items, the serious problem is taking statements directly from source materials. This action is highly discouraged by measurement and evaluation experts (Linn & Gronlund, 2000; Mehrens & Lehman, 1984) for at least two reasons: taking statements directly from sources encourages rote memorization which has little effect on the learning quality of the students, and directly copied materials become so confusing and vague that students may find it difficult to understand the question and to provide appropriate answers. One can observe how the majority of the items in the true-false part are so vague and difficult to answer. For instance, item 15 says, "Emotion is classified into four". This is a directly copied idea from the source material. But the answer could be both true and false depending on the type of information the testees have

or the criteria they use to classify emotions. What criteria were used to classify emotions: pleasantness–unpleasantness; intensity; behavioral and physical components, or any thing else? What did the teacher expect from the students: the types of emotions such as anger, happiness, rage, sorrow, anguish, joy, sadness, etc? Or what? What does "four" indicate in the item?

In the true-false items the majority of the items are picked from the same topic areas. Items 1 and 2 are related items. Items 4, 6, and 7 are similar issues. This encourages guessing and rote memorization, or at worse it puts some students, who for some reasons gave little emphasis to that portion of the material, at a disadvantage. The arrangement of the items is so poor that it sequentially puts the items from the first chapter to the last.

In matching exercise, the directions are so poor, the items include heterogeneous materials, and some items can be answered based on general knowledge without reading or studying the course. All the items are poorly constructed. It is strongly suggested that the responses (lists in Column B) should be shorter than the "solutions" to the problems (premises listed in Column A). But here all the parts of the matching exercise contained a mix of both problems and solutions together in one column. For instance, the word "Id" in Column A had to be interchanged with the phrase "Animalistic nature" in Column B. The items listed here are so confusing to the students. For example, the answer for items 1, 5, and 7, could be "E", which is Freud. The answer for items 2 and 6, could be "B" or Maslow. The items are not mutually exclusive. There are also a number of errors in the list. The last section deals with short answer. One can see how nominal this section is that it only presents one item to the students.

Finally, the test lacks somehow content validity that it focused on only certain segments of the course that the items failed to represent the contents properly and efficiently.

Test 2: Introduction to Psychology

This is also another test developed by another teacher in the same private medical college. The test was a final examination conducted in September 2000 E.C. The test had 10 true-false, 15 multiple choice, 5 keyed response, and 10 matching exercise items.

Compared to the directions of this test with the previous test, this test is by far better in a number of ways. First, the directions told the students what to do, when to do, how to do, and where to write the answers. For example, if we take the true-false part direction it reads: Read each of the following statements and write "True" If [sic] the statement is true or write "False" If [sic] the statement is false on the space provided.

Apart from the problem of language, the true-false items suffered from clues that lead students to the correct answers. For example, if we take item 1, the answer is obviously "False" because the phrase "…every thing in every moment…" makes the statement absolute, which informs students to mark it false. Item 4 has also the same problem. Phrases "All … with one and the same …" qualified the statement in an absolute way which indicate the students to choose "False" as their response to the item.

In Part II, there are multiple choice items each of which involve 4 to 5 options. One of the major weaknesses of this part is it direction. Although almost all the items were correct answer type the direction instructed students to choose the best answer. The other weakness of these items was frequent use of "None of the above" and "All of the above or All". Measurement experts recommended that these options should be used sparingly as there are a number of problems associated with these options. But 12 of the 15 items used either one or both phrases as their options.

The strong side of this test was the use of application items, though the number of items was limited to three. Although the items are so confusing and lack homogeneity as well as characterized by language problem, the use of keyed response (see items 16 – 20) was one positive side of this test. The quality of the items is very poor and there is high probability of getting the correct answer by guessing simply because of the presence of clues in the items. The difference between the number of items and the number of responses is only one which maximizes the possibility of obtaining the right answer by method of elimination.

The other section of the test is the matching exercise items. In this section, there are 10 items. The problem here starts with the direction. The direction is not clear and

informative to the students. The items are so heterogeneous that the students have a good deal of opportunity to score the right answer by partial information about the subject presented. The majority of the problems or premises are put in column B and the short ones are placed in column A.

Still, this test also has a problem of content validity as it presents a few items from a large content area of *Introduction to Psychology* course.

Test 3: Social Science Test

This test was developed by a private teachers college to assess students' achievement of *History*. But, it did not indicate whether the test was designed for the purpose of final examination or for the mid-test. However, from the contents it covers, it looks more of final test. It consisted of 14 multiple choice, 5 true-false items, and 5 matching items, in that order. The majority of the items measure knowledge or lower cognitive level of learning objectives. All multiple choice items but 3 were knowledge items. Item 3 was comprehension. The problem of the test starts with the test direction. The test did not have general direction at all. Even it did not tell the reader which subject test it is, and the time to do the test is not also indicated. The directions for the three parts of the test are so poorly written that they did not show the nature of the items. For instance, the direction for the multiple choice items reads: "Chose the best answer." Though wrongly written, all the items were not "best answer". The directions of the true-false and multiple choice items are also very poor and unclear.

Almost all the items were not written in clear and understandable way that students can easily comprehend their meanings and answer the items based on their knowledge.

Consider item 3. "Prince Henry was given the name 'Henry the navigator' because

A. Due to un reserved [sic] support to navigation & exploration
B. He allocated a budget for making voyages & exploration
C. Due to the aim of circumnavigating Africa & enter in to [sic] Indian ocean [sic]
D. Due to his religious mission
E. All except 'D'"

In this item, there is a mismatch between the stem and its alternatives. In the stem part the word "because" did not grammatically parallel with options "A", "C" and "D". Therefore, there is high likelihood of selecting alterative "B" as this is more in agreement with the stem. This problem arises not from lack of knowledge of students but from the confusing nature of the item itself. The item could have been best written in the following way:

Why was Price Henry given the name 'Henry the Navigator"? That was due to

A. his unreserved support to navigation and exploration.
B. the budget he allocated for making voyage and exploration.
C. his aim of circumventing Africa and entering into Indian Ocean
D. his religious mission
E. all except D

Many items of the test could have been written in a similar fashion than they were actually written. Punctuation, grammatical and other forms of errors are rampant in the items.

In the true-false part, items 1 and 4 were not good items. They were not written clearly that students could understand them. Item 1 could have been rewritten as follows: "Slave trade was the main obstacle for European countries to occupy large territory in Africa in the 19th C." Here the qualitative term "large" is misleading to many students as they interpret it in various ways.

The last section, which is matching exercise, was also very poor in a number of ways. First, it does not have any direction. The list of items dealt with heterogeneous materials which could increase the probability of getting the answers by guessing. The items placed in both Columns A and B are not as per the suggestions of item writing for matching exercise. The short ones are positioned in the Premise part while the long ones are put in the Response section, which is against the suggestions. The test also lacks content validity as the items dealt with a few portions of the course.

Test 4: English Final Test for grade 11

The major purpose of providing English to students at various level of education is to enable them develop their language skills for the purpose of learning other subjects offered in English, and communicate with others using the language. To do so, students are expected to acquire proper knowledge of the subject matter. One way of ensuring their level of competence is the use of carefully crafted assessment techniques. Unfortunately, as you can see the test on Appendix X, this does not hold true. The test is so poor in a number of ways. One major problem is editorial. The test suffered from language problems and other technical aspects of measurement. It is cried by many educators, education bureaus, teachers, and the Government that students have serious problems in English. Can we expect our students to be competent while our teachers fail to teach them properly? Can we attribute the failure of their language competence to themselves? Many say that students' failure in their academic achievements is due to lack of using their mother tongue. However, experiences showed us that they are still weak in their own languages.

The test is not fit not only for the test of English language competence but for any other subjects. Look for instance the directions. The first direction says: "Three of the following statements are true and two them are false, circle only the true states ets, According to the passage." Apart from its obvious language problem it gives students clues to the answers. The students are told that two items are false while three are true, hence, they try hard to find three true answers and two false answers. If students get all correct answers, can we say that they know definitely the answers?! Unfortunately, not necessarily! Consider also the second direction: "Chose the best alternative following sentences /questions according to the passage."

Grammatical inconsistencies, mechanics errors, and other forms of test problems are common in the test. Its failure is not only obvious from measurement points of view but from the English language testing also.

Test 5: Final Test for Professional Course for Private Teachers College Diploma Students

This test was prepared by a private college for diploma students to measure their mastery of general method of teaching, which is one of the courses that provides students with the understanding of test preparation and other evaluation procedures. The test has 5 true-false, 5 short answer type, and 10 multiple choice items. Unfortunately, this test is poor in a number of ways. First, some of the items have little or no relevance to the course being measured. For instance: "Man can influence [sic] the social environment but the natural environment can't" has no significance to the course that dealt with teaching methods. The directions for true-false and multiple choice items are problematic. The direction of the true-false items runs: "Write true for corrects statement and write False for incorrect statement." This direction is vague that does not inform students of how and where to write the answers. The obvious language problem was evident in it. Negative words used in the true-false items were not also emphasized so that students can easily see them and answer the questions without overlooking them. Items 1, 3, and 5 had this problem. Item 4 is tricky that could easily confuse those impulsive students. It says, "Field trip method of teaching method is a planned visit to places inside the regular class room." Most students would mark this item "true" because much of the idea contained in it is correct. The part that makes the item "false" is the phrase "inside the regular classroom". Students sometimes answer a question based mainly on the information they get in the previous section of the item. The phrase "inside the regular classroom" does not add value to the item but confusion to the students.

In the second section, we find the short answer type which comprised confusing and difficult items to comprehend. Consider the first item: "_____ Approach of teaching refers when, students are given more chance to participate." How clear is this item to you as you read it? If I correctly got the central idea, the teacher wanted to ask the students the following question:

A teaching approach that gives students more chance to participate [during instruction] is referred to as _____.

Item 2 of this section has a serious problem. Students can provide a number of answers to this question as it is so vague that did not restrict and direct them to a certain point. The third and the four items also are ambiguous. All the other items did not follow the suggestions to prepare quality items. The students are requested to provide the responses (answers) before they read the questions. This is strongly not recommended by measurement and evaluation experts.

When we come to the multiple choice items, we can see that the majority of the items have one or more problems of punctuation, language, conceptual or principle. Some items are trivial that they measure unimportant knowledge of students. For instance, item 2 asks students the question:

"The grade level of second cycle primary education syllabus include:

A. 6-8 B. 5-8 C. 4-8 D. None"

How important is this question to this level of students? Aren't they the product of the system? The course is not meant to develop the general knowledge level of students but their understanding of the subject matter of teaching methods.

Some items have conceptual problems. For instance, item 8 says:

"Goals that are designed to be attained in longer periods are

A. Objectives B. Aims C. A & B D. None"

The concepts aims, goals and objectives are different and have various meanings. Goals are concepts used in their own right that represent intermediate level of planned actions that bridge objectives and aims. But, here, they are considered as synonymous with aims.

Practically speaking, the last item is not a multiple choice type. It is a true-false type that could have been included in the first section of the test. As this test is a final test, there are only 20 items of different format. Can you imagine how unrepresentative the items could be to the course content? Can we think of a test comprising 20 items to appropriately and sufficiently cover a semester's course? Certainly not! The test is not content valid. The use of "None" and "All" was inappropriate. Although 100% of the test

items measure simple knowledge, the students were instructed to "Choose the best answer from the given choice".

Test 6: An Introduction to Sociology - Final Examination to Medical College Students.

This test is by far better in a number of ways than the rest of those tests analyzed. The test has purpose, general direction, and proper place for students to write their name, department, ID number, and section. Moreover, it communicates to the students the measure to be taken if they cheat during the test. That is great!! The direction for the true-false items is properly and clearly written. The test contained 15 true-false items, 25 multiple choice, and 10 matching exercise items. The increase in the number of items enhances content validity.

Though the test has these strengths, there are a number of weaknesses. Some items in the true-false part are copied directly from the source material which, as stated earlier, encourages rote memorization by the students. Except a few editorial problems, the true-false items are fine and clear.

In part two, we have the multiple choice items. All the items measure simple rote memory of the students or simple learning outcomes at a knowledge level. The majority of the items have options that are circular – that is, ideas are repeated in two options of the same item (see items 22 and 23). Principles of item writing are violated in most of the items. For example, the negative words are not underlined, common words are not included in the stem, and items developed form various sections or units are not randomly distributed in the test. They are placed sequentially from the first to the last section. This could encourage rote memorization.

There are some conceptual, vague, and unclear items in this section. For example, consider item 21 which was presented as follows:

> "the state of keeping of divorcing and remarrying with different serious of partner or spouse is:-"

Getting the central idea of the item would be difficult to the students. Either they try to correct it, ask the teacher for clarification or simply guess the answer. The other problem observed is over-mutilation of the items. For instance, item 20 is written as follows:

"_____ and _____ are the types of cultural innovations."

This is not a good way of presenting items to students as it makes the idea confusing and leads them to provide many possible answers.

Content-wise, the majority of the items focus on sexual and marriage relationships. Other topics seem to be given less weight in the test. This could affect the achievement of the students in a various ways. Some students who gave much emphasis to this section may have the chance to score more than others who focused on other topics. Achievement difference would be the result of variation in emphasis rather than knowledge.
The direction is also misleading that it instructed the students to choose the best possible answer. Each item has one correct answer and students are required to identify that correct answer.

The last part of the test is a matching exercise. The strength of this section is that it places short items on the right and the longer ones on the left columns. Nonetheless, the items included here are so heterogeneous that the teacher cannot confidently ensure that the students' answers represent their knowledge on the subject matters presented to them. As heterogeneity increases, the possibility of getting the correct answer through guessing and elimination also increases. The direction is not as per the suggestions of writing matching exercise items.

Test 7: Civics and Ethical Education Final Exam for TVET Night Students

This test was designed to students in a TVET program. The test had 5 true-false, 12 multiple choice items, 3 short answer, and 5 matching exercise items. The good thing of this test is the use of different item formats like the other ones. The true-false items are characterized by general knowledge rather than knowledge based on the taught course. Anyone who can read and understand the items can answer them correctly without going to the class to attend the lessons. The direction is also a problem for students. The

students are not instructed where and how to write their responses to the questions. In multiple choice items, the items deal entirely with simple knowledge of facts, names, and definitions. But the students are told to choose the best answer from the list of alternatives which does not fit to the nature of the items. The majority of them treat general issues that students do not need to go to school to learn. Take the first item (item 6), the answer can be obtained by the students simply based on their general knowledge. Similarly items 7, 8, and 12 can be answered without attending a course in classrooms.

All "the fill in the blank" items did not follow the principles of item writing for short answer type. The blanks are placed at the beginning which would affect the response accuracy of the students.

Matching exercise items also suffered from the problem of heterogeneity of materials and improper positioning of the responses and premises. The direction is also poorly prepared. Some items deal with names, others with concepts and still others with practices. This doesn't allow students to study hard and focus on higher order leaning outcomes.

The validity of the test is also so poor that it did not represent the whole course content in an appropriate and representative manner.

Test 8: Geography Final Exam for Higher Education Preparatory Secondary School Grade 11 Students

In this test, there were 40 items partitioned into 3 sections of which the first comprised 5 true-false items, 6 matching exercise items followed by 22 multiple choice items. In this test, the true-false items had no major problems except items 3 and 5. In item 3, there were confusing statements on top of spelling errors. The item says: "The Atlantic Ocean drainage system is the largest drainage system interms [sic] of annual discharge but not interms [sic] of drainage density." In this item, the comparison reference is not given. It said the "largest" but in comparison to what? Against what is it the largest? Such use of qualitative terms has also been employed in item 5. The words "tremendous" and "large"

are vague and could confuse the students to give the proper answer. In general, however, in this test errors are minimized in the true-false items.

The direction for matching exercise items is somehow informative and instructs the students how they use the responses in column B, though, still, it lacks clarity. The good side of this part is, unlike the other tests analyzed so far, the use of homogeneous materials. All items deal with drainage systems. The problem here is the placement of the premises and the responses. They are interchanged their positions.

The last section of the test involves multiple choice items each of which have four options except item 18, which is a true-false item. All items, except item 28 which assesses comprehension, measure simple knowledge of facts, names, places, and specific figures. As result, the direction used is not proper. With the exception of some punctuations and editorial problems, the items are written clearly and the alternatives are placed in a proper order. The majority of the destructors are plausible in that they attract uninformed students as equally as the correct answers. This is one strong side of the items. The grave weakness of the test is its emphasis on only two content areas, namely; water bodies and drainage systems, and population. A few items deal with soil and wildlife as well as rainforest. This condition made the test to gravitate towards the measurement of limited course contents which could affect the learning and study habits of students. Hence, the problem of content validity is seriously put at risk.

Test 9: Geography Mid-Test for Higher Education Preparatory Secondary School of Grade 11 Students

This is almost similar in many respects with the previous test type. It consisted of 2 true-false, 7 matching exercises, 7 multiple choice, and 4 short answer items. Although it is good to have more types of item formats, it is not advisable to have two items in one section of a test. This would increase the probability of getting the correct answers by guessing. It has been indicated that item-for-item reliability of true-false items is low because of guessing effect (Mehrens and Lehman, 1984). The next section presented matching exercise items. The direction is unclear and could be confusing to the students. Consider the direction: **"**Match Item 'B' With Item 'A' and Note that item 'B' can be

repeatedly match with item 'A'." How clearly could you understand this direction? But fortunately students used to work on various matching exercise items in their past schooling and try to answer the items based on their experiences rather than on the direction given to a test. The items are copied from the previous test with some modifications in the Premises part (Column A). This is a bad practice in the evaluation process as students will be forced to search aggressively for previous test papers to get ready for the test instead of preparing themselves on the course contents they learned.

The multiple choice section started with an erroneous direction which instructed students to "choose the best answer from the given alternatives," while the items are all correct type ones. The test writer did not consider the majority of item construction principles such as avoiding the use of negative words or emphasizing them, avoiding or minimizing the use of "all of the above" and/or "none of the above", minimizing differences in the lengths of options, etc. Some items were bewildering to the students. See for example item 4. Spelling, punctuation and other errors are common in the test items of this section.

The last section was short answer type which did not completely follow test item construction suggestions. The blank spaces are put at the beginning of the test items. The last item looks vague and gives clue to the students to get the correct answer. The number of dashes informs the students of the number of required responses.

In this test, the items were pooled from the same content area: Water and related factors. Thus, the test has a limited possibility of enhancing the understanding of various course contents by the students. This would restrain their cognitive development in the said course.

Discussion and Implication

The major purpose of this study was to critically analyze test papers prepared by classroom teachers at preparatory schools and higher learning institutions to identify the strengths and weaknesses they had and to unveil the effects they could have on the overall quality of education. The assumption is that the quality of tests and other

assessment techniques employed by teachers ultimately determines the quality of education in general, and the teaching-learning processes as well as students' performance in particular. As stated earlier, almost any type of decisions made by students, teachers, educators, parents, and policymakers depends to a greater extent on the level and quality of students achievement, which in turn, relies entirely on the quality of assessments employed.

When we intend to develop and use any assessment technique, we have always some purpose in mind. We do not do assessment for the sake of doing it. We want to make sure we have achieved the objectives we set, to determine the extent to which students mastered the subjects they learned, to assure which part of the curriculum was challenging to the students and which was not, to ensure the effectiveness of the instructional process, etc.

When we make decisions based on the data gathered through various assessment techniques, our primary concern must be on the usability, reliability and validity of the data. We ask: Do the assessment techniques we developed measure what they are purported to measure? Do they generate information that enable those concerned to make sound decisions? To answer these questions affirmatively, the tests used to collect the information should be prepared in such a way that errors are minimized, content validity is maintained, principles and suggestions of test development are carefully observed, and purpose are well defined. If these conditions are not met, the whole endeavor we exert in the educational process will be doomed to failure. This could be one of the most principal factors that jeopardizes the quality of education. Graduates who pass through weak systems of evaluation would never demonstrate their potentials during learning and develop skills, attitudes, and knowledge which are consequential in the world of work. If the tests do not demand them to practically exhibit their potentials and challenge them to master the subjects they learn, they will never be effective citizens who can contribute to their country's development as stipulated in the Education Training Policy of Ethiopia (1994).

This study raised some basic issues concerning the quality of tests prepared by teachers. The results of the test items analysis showed, however, serious pitfalls in a number of expected practical considerations. As indicated in the analysis part, it was found that the test papers had the following major problems.

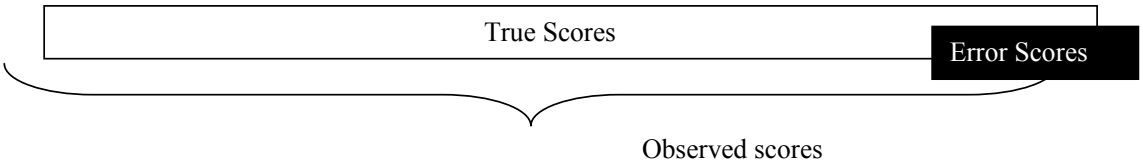
1. Technical errors
    a. Item writing principle-related problems
    b. General item writing consideration problems
    c. Content validly problem
2. Skill-related problems
3. Negligence problems
4. Language competence problems

Test experts and measurement specialists have formulated important principles and suggestions that could guide teachers in the development of test items of various formats. The ultimate goal of those principles and suggestions is to produce items of high quality that measure the desired learning outcomes on the bases of the learned materials. They benefit the teachers to minimize unintentional errors or irrelevant clues that would enable uninformed or unprepared students from getting the correct answer and to avoid any obstructions for those informed students not to miss the right answer. In other words, teachers should avoid any clue that could enable students to get the right answer. Students should not get the right answer if they did not know the subject well. This would lead us to erroneous decisions about the performance of students. If, for instance, a student scores 50% of a test simply by the help of clues in a test, we will make a value judgment that the student has an average "knowledge" on the subject tested. This is an invalid decision that depends on data that are not objective. If, on the other hand, we prepare ill-defined items that are characterized by vagueness and lack of clarity, students will miss the correct answers even if they may have a good deal of knowledge on the subject. In this case their failure could not be attributed to their lack of knowledge but to the quality of the items used to tap their academic behavior. In both cases the nature of the tests could affect the performance of students. Any score students get is a combination of two factors. They are results obtained through the knowledge students

have which are called true scores and those scores students get by guessing, test-wiseness, clues, or miss because of the confusing nature of the items, which are called error scores. This can be put conceptually as follows:

Students' Obtained Scores = True Scores + Error Scores

Graphically, this can be illustrated in the following way.



If a student scores 45 out of 50, can you be dare to say he/she knows all those items correctly answered, i.e., the 45 points? Can you say the student missed the remaining 5 items because he doesn't know the ideas presented in those items? Definitely you can't! Why there are a number of factors that contaminate the student's achievement which we call sources of errors? Our effort in education is to minimize these errors through the use of carefully crafted items and controlled test administration sessions. Cheating, guessing, poorly constructed items, test-wiseness, low content validity tests, and many other factors account for error scores. In this case one of the responsibilities of teachers is developing quality items that measure students' learning competence and their achievement behaviors. No matter how carefully the tests are administered, no matter what the students know about the subject matter, and other examinee-related factors are controlled, if the test items are not as quality as they are expected to be, the intended learning outcomes will not be successfully measured. This evidences that evaluation is at the heart of the whole educational system.

From the analysis of the tests, it looks that the teachers seem to have a number of problems: lack of commitment to the profession, carelessness, lack of understanding of the subject matters they teach, lack of competence in item writing, and serious language problem. Almost in all tests there were a number of errors and mistakes, some of which could be avoided through careful editing of the tests. The teachers did not use items that

could be used to assess higher order learning outcomes. This could be accounted for by their lack of understanding of the subject matter or lack of skill the construction of items or, at worst, due to leniency error. When the teachers feel that their students should not fail (without understanding the subjects they teach) they develop items that are simple. The goal of these types of tests is to promote students from one year or grade level to the next.

The tests analyzed so far were poor in terms of measuring various learning outcomes. In other words, the total emphasis of the tests was on rote memorization or simple knowledge of facts, names, places, figures and conditions. This, however, will not help us reach our educational goals as stated by the Education and Training Policy (1994). The Policy envisioned that our educational system would be geared towards producing problem solver, creative, innovative, competent and critical thinker citizens. In order to meet these national objectives in a practical manner, the evaluation system should be made comprehensive enough to cover a wide area of cognitive, affective and psychomotor domains. If we focus on the cognitive domain which is measured by teachers using paper-and-pencil test, it has six levels. The simplest is the knowledge level which refers to rote memorization of facts and figures that depend mainly on surface level learning or maintenance rehearsal. Knowledge acquired through this procedure "evaporates" before it reaches the long term memory center of the students. The highest and complex level of the cognitive domain is evaluation which focuses on enabling students to develop deep level of understanding using elaborative rehearsal. The relationships of the levels are indicated in Figure 4.

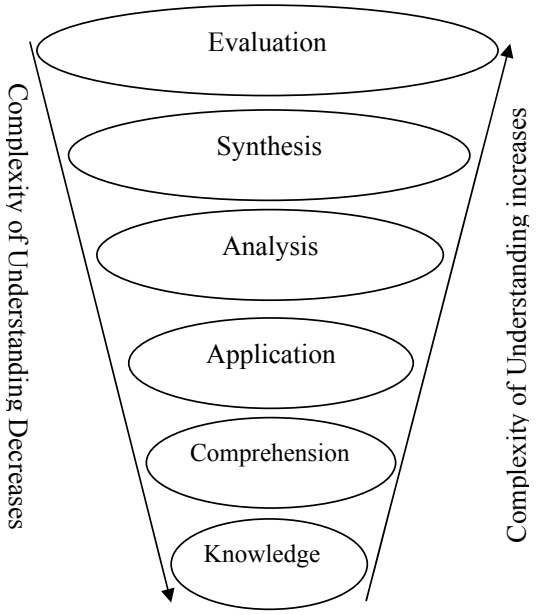


***Figure 4**: The Levels and Complexity of Cognitive Domain.*

As repeatedly stated in the previous sections, according to Ebel and Frisbie (1991) and Capper (1996), the major function of teacher-made test is to measure students' achievement and to contribute to the evaluation of their educational progress and attainments. The second function is to motivate and direct student learning.

As teacher-made tests are the sole means of measuring instructional outcomes in our school systems, where standardized tests are never employed, they provide data to make decisions on important learning outcomes.

Hence, if tests are poorly constructed, the information they provide lead teachers to wrong judgment about their students' academic achievement. The validity and reliability of the information that the tests provide depend on the quality of the tests, which in turn, depend on principles underlying their construction and use. Regarding this point, measurement experts like Cole and Chan (1994) and Ebel and Frisbie (1991) suggested that teacher-made tests lead to effective teaching if they are carefully crafted in a positive

and constructive way following the principles of test construction and if teachers are knowledgeable about and skilled in the use of educational tests. Tests have an impact on what is taught and learned in classrooms, especially, when the results of the tests are used to make important decisions.

Unfortunately, however, educators and teachers in schools and higher learning institutions give less attention to this issue. In relation to this point, Mehrens and Lehman (1984:189) stated, "one of the most common mistakes of teachers is that they do not check the effectiveness of tests." Sometimes students fail not only due to lack of ability but as a result of the poor quality of the tests themselves. The quality of tests can be improved by using general and specific principles (suggestions) of test construction (Gronlund, 1972).

From the above figure (Figure 4), we see that the level and complexity of understanding increases from the knowledge level to evaluation. As we have seen in the results' section, almost 100% of the items of all test measure the simplest level of learning outcomes as it is easy to construct and correct items of this nature. But can we achieve the goals we set in the Educational and Training Policy through these types of items? Unfortunately not! The tests are poor not only in terms of their emphasis on the lowest level of cognitive domain, but they also lack content validity. The tests cover restricted areas which did not represent the courses they are meant to measure. In this case the decisions teachers make not only affect the students (negatively and positively) but they endanger the quality of education in a number of ways. As indicated earlier, in Figure 2, educational process involves four fundamental elements that are interrelated to each other and impact the major part of the quality of education. The interactions of these elements could be best represented as in Figure 5. Objectives are the expected outcomes we wish to attain at the end of the course delivery. The objectives could be cognitive, affective, or psychomotor which focus on a diverse nature of behavioral outcomes that we anticipate from our students. Unfortunately, however, the tests focus 100% on the cognitive aspect of learning. The other element is the teaching process through which teachers assist or support their students to get the most out of those courses taught to them. The contents

are the vehicles that help reach the objectives, and evaluation is a means to ascertain the effectiveness of the teaching process and the attainment of the objectives. In this case the quality of the tests (assessment techniques or evaluation) developed by teachers is a determining factor that either positively or negatively impinges on quality of education.

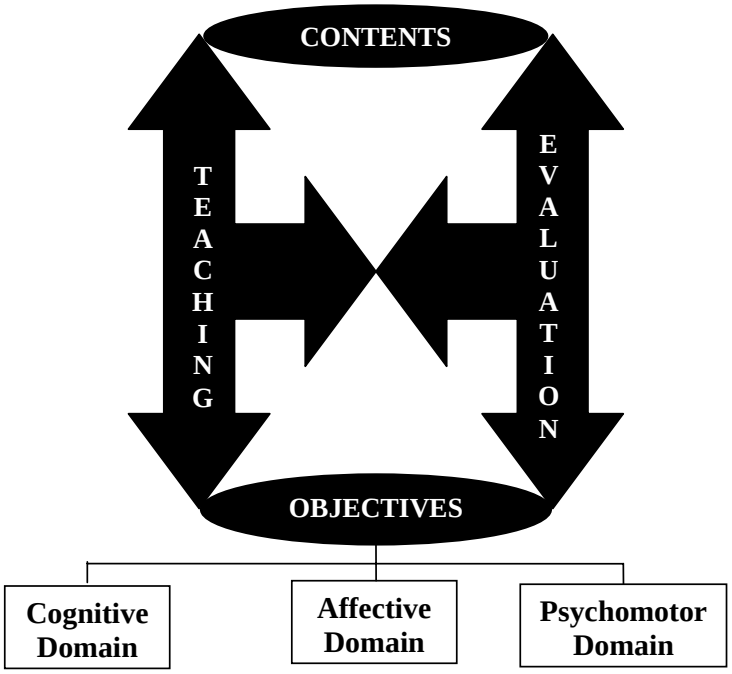


Figure 5*: The Interactions among the Elements of Educational Process.*

The implication of this study is, therefore, the quality of the tests directly affects the quality of education. If teachers have no competence to measure their students using properly crafted tests, the data they gather will not represent the intended learning outcomes, as a result the decisions they make about their students will not be valid and reliable. Hence, serious attention should be paid to this grave situation by various education officials to design a way to overcome the problem through improving teachers' language competence, their test development skills, and raising their motivation as well as enhancing their commitment to their profession. Most importantly, broader scope of research should be carried out to examine the extent to which the poorness of the tests affects the quality of education as a whole.

References

Capper, J. (1996). *Testing to learn, learning to test: Improving educational testing in developing countries.* International Reading Association and the Academy for Educational Development.

Carey, L.M. (1994). *Measuring and evaluating school learning*. (2nd ed.). Boston: Allyn and Bacon.

Cheng, Y.C. (1996). *School effectiveness and school-based management: A mechanism for development*. London: The Falmer press.

Cohen, R.J. and Swerdlick, M.E. (1999). *Psychological testing and assessment: An introduction to tests and measurement* (4th ed.). Mountain view: May Field publishing company.

Cole, P.G. and Chan. L.K.S. (1994). *Teaching principles and practice* (2nd ed.). New York: Prentice Hall.

Crowl, T.K., Kaminsky, S., and Podell, D.M. (1997*). Educational psychology: Windows on teaching.* Madison: Brown and Benchmark publishers.

Cruickshank, D.R., Bainer, D.L., and Metcalf, K.K. (1995). *The act of teaching*. New York: McGraw-Hill, Inc.

Ebel,R., and Frisbie, D.A. (1991). *Essentials of educational measurement* (5th ed.). London: Prentice-Hall, Inc.

Eggen, P. and Kauchak, D. (1997). *Educational psychology: Windows on classroom.* (3rd ed.). New Jersy: Prentice-Hall, Inc.

Elliott, S.N., Kratochwill, T.R., Cook, J.L., and Travers, J.F. (2000). *Educational psychology: Effective teaching, effective learning* (3rd ed.). Boston: McGraw Hill.

Ferrara, D. L. (2007).The school improvement and transformation system. *Educational Planning*. 16(1), 18 – 30.

Gronlund, N.E. (1971). *Measurement and evaluation in teaching* (2nd ed.). New York: Macmillan Company.

Gronlund, N.E. (1982). *Constructing achievement tests* (3rd ed.). London: Prentice Hall International, Inc.

Hounsell, D. (2003). The evaluation of teaching. In H. Fry, S. Ketteridge and S. Marshall. (2003). *A handbook for teaching and learning in higher education: Enhancing academic practice* (2nd ed.) pp. 200 – 212. London: RoutledgeFalmer, Tyler & Francis Group.

Kubiszyn, T. and Borich, G. (2003). *Educational testing and measurement: Classroom application and practice* (7th ed.). New York: John Willy and Sons,Inc.

Linn, R.L. and Gronlund, N.E. (2000). *Measurement and evaluation in teaching* (4th ed.). New York: Macmillan publishing.

McAllister, G., and Alexander, S. (2003). Key aspects of teaching and learning in information and computer sciences. In H. Fry, S. Ketteridge and S. Marshall. (2003). *A handbook for teaching and learning in higher education: Enhancing academic practice* (2nd ed.) pp. 278 – 300. London: RoutledgeFalmer, Tyler & Francis Group.

McCormik, C.B. and Pressley, M. (1997). *Educational psychology*: *Learning, instruction, assessment*. New York: Longman.

McDaniel, E. (1994). *Understanding educational measurement*. Madison: Brown communications, Inc.

McKimm, J. (2003). Assuring quality and standards in teaching. In H. Fry, S. Ketteridge and S. Marshall. (2003). *A handbook for teaching and learning in higher education: Enhancing academic practice* (2nd ed.) pp. 182 – 199. London: RoutledgeFalmer, Tyler & Francis Group.

Mehrens, W., and Lehman (1984).*Measurement and evaluation in education and psychology* (2nd ed.). Japan: Holt Rinehart and Winston, Inc.

Muijs, D.and Reynolds, D. (2005). *Effective teaching: Evidence and practice*. London: Sage publications.

National Center for Education Statistics (NCES) (December 2007). *Highlights from PISA 2006: Performance of U.S. 15-year-old students in science and mathematics literacy in an international context*. U. S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. NCES 2008-016.

National Center for Education Statistics (NCES). (June 2005). *The condition of education 2005.* U.S. Department of Education. Institute of Education Sciences., NCES 2005–094.

National Center for Education Statistics (NCES). (December, 2000). *Monitoring school quality: An indicators report. Statistical analysis report.* U.S. Department of Education. Office of Educational Research and Improvement, NCES 2001–030.

Nitko, A.J. (1996). *Educational assessment of students* (2nd ed.). New Jersey: prentice-Hall, Inc.

Nitko, A.J. (2004). *Educational assessment of students* (2nd ed.). Upper saddle River: Pearson Education, Inc.

Plake, B.S. and Impara, J.C. (1997). Teacher assessment literacy: What teachers know about assessment? In G.D.Phye (Ed.), *Handbook of classroom assessment: learning, adjustment, and achievement* (pp.55-67). San Diego: Academic Press.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F.. (1998). *Teachers, schools and academic achievement.* Paper presented at the Association for Public Policy Analysis and Management, New York City.

Sax, G. and Newton, J.W. (1997). *Principles of educational and psychological measurement and evaluation* (4th ed.). Belmont, CA: Wadsworth publishing company.

Stapleton, M. (2001). *Psychology: in practice education*. Sunderland: Hodder and Stoughton (a member of the Hodder Headline Group).

Transitional Government of Ethiopia*. (1994). Education and training policy of Ethiopia*. Addis Ababa: EMPDR.

UNICEF. (n.d). *Quality of primary education: The potential to transform into a single generation.*

Wakeford, R. (2003). Principles of student assessment. In H. Fry, S. Ketteridge and S. Marshall. (2003). *A handbook for teaching and learning in higher education: Enhancing academic practice*. (2nd ed.) pp. 42 – 61. London: RoutledgeFalmer, Tyler & Francis Group.

The World Book Encyclopedia. (1985). *Education*. 6, 59-72. Chicago: World Book, Inc.

Worthen, B.R., White, K.R., Fan, X., and Sudweeks, R.R. (1999). *Measurement and assessment in schools* (2nd ed.). New York: Addison Wesley Longman, Inc.