AJOLD

African Journal of Leadership and Development

2019, Vol. 4 No. 1, 68-79

DOI: https://doi.org/10.20372/ajold.2019.4.1.5

The Effect of Teachers' Classroom Test Writing Practices on Students' Perceived Ratings of Continuous Assessment and College Cumulative Grade Point Averages: The Case of Ethiopian Civil Service and Adama Science and Technology Universities

Desalegn Sherkabu[1]

Abstract

This paper aims at investigating classroom test Item Writing Practices and challenges of evaluation in the Ethiopian Civil Service University (ECSU) and Adama Science and Technology University (ASTU) with their 370 and 172 sampled university students and instructors respectively. A concurrent quantitative qualitative (QUAN-QUAL) mixed descriptive design was employed. While the universities and faculties and/or departments were chosen by using purposive bases of relative similitude, the sixteen academic leaders and the 13 exam booklets were sampled on accessibility basis. Thus, the study was based on purposive sampling. Data were collected using two sets of questionnaire for the former two sets of respondents, and FGD and interviewees were arranged with the other two groups of respondents. In the questionnaire, the researcher designed a five point Likert scale with 49 items under 11 categories for student respondents, and a five point Likert scale with 61 items under 6 major categories for instructor respondents. Sample examination booklets were closely studied qualitatively. The quantitative data were analyzed using multiple linear regression technique. The study depicted a moderately high rate of standard- about 65.2-80% on the five point scale questionnaire filled in by students. It also revealed (r=0.47), a moderately positive relationship pointing out that a unit of change in college entrance examination (ESLCE) resulted in forty-seven units of change in college CGPA which is direct substantiation of the trend to be in its normal course. Students rated ethical standards as low which verifies the qualitative evidence to the sagging quality of the assessment, contrary to the quantitative data. Accordingly, the assessment and evaluation practices demonstrated both positive achievements and deficiencies. Finally, recommendations were made including a call for a firm follow-up of the process.

Key Words: assessment, evaluation, testing, higher education institutions

1. Introduction

There is almost every reason for studying the evaluation practices of HEI'S as one of the major activities to be accomplished. There has been a growing frustration that workplace-competences of graduates are getting lower and lower (Mehrens and Irwin, 1991; 2000). There are also growing concerns from such stake holders as the MOE, instructors and students about problems

[1] Lecturer, Ethiopian Civil Service University, Email:desalegnsherkabu08@gmail.com

ISSN 2519-5255(print) ISSN 2957-9104(online)

of cheating at examinations and plagiarism. In addition, issues of high concern are irregularities in student grades, while it is also known that there are always high achievers, averages and low achievers in every group of students. A bird's eye view of previous semesters' grade reports elsewhere for some courses; however, depicted that all A^- and above grades; all grades less or equal to C^+, etc. in the sampled universities were hard to swallow.

The assessment and evaluation practice of any educational institution should not be left unattended. Thus, what the practice looks like, how much the teaching learning process are and the assessment and evaluation process influencing each other need to be studied closely to pin point some ways forward out of the as is. To this end, the researcher set the following four basic questions.(i) How much of the nature of the evaluation in the HEI's under study up to standard? (ii) How much is the practice contributing to the success of the teaching-learning process?" (iii) What does the ethics look like? (iv) How can the existing practice be further improved? This paper aims at investigating the challenges that higher HEIs in Ethiopia today have in evaluating their students, and finding out the problems that the students have in relation to the evaluation.

This study is supposed to have both theoretical and practical importance. The theoretical importance is that the findings of the study might draw the attention of the instructors themselves and/or other researchers to conduct further in-depth action research in the area to bring about immense changes in the quality of the educational programs. The practical advantage of this study for students, teachers and educators is that the study would reveal the extent of the challenges with evaluation practices so that corrective measures are taken step by step and each stakeholder contributes to and hence benefits from the improvement he/she aspires to have.

This study would have been better reliable had there been more HIE's in it. Nevertheless, the writer has limited the study to two HEI's for the sake of feasibility with respect to time and resources. Evaluation is a challenge to all the conventional, distance, and evening educational programs. Nevertheless, the findings of this study were directly attributed to evaluating the assessment and/or evaluation practice of mainly the conventional programs in the two HEI's though this may have implications to all other HEI's in Ethiopia. The data from the HEI's were brought to comparison even though classroom teacher made tests do have their own problem of subjectivity. The universities might somehow differ from each other in such matters as the nature of students they admitted in terms of their age and job experience. Besides, the students' examination booklets and their parallel answer-keys was found very disorganized, and hence, the researcher refrained from insisting on getting this issue at least for the present study with the intention of leaving that part of the study to be carried out by the instructors of the particular departments. Therefore, the present study was not meant to address the details about the actual exam and test items such as item analyses and item difficulty levels.

This study followed the following principles-driven conceptual framework based on the works of Ornstein (1995). According to this model, the fulfillment of the conditions listed in the column to the left would automatically lead to the fulfillment of the outcomes listed in the column to the right.

2. Review of Related Literature

Assessment, measurement, testing and evaluation are seemingly very much interrelated terms. Assessment and evaluation can be explained as having same purpose but to take place at different stages of a given measurement of educational process as depicted below.

> *Assessment is a process by which information is obtained relative to some known objective or goal. .. Assessment would be a review of journal entries, written work, presentation, research papers, essays, story writing, tests, exams etc. and will demonstrate a sense of more permanent learning and clearer picture of a student's ability (Kizlik, 2012,PP.1-2).*

While evaluation refers to the process of delineating, obtaining, and providing useful information for judging decision alternatives, some prefer the term assessment to this process, while others emphasize the aforementioned distinctions. These terms are certainly connected, but it is useful to think of them as separate but connected ideas and processes.

Measurement and evaluation help students in communicating teachers' goals, increasing motivation, encouraging good study habits, and providing feedback that identifies strength and awareness (Mehrins and Lehmann, 1991;Ward and Mildred,1999). While, Guskey suggested an alternative to use grades as negative reinforces. Critics contend that by encouraging teachers for years to breakdown learning into "factoid" and then to test those "factoids" the result was to deemphasize teaching or assessment aimed at high-order thinking (Herman, Pamela, and Lynn, 1992; Neil 1992). What performance assessment basically refers to has been given by Ornstein as indicated below.

> *...The idea behind performance assessment is that is students are supposed to conduct a scientific experiment, then have them to do it and assess them while doing it…Moreover, that demonstration should be in a performance context, dealing with the process and operation not simply the name or definition of a fact or concept, and there should be an intended outcome toward which the student plans, organizes, and works from start to finish (Ornstein, 1995, PP.49-61).*

Nevertheless, performance assessment, with all its merits mentioned earlier is not without demerits. Therefore, teachers might need to change their way of teaching: moving from specific time *blocks to flexible scheduling, from focusing on how well students do the first time to how well they eventually do, and from less individual learning to more cooperative learning (*Spady, 1992) and Ornstein 1995).

In brief, the details of these two decision bases have been discussed by the next few principles. The two most common criteria for choosing tests are test validity and reliability. Test validity refers to the extent to which a test has served the purpose it was supposed to serve. Test validity can take the form of content, curricular, construct, criterion or predictive. According to Ornstein (1995), of all the forms of validity, content validity is perhaps the most important one. Test content validity refers to the extent to which the test measures what it is intended to measure.

Test reliability refers to the quality of the test to yield similar results when it is repeated over a short period of time or when a different form is used. A reliable test can be viewed as consistent, dependable, and stable. Test reliability can be expressed numerically–in terms of coefficient of correlation. A reliability coefficient is often the statistics of choice to test consistency among different administrations. A coefficient of .80 or higher indicates high reliability, .40 to .79 fair reliability, and less than .40 low reliability. A respective coefficient of correlation of 1.00 and -1.00 represents complete/perfect positive or perfect negative relationships.

Test coverage /sufficiency refers to how large should test items refer to test coverage /sufficiency. It answers the question, “can the classroom teacher check the knowledge aspects that students have difficult in mastering them point by point?” An instructor who sets sufficiently large number of questions that are representative samples would normally achieve better result. A good test has to be set in such a way that test items are fairly distributed to the three knowledge domains namely, KSA- knowledge, skill and attitude domains. Test relevance asserts that teaching and assessment and/or evaluation have to be one - the immediate reflection of each other.

Another important criterion for a good test is that it must take into account the language, and other cultural elements of the learner. Test variety refers to the need for making use of different test types and test items. Not all teachers, like any person, are good at every skill. Some are good at oral presentations of contents, while others are better successful in written tests. Some have an exceptional performance in analyzing concepts, while others prefer to produce something, instead. Hence, instructors have to avoid a hollow-effect syndrome in the course of assessing and/or evaluating students. Test and examination items that are put in item banks are those which are found to be with best discrimination power (DP) on the one hand, and proper level of difficulty (LD) on the other hand.

Classifications of teacher made tests can be made based on different criteria. For example, Ornstein in his book titled “Strategies for Effective Teaching,” states that tests can be of Short-Answer or Essay type, each of which has its own merits and demerits.

The following points by Mehrins and Lehman include points of a checklist that a teacher should consider when preparing classroom tests.

> *What is the purpose of the test; why am I giving it? What skills, knowledge, attitudes, and so on, do I want to measure? …How are the test scores to be tabulated? How are scores (grades, or level of competency) to be assigned? How are the test results to be reported? (Mehrens and Lehmann, 1991, PP.87-119)*

Test classification by stimulus materials depends on the nature of the course verbal picture, tools etc.; open-book versus closed –book examinations. Good tests do not just happen. They require adequate and extensive planning so that the instructional objectives, the teaching strategy to be employed, the textual material, and the evaluating procedure are related in some meaningful fashion. Inadequate planning is, however, one of the most common errors teachers commit in preparing teacher-made tests.

Developing test specification is different from table of specification. It refers to answering such questions as why do I test and what is the best way to do what I want to do. A table of specification refers to what is to be tested. It’s content and objectives in a matrix, item difficulty, and when to test- frequency is referred to here.

Principles for Writing Short-Answer Tests: Short-answer items include multiple choice, matching, completion, and true-false. The following 14 by *Ornstein* include principles to be considered when preparing and writing short-answer test.

> *… Writers of short-answer test/test items should (1) measure all the important objectives and outcomes; reflect the approximate emphasis given the various objectives and content of the subject or course… (14) not to be the only basis for evaluating the students’ classroom performance or for deriving a grade for a subject…(Ornstein 1995, PP.111-114).*

Principles for writing Multiple Choice Questions*:* The basic form of the *multiple-choice test* item is a stem or lead, which defines the problem, to be completed by one of a number of alternatives or choices. There should be only one correct response, and the other alternatives should be plausible but incorrect. For this reason, the incorrect *alternatives* are sometimes referred to as "distracters." In most cases, three to four alternatives are given along with the correct item. The effect of guessing in multiple choice test item is reduced, but not totally eliminated, by increasing the number of alternatives. The following 14 by Mehrins and Lehman include among suggestions for writing multiple choice items.

> *… Writers of multiple choice test/test items should (1) state the central problem in the stem… (14). use the alternatives "All of the above" and "None of the above", only sparsely (Mehrins and Lehman, 1995,PP.129-149).*

Principles for Writing Matching Questions: In a matching test, there are usually two columns of items. For each item in one column, the student is required to select a correct (or matching) item. The items may be names, terms, places, phrases, quotations, statements, or events. The basis for choosing must be carefully explained in the *directions*. The following 10 by Mehrins and Lehman include among suggestions for writing matching test items.

> *… Writers of matching tests/test items should (1) use directions which briefly and clearly indicate the basis for matching,… (9) avoid Negative statements in either column (10) Many multiple-choice questions can be converted to matching test; therefore, many of the suggestions are applicable to both… (Ornstein, 1995,PP.115-118).*

Principles for Writing Completion Questions: In the completion test, sentences are presented from which certain words have been omitted. The student is to fill in the blank to complete the meaning. This type of short-answer question, sometimes called a fill-in or fill-in-the-blank question is suitable for measuring a wide variety of content. The following 10 by *Ornstein* include among suggestions for writing Completion test items.

> *Writers of completion-test/test items should (1) inform students all the details about the direction…, (12) When combining multiple-choice and completion formats, use homogeneous alternative responses in form, length, and grammar …(Ornstein, 1995,PP.107-114).*

Principles for Writing True-False Questions: *The true-false question* is the most controversial of all types of short-answer questions that are used in education. The following 12 by *Ornstein* include among suggestions for writing true-false questions.

> *Writers of true-false tests/test items should (1) test an important concept or piece of information, not just a specific date or name... (10) use simple grammatical structure. 11) be clear and concise. (12) place the idea being tested at the end of the statement… (Ornstein, 1995, PP. 119).*

According to Tuckman (1991), the essay is considered to be the most authentic type of testing for, among others, college students, and is, perhaps, the best one for “measuring higher mental process.” Ornstein 1995 strengthens the same to say to learn how a student thinks, attacks a problem, writes, and utilizes *cognitive* resources, i.e. something beyond the short-answer test such as essay test is needed- especially this is a must where there is no specific right answer. Scholars of testing identified three types of essay test items.

The following 14 by Ornstein are **suggestions for writing essay questions**.

> *Writers of Essay tests/test items should (1) Make directions specific …, and (14) Write comments on the test paper for the student…(Ornstein, 1995,PP.87-106).*

The following include additional principles for writing Good Classroom Tests.
According to Tuckman (1993), one does not want to reward test-taking skills as a substitute for acquiring knowledge through hard work such as coming to class *and* studying.

Both administration and post exam-administration activities worth considering. It is recommended that teachers announce tests well in advance. Conditions other than students’ knowledge can affect students’ performance in tests. Tests should be returned to students as quickly as possible. As the papers are returned, the teacher should make some general comments to the class about awareness of the group effort, level of achievement, and general *problems* or specific areas of the test that gave students trouble. Each question on the test should be discussed in class, with particular detail given to questions that many students missed.
Grading has the following purposes in general: (1) certification, or assurance that a student has mastered specific content or achieved a certain level of accomplishment; (2) selection, or identifying or grouping students for certain educational path or programs; (3) direction or providing information for diagnosis and planning; (4) motivation, or emphasizing specific material or skills to be learned and helping students to understand and improve their performance; and (5) evaluation, or comparing schools and school districts to establish or justify local state programs and policies.

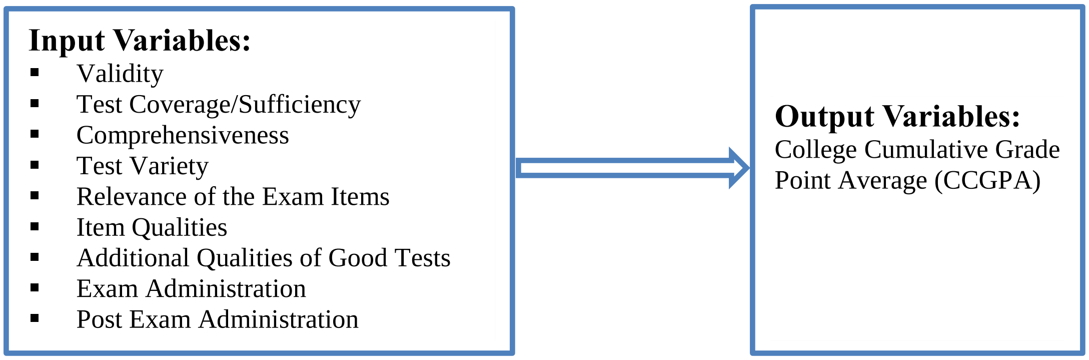
Input Variables:
Validity
Test Coverage/Sufficiency
Comprehensiveness
Test Variety
Relevance of the Exam Items
Item Qualities
Additional Qualities of Good Tests
Exam Administration
Post Exam Administration
Output Variables:
College Cumulative Grade Point Average (CCGPA)

Figure 1: Conceptual Framework
Source: Adapted from Ornstein (1995)

3. Research Methodology

3.1 Research Design

The study was made to use a combined/mixed concurrent qualitative and quantitative descriptive (i.e., explanatory and exploratory) approach (Creswell, 2009).

3.2 Sampling Techniques

It was also conducted through a multiple of sampling techniques such as convenience, stratified random, purposive and accessible sampling techniques where appropriate depending on the nature of the respondents and the researcher. Accordingly, 370 (i.e. a respective number of 280 and 90 student respondents from the respective universities of ASTU and ECSU) were selected to represent a sample framework of 5009 which intern were purposively chosen by the researcher considering the relative program similitude between the institutes, faculties and departments within the two universities using proportionate random sampling or. Similarly, 172 instructors from a 303 sample frame were participated using same formula above. The simple formula next were employed to determine the sample size by taking a confidence level of 0.05.

n = N/ [1+N(e)2] Slovin (1960)

Slovin (1960) as cited in Guidford | (1974) is the commonly accepted sample-size determination formula. Where, n= the desired sample size; N=the sample frame. In this case, the researcher also involved eight instructors (i.e. one from the chosen departments in each of the two HEI's). The researcher sampled out a respective number of four and nine exam booklets from ECSU and ASTU. Sixteen academic leaders were sampled out on accessibility basis.

3.3 Data Gathering Tools

Based on 30 pages review of Basic Concepts of Assessment, Measurement, Testing and Evaluation, the researcher prepared Two-sets of Interviews and Questionnaires and a series of FGD's. The researcher designed a five point Likert scale with forty-nine items for students in the following eleven categories: validity, test coverage/sufficiency, comprehensiveness, test variety, relevance of the examination items general evaluation, additional qualities of good tests, item qualities, examination administration, post examination administration, ethical issues. The researcher also designed a five point Likert scale with sixty-one items for the instructor-respondents in the following six categories: planning and preparation for tests/exams, perception, writing test/exam items, monitoring the examination administration, post examination administration activities, and general evaluation.

3.4 Data Analysis and Discuss Each in Detail

Then, the quantitative data from the two questionnaires were put into SPSS V-21 and arithmetic mean were computed for selected categories of the rating scale and, then, converted into percentages for ease of explanation. Multiple linear regression data analyses technique was employed, to also regress students' college cumulative grade point which is commonly referred to as "the CGPA." against their average college entrance results. An in-depth exploration using a strategy called qualitative matrix analysis was concurrently employed to analyze the verbal responses from students, instructors and department heads. $Y=Ƀ_0X_0+ Ƀ_1X_1+ \ldots Ƀ_nX_n$

4. Results and Discussion

Students who participated in filling in the questionnaires were 370 (i.e. 90 from ECSU and 280 others from ASTU). Thus, the over 11% sample size is well taken for descriptive surveys in most social science studies to which this one is also the case. Considering possible non-returnees, 5% plus questionnaire was distributed to both student and teacher respondents. Consequently, 370 students and 172 instructors filled in and returned the questionnaires. The following is the detailed account of the respondents, interviewees, discussants and documents (examination booklets). On the other hand, of the 172 total teacher respondents who returned the questionnaires with missing 5% contingency.

The codebook in the shaded paragraph below was prepared for the qualitative data analyses to make better referencing of facts for readers. The codes helped the researcher to manually organize and handle the vast qualitative data.

E-FGD-Gn-Xn stands for the respective variable names: University, Focus Group Discussion made, Focus Group Discussion “Group”, the Particular discussant in the FGD, and the last the number given to the particular remark made by the discussant. Thus, E stands for Ethiopian Civil Service University; ASTU stands for Adama Science and Technology University. For example, E-FGD-G1-T refers to respondent named by his/her first name initial T, in FGD with Group I at ECSU quoted for his particular remark labeled as 1st. **SQCN** stands for Student Questionnaire Code Number. For example, SQCN1 refers to “a questionnaire filled by a student given the code number 1 on the top of the front page and centered using hand written ink.” **TQCN** stands for Teacher (Instructor) Questionnaire Code Number. For example, IQCN1 refers to “the questionnaire filled by a teacher given the code number 1 on the top of the front page and centered using hand written ink.” **I-Int** stands for Individual interview

The assessment and evaluation practice is of a moderate standard with a 69.8% variety, relevance, … (See Table 1).

Table 1: Students’ Opinions about Continuous Assessment and Evaluation of their Performance

No.	Items	Rating	%
1	Validity	3.35	67.0
2	Test Coverage/Sufficiency	3.38	67.6
3	Comprehensiveness	3.15	63.0
4	Test Variety	3.50	70.0
5	Relevance of the Exam Items	3.36	67.2
6	Item Qualities	3.21	64.2
7	Additional Qualities of Good Tests	3.24	64.8
8	Exam Administration	3.14	62.8
9	Post Exam Administration	4.38	87.6
10	Ethical Issues	4.98	71.14
11	General Evaluation		
	Overall the nature of the continuous assessment is up to standard.	3.00	60.0
	Overall the nature of the final exam is up to standard	3.24	64.8
	Average	3.49	69.8

Source: Own Survey –June 2016

The assessment and evaluation practice was of a moderately high standard level as per the students' rating; and was rated as high as to the instructors' rating, i.e. 82.6% (see Table 2). So, it can be taken that the assessment is somewhere between moderate and moderately high status when seen as a whole.

Ethical elements were rated as moderately high as 71.14% (See Table 1 below) despite issues of copying, plagiarism, free-readership in group works, etc.(M4, Bi1 & A3); (E-FGD-G1-B2). Thus, the overall picture seems that ethical issues are still formidable especially when the qualitative data from the instructors is incorporated in it.

Table 2: Instructors' Students assessment of their continuous assessment and evaluation

No.	Items	Rating	%
1.	Planning and Preparation for Tests/Exams	3.92	78.4
2.	Perception	3.99	79.8
3.	Writing Test/Exam Items	3.71	74.2
4.	Monitoring the Exam Administration	3.92	78.4
5.	Post Exam Administration Activities	3.99	39.9
6.	General Evaluation		
	Continuous assessments are up to standard	4.72	94.4
	Continuous assessments have positive influence	4.13	82.6
	Final exam are up to standard	4.45	89
	Final exam have positive influence	4.36	87.2
	Average	**4.13**	**82.6**

Source: Own Survey–June 2016

As can be noted from Table 3 next, the independent variables in most case are little significant separately, except few cases like Test Item formatting, followed by item sufficiency, item variety and ethics. However, it is also important to raise a question of why are discrimination power, exam environment and ethics are negative. While the descriptive statistics resulted in Ethical elements were as moderately high, Ethics in the regression table next depicted a slightly negative correlation with Beta coefficient of about -.027.

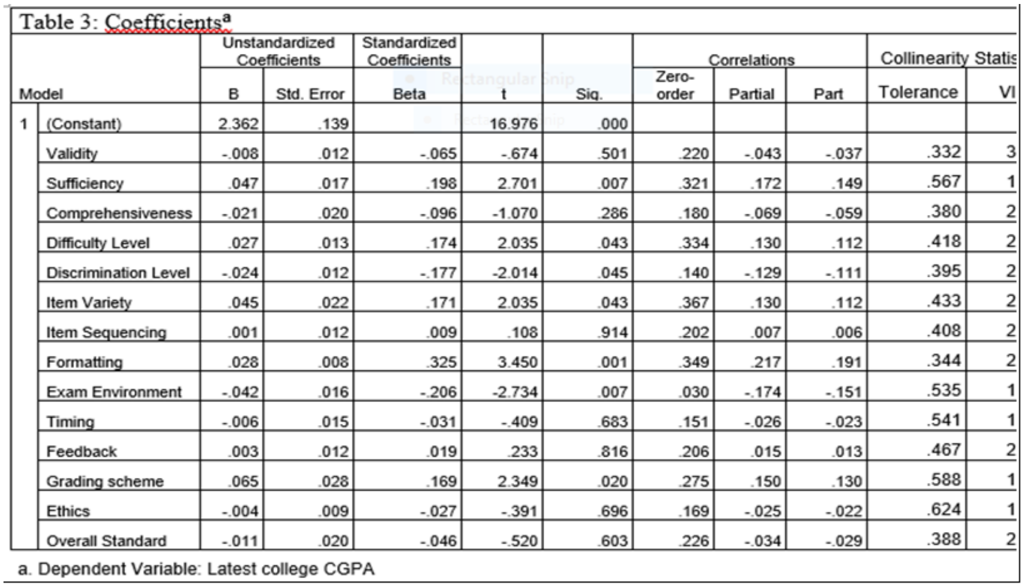

Table 3: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statis	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VI
1	(Constant)	2.362	.139		16.976	.000					
	Validity	-.008	.012	-.065	-.674	.501	.220	-.043	-.037	.332	3
	Sufficiency	.047	.017	.198	2.701	.007	.321	.172	.149	.567	1
	Comprehensiveness	-.021	.020	-.096	-1.070	.286	.180	-.069	-.059	.380	2
	Difficulty Level	.027	.013	.174	2.035	.043	.334	.130	.112	.418	2
	Discrimination Level	-.024	.012	-.177	-2.014	.045	.140	-.129	-.111	.395	2
	Item Variety	.045	.022	.171	2.035	.043	.367	.130	.112	.433	2
	Item Sequencing	.001	.012	.009	.108	.914	.202	.007	.006	.408	2
	Formatting	.028	.008	.325	3.450	.001	.349	.217	.191	.344	2
	Exam Environment	-.042	.016	-.206	-2.734	.007	.030	-.174	-.151	.535	1
	Timing	-.006	.015	-.031	-.409	.683	.151	-.026	-.023	.541	1
	Feedback	.003	.012	.019	.233	.816	.206	.015	.013	.467	2
	Grading scheme	.065	.028	.169	2.349	.020	.275	.150	.130	.588	1
	Ethics	-.004	.009	-.027	-.391	.696	.169	-.025	-.022	.624	1
	Overall Standard	-.011	.020	-.046	-.520	.603	.226	-.034	-.029	.388	2

a. Dependent Variable: Latest college CGPA

$Y=\beta_0X_0+ \beta_1X_1+ \dots\beta_nX_n$

However, the regression model represented by Table 5 next depicted a standard Beta coefficient of ,472, which is ECEE predicted CCGPA, which in turn depicted that the assessment was serving its purpose.

Table 4: Coefficients[a]

		Unstandardized Coefficients	Standardized Coefficients			95.0% Confidence Interval for B		Correlations		
	Model	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part
(Constant)	1.899	.141		13.481	.000	1.622	2.177			
College Entrance result	.481	.057	.472	8.454	.000	.369	.593	.472	.472	.472

a. Dependent Variable: Latest college CGPA

Accordingly, a regression analyses of r=0.47 depicting that ECEE predicted CCGPA, which in turn depicted that the assessment was serving its purpose. The model is significant at .000 level of ά. Courses for the most part were, too, theoretical lacking real-life experiences because of missing design equipment's and computers-(A-Q-St-16);E-Q-St-47) ((A-Q-St-17 (E-Q-St_45). Problems include the nature of influence between the T-L process and the assessment: handouts also meant- "HANDS OUT," and the continuous assessment both influence… (A-I-Int 2).

Less attention was given, and hence, little or no monitoring and supervision activities…almost by all parties-load, experience, lack of pedagogical issues (E-FGD-G1-F3; (E-FGD-G1-B2) E-Q-St-2) (E-Q-St-2) The inattentiveness problems were added with possible copy-paste effects of tests, and examination writing can result in the worst of the scenarios that students reported that they even experienced to have once been sitting for a test which was not meant for themselves. Moreover, the header of the final examination for one of their advanced major area course was depicting that it was rather an examination booklet which was given to some other university students in a previous semester (E-FGD-G2+Y6).

Confounding nature of the questions can be epitomized by "What a bread is made of?" Vs "What wheat is made up of?" Post examination administration points were rated as the lowest - about 58.8%...feedback and options of crosschecking results. Too much of the continuous assessment, for example, when sometimes there are about eight courses in a semester, was deficient-(E-FGD-G1-Bi2), and a single test and final examination mode of assessments were uncommon (E-I-Int-E4). The less creative usage of the 1:5 student cluster formation poses pedagogic deficiency in that it would make classroom interaction too boring, limiting and less natural- (E-FGD-G1-F3; E-Q-St-2). Checking content validity, item level of difficulties and discrimination power were found difficult or impossible. Lack of individual assignments, and poor sense of coordination among instructors teaching in the same classes resulted in havoc. and so did poor editing of final examination papers. - (E-FGD-G2+Y6).

5. Conclusions and Recommendations

5.1. Conclusions

The assessment and evaluation practice taken as a whole was found to be of a fairly below moderate standard leaving some room for improvement endeavors. These include limited attention given to the follow-up which can result from less exposure and, hence, insensitivity to related pedagogical principles, low ethical standards. The less emphasis given to performance evaluation has affected the quality to some extent. Consequently, A considerable number of instructors were involved in writing factoid examination and test items which in turn were supposed to have encouraged students to depend on their handouts only. The handouts themselves were found discouraging students to go for references in the library. The continuous assessment was found to have been positively affecting the teaching learning process and it clearly enhanced student-retention rate. However, malpractices of same continuous assessment in various instances were observed in relation to many of the students and the instructors themselves. The overall process seems to demand firm follow-up from all those who have vested interest in the business of the school: students, teachers, department heads, academic leaders in general and the public at large.

5.2. Recommendations

- The academic leadership should ensure effective implementation of the right pedagogical principles across the departments and programs. It should also ensure that all students have no chance of committing any form of "academic misdemeanor which, in some cases, may go as farther as not implementing the written curricula, for example, by getting tests cancelled through some form of pressure on the teacher.
- Department heads should supervise the aforementioned task/process very carefully and the practice be strengthened by also ensuring sufficiently large individual student-assignments, and enhancing coordination among instructors teaching same classes to consider overloads and overlaps. The academic leadership should ensure that tests as part of a continuous assessment should not weigh more than 20% in order to give space to varieties of options that lead towards a wholistic and more objective evaluation of the students.
- Instructors should ensure that the teaching-learning process be made more skill-based by all means including checking teachers' competence. Instructors should conduct a full-scale and an in-depth exploration of the practice of their assessment and evaluation, especially when seen from validity and reliability perspectives, and the academic leadership, in this regard, has to encourage action research. Instructors among others should be advised to edit tests and final examinations time and again in a way it guides students towards deep learning approach.

References

Arasian, P. (1991), **Classroom Assessment.** New York: McGraw-Hill

Anderson, J.O. (1987 June). Teacher practices in and Attitude towards Student Assessment. ***Annual paper presented on Canadian educational research's association.*** McMaster University, Hamilton, Ontario.

Anderson, L.W. (1981). Assessing Affective Characteristics in the School. Boston Allyn and Bacon.

Creswell, John W. (2009). Research Design Qualitative, Quantitative, And Mixed Methods Approaches. 3rd Ed. Los Angeles. SAGE Publications. Inc.

Desalegn, Sherkabu; Messele Getachew and Temesgen Dagne (2022). "Using Group Work as an Active Learning-Teaching Strategy at College Level: The Case of Social Science Departments of the Ethiopian Civil Service University" from you. Unpublished.

Kizlik, Bob (2012). Measurement, Assessment, and Evaluation in Educationhttp://www.adprima.com/measurement.htm

Mehrens, William A. and Irwin J. Lehmann (1991). Measurement and Evaluation in Education and Psychology.4th ed. Michigan State University. Wadsworth Thomson Leaning

Melkam Zina (2007). "Academic Dishonesty Among Third Year Business Education Faculty Students of Adama University." AAU. M.A. Thesis. Unpublished.

MOE (1994) The New Education and Training Policy. FDRE Addis Ababa

MOE (2015). "Annual Report." …. Unpublished.

Neil, O. (1992). "Putting Performance Assessment to the Test," ***Educational Leadership***, 49, 8. PP.14-19 May 1992

Nolan, Susan B., Thomas, M. Haladyna, and Nancy S. Hass. (1992). "Uses and Abuses of achievement tests," in *Educational Measurement. 72, 1. Phi Delta Kappa International.USA.*

Ornstein, Allan C. (1995) Strategies for effective Teaching. USA

Slovin,M. (1960). In Guidford, J.P. and Fucher, B. (1973). Fundamental Statistics in Psychology and Education. New York, Mc Graw-Hill.

Spady, W. G. (1992a). It's time to take a close look at outcome-based education. Communiqué,. 20(6), 16-18. Spady, W. G. (1992b).

Tuckman, Bruce W. (1991). "Evaluating the Alternative to Multiple-Choice testing for Teachers" in Contemporary Education. Summer

Tuckman, Bruce W. (1993). "The Essay Test: A Look at the Advantages and Disadvantages" NASSP Bulletin