Students' Ratings of Their College English Instructors Before and After the Issue of Grades[*]

Amanuel Gebru[**]

ABSTRACT: *A study was conducted into the reproducibility of first year students' ratings of their College English instructors at Addis Ababa University Social Sciences College . Official Addis Ababa University Teacher Evaluation Forms consisting of 30 close-ended items classified into six evaluative units were administered to 4 randomly selected first year sections of College English 1, before course grades were released and after they were issued i.e. in the late first and early second semesters. Multivariate Analysis of Variance (MANOVA) showed evaluation scores were significantly different in all sections and across all evaluative variables. Considerable differences also occurred between grades expected and grades earned. However, there was no correlation between mean class grades expected or earned and mean class ratings given. The implications of the major findings are discussed as they relate to College English and the Addis Ababa University situation.*

Introduction

Addis Ababa University initiated a Teacher Evaluation Program in 1996 as part of its structural adjustment package. Though teacher evaluation was in use in much earlier days, its reintroduction for earnest use is evidently quite recent. In the period before 1396, Teacher Evaluation Forms were used, but were of limited consequence since they were often administered almost for mere formality. , They raised little faculty concern as they were often used for less crucial administrative decision in pay rise and promotion, but not in the termination and renewal of contracts. With the

[*] Lecturer, Department of Foreign Languages and Literature, Addis Ababa University'

[**] This research was supported by a grant from the Ethiopian American Foundation.

Students' Ratings of Their College English Instructors Before and After the Issue of Grades[*]

Amanuel Gebru[**]

ABSTRACT: *A study was conducted into the reproducibility of first year students' ratings of their College English instructors at Addis Ababa University Social Sciences College . Official Addis Ababa University Teacher Evaluation Forms consisting of 30 close-ended items classified into six evaluative units were administered to 4 randomly selected first year sections of College English 1, before course grades were released and after they were issued i.e. in the late first and early second semesters. Multivariate Analysis of Variance (MANOVA) showed evaluation scores were significantly different in all sections and across all evaluative variables. Considerable differences also occurred between grades expected and grades earned. However, there was no correlation between mean class grades expected or earned and mean class ratings given. The implications of the major findings are discussed as they relate to College English and the Addis Ababa University situation.*

Introduction

Addis Ababa University initiated a Teacher Evaluation Program in 1996 as part of its structural adjustment package. Though teacher evaluation was in use in much earlier days, its reintroduction for earnest use is evidently quite recent. In the period before 1996, Teacher Evaluation Forms were used, but were of limited consequence since they were often administered almost for mere formality. , They raised little faculty concern as they were often used for less crucial administrative decision in pay rise and promotion, but not in the termination and renewal of contracts. With the

[*] Lecturer, Department of Foreign Languages and Literature, Addis Ababa University

[**] This research was supported by a grant from the Ethiopian American Foundation.

reintroduction of evaluation came a revision and resigning of contractual agreements between the University and faculty. In the revised 1996 contractual document for academic staff, which has now taken effect, Article 9 Cancellation and Termination states:

> *The university reserves the right to cancel this contract without any prior notice where there exists good cause for so doing. 'Good cause' under this subarticle includes "The incompetence and/or inefficiency of the employee as evidenced in his/her rating in the system of evaluation of Academic Staff Members currently employed by the university (P.5).*

Half of the evaluation score any faculty member may receive is accounted for by student evaluation.

This reintroduction of Teacher Evaluation for serious administrative purposes triggered a considerable debate among the academic community in several campuses. The reliability of students as sources of evaluative information was called into question and the consequences of their suspect ratings feared. This caused anxiety similar to that produced in the West that student ratings would have an inflationary effect on course grades. Nevertheless, though there is fear that student evaluations may result in inflation of grades student ratings have been in use in a considerable number of colleges and universities world-wide. However, the large body of literature remains inconclusive about their validity and reliability despite an unabated research on the subject since the 1930s. Perceptions about student evaluation are contradictorily "reliable, valid, and useful" and "unreliable, invalid, and useless" (Aleamoni; 1981:88).

A commonly held belief is that there is a positive correlation between earned or expected graders of students and ratings given to instructors. Thus taskmasters and tough graders may expect to receive more negative ratings than lenient graders, irrespective of their instructional effectiveness. For this reason researchers who have come up with the most stringent criticisms of student evaluations suggest that a lecturer need only give generous grades and disburden students to receive generous' returns in student ratings (Overall, Marsh, and Thomas in Howard and Maxwell, 1980). There is also research corroborating the "grades-affect-ratings " theory. Feldman (1976) and Pratt and Pratt (1970) show a significant positive correlation between student evaluations and instructional grades, possibly demonstrating grading behavior affecting evaluative behavior in students.

Further research also supports a bias interpretation of the correlation between grades and evaluations. Holmes (1972), Kennedy (1975), Synder and Clair (1976), Vasta and Sacraminto (1979) and Stumpf and Freedman (1979) found varying levels of causal relationships between grades expected and ratings given.

Despite these findings collectively signifying the contamination effect of grades on evaluation, there is an interpretation that the substantial correlations between grades and ratings are better explained by the Teaching Effectiveness Model which posits that a third variable correlates with both course grades and student ratings. Considering the complexity of the instructional process it seems reasonable to assume that the teaching process is multidimensional and that evaluation instruments should attempt to measure these dimensions, what ever they may be (Ducette and Kenney, 1982:309). As a component of the instructional sphere student motivation has also been considered as a significant variable in student evaluation, with better motivated students likely to give better ratings. The instructor effectiveness model and the student characteristics model imply a causal relationship between course grades and student evaluations. The models encompass several other causal relationships.

- effective instruction produces better learning

- higher learning motivation produces better learning
- better student learning leads to higher course grades and greater student appreciation of instruction (Howard and Maxwell, 1980).

In light of this critical of any monodirectional interpretation of correlations between evaluations and expected or actual grades, Ducette and Kenney (1982) reason in their study of causal connection between grading standards and student evaluations that students tend to have unrealistic grade expectations with over 90% expecting an end of term grade of A or B. They report, however, that they failed to replicate earlier results showing a contamination of grades by ratings.

Much of the literature seems to support the reliability and validity of student testimonials against several individual variables. A review of studies (Muray, 1980) suggests that student evaluations remain significantly unchanged over time, show significant interrater agreement, and are insignificantly affected by a number of variables like grading behavior and classroom observers and teaching supervisors. Student evaluations are causally linked to purported more objective indications of student achievement on standardized tests (Marsh 1984).

A significant correlation has been found between end-of-course rating and ratings obtained from the same students several years after graduation (Overall and Marsh, 1970). In this study, end-of-course ratings and ratings obtained from the same students several years after graduation showed positive correlation (Overall and Marsh, 1979). End-of-term evaluations in 100 courses showed a correlation of .83 with retrospective evaluations. In another cross sectional longitudinal study Firth (1979) reported a correlational agreement in ratings of same students at graduation and several years after graduation. Correlation studies indicate that higher ratings are attributes of rigorous teaching, not generous grading (Marsh 1987; in Webb 1994). However, commenting on the grades-affect ratings hypothesis, Fieldman (in Peterson and Cooper, 1980:683) cautions that "alJ currently available evidence cannot be taken as definitively establishing a bias in teacher evaluation due to the grades students receive or expect to receive in their courses, but neither is it presently possible to rule out such a bias." The contradictions in the area of evaluation may indicate the intricate multidimensionality of evaluation and help to shed light on the wide reliability range reported in the literature from -.75 to .75 (Stumpf , and Freedman, 1979). Also methodological differences may explain some of the substantial differences in the reliability scores reported in the area of teacher evaluation research. In this Ethiopian study which uses the test-retest method, an attempt is made to investigate the effect of received grades on how College English students in Social Sciences College rate their instructor's instructional competence.

Instrument

The forms, which had an internal consistency of .65, were the recently developed Addis Ababa University Teacher Rating Forms (TRF) used by all departments at all levels and in both regular and extension divisions. They consist of 30 items measuring various aspects of instructional effectiveness , with all items rated on a 5-point scale ranging between *very good*(5) and *very poor* (1). Examples could be items 4 and 5 which measure instructors' knowledge of subject matter and preparation for classes respectively. The forms also contain two-open ended items concerned with course evaluation and two about attendance record and expected grade pertaining to the course. For a full reading of the items, see annex.

Subjects

The subjects were first year students (n=115) taking College English, which is a multiseciton course at Addis Ababa

University. They were randomly selected male and female students of four College English sections out of thirty two in the regular program in the first and second semesters of 1997/98 academic year. All were taught by full-time faculty with the rank of lecturer. Eight students who did not evaluate their instructor before the issue of grades were excluded to make possible the pairing of evaluations. Also 12 students who evaluated their instructors in the first treatment, but failed to turn up in the second treatment were excluded from the study.

Variables

In an attempt to measure the specific ratings College English students gave to individual instructional aspects, a taxonomy of teaching behaviors was formulated. Thus, the 30 evaluative items on the Teacher Rating Forms were categorized into six roughly distinct instructional units treated as variables. These were *Intellectual Preparation and Organization* (coded x1), *Presentation and Exposition Skills* (x2) *Management Skills and Affective Factors* (x3) P*rofessional Ethics* (x4), *Assessment Skills* (x5) and *Global Teaching Effectiveness* (x6).

Procedure

The evaluation forms were administered at the end of semester one (before grades were released) and readministered at the start of semester two (after grades were released) in the same academic year. As the course instructors left the rooms, the instructions were read out to the subjects with an emphasis of the need for frank and genuine responses. The subjects were also told that their responses would be used to improve the quality of English Language instruction at Addis Ababa University . In an attempt to justify the repeat evaluations of their first semester instructors, they were further told that the Department of Foreign Languages and Literature was unable to trace their first semester ratings. They were not informed that they were participating in research. In the second administration subjects were asked to indicate grades actually earned. Also as in similar studies, in the last minutes of the evaluation sessions the subjects were orally asked to supply their dates of birth as this would help to pair the evaluations.

The distribution and collection of the Teacher Rating Forms and the supervision of the evaluation was done by department colleagues who volunteered to assist in the research. The evaluations took an average of 20 minutes. The subsequent analysis of the codified evaluations employed a statistical software called **S plus**, and a computer program written specifically for the research. It was found that Statistical Programs for the Social Sciences (SPSS) (Norusis 1990) does not have a facility for Paired MANOVA tests. Likewise the SAS package (Helwig and Council 1979) despite its computational sophistication was found to be lacking in the relevant MANOVA facility.

Operational Definition of Terms

- *Teacher evaluation*: In English Language Teaching, this is a performance review by students which has both developmental and summative functions.

- Reliability: This refers to the reproducibility of ratings given by students on a readministration of evaluation.

- *Academic Responsibility*: The belief that one's academic success or failure is due to the quality of one's effort , not due to external factors like chance or teachers.

Method

A Multivariate Analysis of Variance (MANOVA) is a statistical test employed in research involving many variables. In English teacher evaluation research, the multifaceted nature of evaluative processes demands that measurements should be made on several variables and that the analytical system used should make possible a concurrent examination and analysis of several variables on which evaluation data have been gathered. In this study, MANOVA is used to help bring out relationships between a typology of six instructional variables formed out of 30 evaluative items on the Addis Ababa University evaluation forms. The variables are also paired in line with paired MANOVA principles to indicate the differences in ratings given to each instructor by each student-evaluator before and after the issue of grades. A paired MANOVA test is also conducted to demonstrate the variability/ consensus in the ratings supplied by students.

As concerns the analysis of this paired MANOVA test, the follwing steps were taken and the following analytical procedure was used.

Let X_{1ij} denote the response to treatment 1 (semester one) and X_{2ij} denote the response to treatment 2 (semester two) for the i^{th} variable (I = 1, 2, - - - , p) and j^{th} student (j = 1, 2, - - -, n). That is for the i^{th} variable the pair (X_{1ij}, X_{2ij}) are evaluations of the j^{th} student before and after the student receives his or her grade. Let us define $D_{ij}=X_{1ij}–X_{2ij}$, as the paired difference on the j^{th} unit, which only reflects the differential effect of the grade the student earns on evaluation.

If D_j is the vector of differences for the responses with

$$E(D_j)=\begin{bmatrix} \mu_{1d} \\ \mu_{2d} \\ \cdot \\ \cdot \\ \cdot \\ \mu_{pd} \end{bmatrix} \text{ and } \mathrm{Cov}(D_j)=\Sigma_d=\begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdot\ \cdot\ \cdot & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \cdot\ \cdot\ \cdot & \Sigma_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \Sigma_{p1} & \Sigma_{p2} & \cdot\ \cdot\ \cdot & \Sigma_{pp} \end{bmatrix}$$

and assuming D_1, D_2, - - -, D_n to come from a multivariate normal distribution with a mean vector μ_d and covariance matrix, inference about the mean differences can be based upon Hotelling's T^2 statistic given by

$$T^2 = n(\bar{D} - \mu_d)^T S_d^{-1}(\bar{D} - \mu_d)$$

$$\text{Where, } \bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j \quad \text{and } S_d = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})(D_j - \bar{D})^T$$

That is, given the observed differences:

$$d_j^T = \begin{bmatrix} d_{1j} \\ d_{2j} \\ . \\ . \\ . \\ d_{pj} \end{bmatrix}$$

To test the hypothesis

$$H_o : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

It is known that T^2 is distributed as $[(n-1)/(n-p)]\, F_{p,\, n-p}$ whatever values μ_d and Σ_d assume.

Hence for a multivariate normal population with a mean vector μ_d and covariance matrix Σ_d, the decision rule is given by:

$$\text{Reject } H_0 \text{ if } T^2 \phi \frac{(n-1)}{(n-p)} F_{p,n-p}(\alpha)$$

where $F_{p,\, n-p}(\alpha)$ is the upper $(100\ \alpha)^{th}$ percentile of the F – distribution with p and n-p degrees of freedom (Johnson and Wichern, 1992; Hand and Taylor, 1987).

Results

The analysis is done firstly for each of the instructors evaluated and then a global analysis is presented. Results of the paired MANOVA analysis showed a significant paired difference at $p< 0.05$ in each of the 6 instructional variables.

Table 1: Mean and Variance for Evaluative Data of Instructor A

Variable	Mean difference	Variance
[X1]	0.3628571	0.4569
[X2]	0.3519048	0.5944
[X3]	0.5071429	1.3320
[X4]	1.3338095	15.311
[X5]	0.4423810	1.2757
[X6]	0.5714286	1.9571

For Instructor A the mean difference vector is given in column two of Table 1. The computed value of the t^2 statistic is 0.3299 which is significantly greater than that of the tabulated value at $\alpha=0.05$ which is 0.0967. Table 1 also shows that the smallest mean difference is in variable 2 measuring Presentation and Exposition Skills. In contrast, the biggest mean difference for the same instructor is in the variable x4 which stands for Professional Ethics. What this means is that the evaluation given in the first semester for this item is much greater than the evaluation given after the issue of grades for this same item. Also, the biggest evaluative inconsistency is observed in x4 which measures Professional Ethics whilst the lowest inter-rater variability is seen in item x1 which pertains to Intellectual Preparation and Organisation.

Table 2: Mean and Variance for Instructor B

Variable	Mean difference	Variance
[X1]	0.3365385	0.31745
[X2]	0.4911538	0.46312
[X3]	0.6223077	1.94725
[X4]	0.3080769	0.84494
[X5]	0.5042308	1.02720
[X6]	0.8846154	3.46615

For Instructor B the mean difference vector is presented in column two of Table 2. The computed value of the t^2 statistic is 0.8524 while the tabulated value at $\alpha= 0.05$ is 0.0744, which indicates that there is a significant difference in favor of the tabulated value. The biggest mean difference observed is in item 6 which represents global assessment of teaching effectiveness while the lowest mean difference is observed in item 4 which stands for Professional Ethics. Similarly, the highest variance in the ratings occurred in the area of Global Teaching Effectiveness while the highest consistency is observed in x1 which represents Intellectual Preparation and Organization.

Table 3: Mean and Variance for Instructor C

Variable	Mean difference	Variance
[X1]	0.6094	1.00788
[X2]	0.7458	0.98642
[X3]	0.5208	1.24005
[X4]	0.5694	1.52215
[X5]	0.2917	1.49463
[X6]	0.9167	0.89583

For Instructor C, the mean difference vector is presented in column two of Table 3. The computed value of the t^2 statistic is 2.0421 which is considerably greater than the tabulated value at α=0.05 which is 0.0819. This table also shows that the smallest and biggest mean differences observed are in variable 5 and variable 6 which stand for Assessment Skills and Global Teaching Effectiveness respectively. In addition, the table shows that x4 i.e. Professional Ethics and x6 Global Teaching Effectiveness represent the biggest and smallest degree of consensus in the ratings supplied by students.

Table 4: Mean and Variance for Instructor D

Variable	Mean difference	Variance
[X1]	0.9695833	0.39134
[X2]	0.8466667	0.51698
[X3]	1.3416667	0.76655
[X4]	1.1387500	1.24813
[X5]	0.9112500	0.61168
[X6]	1.0833333	0.86232

For Instructor D, column two of Table 4 presents the mean vector difference. Accordingly, the computed value of the t^2

statistic is 3.8918 which is significantly greater than that of the tabulated value at α=0.05 which is 0.0819 . The smallest mean difference observed is in variable 5 which represents Assessment Skills while the biggest mean difference observed is in variable 3 which stands for Management Skills and Affective Factors. Also, the smallest degree of inter-rater agreement is observed in x4 (Professional Ethics) while the highest inter-judgmental consensus can be seen in x1 representing Intellectual Preparation and Organization.

Table 5: Mean and Variance for All Instructors

Variable	Mean difference	Variance
[X1]	0.5713684	0.58984
[X2]	0.6144211	0.65525
[X3]	0.7531579	1.41264
[X4]	0.8107368	4.33352
[X5]	0.5397895	1.11336
[X6]	0.8736842	1.96260

Overall, column two of Table 5 shows the mean vector difference for all instructors as a group. Thus the computed value of the t^2 statistic is 0.7100 which is significantly greater than the tabulated value of 0.0179 at significance level at $p<0.05$. The biggest mean difference is observed in variable 6 (Global Assessment of Teaching Effectiveness) while the smallest mean difference observed is in variable 5 (Assessment Skills). This indicates that before grades were released the instructors were rated most highly as being globally effective, but the relevant ratings were considerably reduced after grades were issued. In contrast, the fact that the lowest global mean difference is observed in the area of assessment indicates that even before grades were issued the subjects tended to be unhappy with the instructors grading policies. On the other hand, the global variance indicates that there was a high degree of inter-rater disagreement in the area of Professional Ethics as opposed to the area of Intellectual Organisation and Preparation which witnessed the lowest degree of inter judgmental concurrence.

As one can see in Table 6, the majority of the subjects had great grade expectations. In Section A, 43.4% and 30.4 % expected their grade to be A and B respectively. Only 8.6 % expected a C. In Section B, 64 % expected to earn A or B; interestingly none expected a C. In Section C, 30.76%, 38.46% expected their grade would be A and B respectively. Only 11.53% had a modest expectation of a C. Section D had even higher expectations with as high as 42.85% and 33.33 % expecting to earn A and B respectively. Only 9.52 % had an expected grade of a C. However, as this Table and Table 3 (Annex) show there is no systematic relationship among sections between mean class grades expected by students and mean class evaluations assigned to instructors.

Table 6: First Year Students' Expected Grades in College English

	No. of	No. of Students Expecting										Unde-	
	Students	A	%	B		C	%	D	%	F	%	cided	%
A	23	10	43.40	7	30.40	2	8.60					4	17.30
B	25	8	32.00	8	32.00							9	36.00
C	26	8	30.76	10	38.46	3	11.53					5	19.23
D	21	9	42.85	7	33.33	2	9.52					3	14.28

As Table 7 shows, most College English students in the sections covered received a C grade in the course. In section A, only 8.69% and 13.04% received ***A*** and ***B*** although 43.4 % and 32 % had expected to receive A and B respectively. Only 8% expected they would score a C but 69.56 % of the students in the Section received this grade. In Section B, 8%, 16% and 72 % received A, B and C respectively. Whole class expectations were much higher in Section C. 64 % of the respondents expected they would score A or B but none in the section scored A; 11.53% and 84.61 % scored B and C respectively. In Section D, 9.52% and 19.04 % received A and B respectively; 61.90 % received a C, but 42.85% and 33.33 % had expected to obtain A and B respectively. Only 9.25 % expected to earn a C. However, as Table 7 and Table 3 (Annex) indicate, there was no systematic difference among the sections between mean class grades earned and mean class ratings given.

Table 7: First Year Students' Self-reports of Received Grades In College English

Section	Total No. of Raters	A	%	B	%	C	%	D	%
A	23	2	8.69	3	13.04	16	69.56	2	8.69
B	25	2	8.00	4	16.00	18	72	1	4.00
C	26	-		3	11.53	22	84.61	1	3.84
D	21	2	9.52	4	19.04	13	61.90	2	9.52

Discussion

The findings of the present study provide no support for the hypothesis that students as evaluators are stable but suggest instead that students may be significantly influenced by the course grades they receive. A comparison of ratings of students obtained before and after the treatment shows that there is a statistically significant difference across all evaluative variables and with all sections in the study.

These results failed to replicate studies by Muray (1980) and Overall and Marsh and (1979) who found significant short-term and long-term stability of student ratings.

The study seems to corroborate biased interpretations of student ratings reported in the literature. The subjects, with 72.14% of them receiving C and below in accordance with the centralised grading system of College English which allows A and B scores in College English to about 16% of all students. However, students may have perceived that they have been unfairly graded. They had high expectations, with 71% of them expecting that they would score ***A*** or ***B.***

According to Addis Ababa University banding instructor evaluation scores fall into the following evaluation categories.

- 4-5 = good

- 3.5-3.99 = satisfactory
- <3.5 = unsatisfactory

The results of the study show that in administrative terms all the College English instructors have received one banding less in the second administration than in the first, with all of them scoring between 3.86 and 3.50 (Annex Table 3) evidently due to their standards of assignment of grades which is centrally decided by the College English Test Committee in the Department of Foreign Languages and Literature. Instructors giving the course are expected to observe the departmental grading norms.

This study established that there is a significant disparity in ratings before and after the issue of grades and this may be explained by the students Ethiopian School Leaving Certificate Exam[1] scores, their high school[2] self-ratings relative to college conditions, their disappointment with college grades, with over 71% expecting ***A*** or ***B*** in the course.

Thus to the average aspiring student with a dogged determination to score good grades and even to the modestly ambitious, an earned 'C'[3] may be taken as inauspiciously signaling a similarly unsatisfactory grade in the second semester (Part II of the course). Even worse, it may raise associated safety concerns as an average grade in the first year suggests or may be taken as suggesting academic insecurity. Students may experience fears of a compulsory withdrawal or even an academic dismissal.

To the Ethiopian freshman, first year first semester grades may be the most important grades in a student's career at Addis Ababa University since grades in the first year determine which most competitive and marketable field one can study in the years ahead. There is usually a fierce competition to join business studies departments as these usually are believed to lead to an easier employment and a higher pay. Whereas in North America (from where most evaluation research comes) a college student does normally have a wide choice of universities, and a far higher possibility of joining departments of their own first choice, the average Ethiopian first year student may not have a similar opportunity. Subsequently, students may be unable to internalise failure to realise aspirations. Thus they may have to form a self-serving bias i.e. attribute failure to external factors, normally the "tough grader", as is customary with many first year students who experience academic disappointments[4].

Indeed in this context there seems to be a tendency towards attribution of success to internal and failure to external matters. In a study of degrees of assumptions of intellectual and academic responsibility of Ethiopian adults, Belay Hagos (1994) found a significantly lower internal responsibility for failure than for success in both average and above average students. Normally, evaluations obtained in a re-evaluation should be basically similar to those obtained in the first evaluation, i.e., if they are a true reflection of the instructor's teaching competence. This failing, there is a case for the occurrence of a bias due to the less than expected grades received. Indeed in our case the consistently low ratings produced by the repeat evaluations in all sections covered could not have been due to a memory lapse, or a changed perspective but due to grade-related mass discontent explained by the observed discrepancy between expected and received grades of the participants in the study.

The possibility of a contamination in at least a revaluation by student populations that attach utmost importance to grades received thus looks all too evident.

Conclusion

Unlike many North American studies, this study shows that a bias can occur in a read ministered evaluation. As the findings of the study seem to suggest, College English students were not particularly satisfied with the instructors' grading standards.

A global analysis of the evaluative variables would show that the lowest rating received is in the area of student assessment which is true for the instructors as a group and for two instructors individually except for Instructors One and Two who had their lowest means in other areas. This discrepancy appears to be explainable in relation to College English Testing and Assessment. Thus there are some variables which appear to be measuring examination and grading related dissatisfaction. College English testes appear to feel that exam time was not enough; that items taught in class did not appear in exam directly; that instructors did not return tests and assignments (which are challengingly too many for the typical tightly busy College English instructor)[6], and above all that "instructors are not fair in marking". Fairness to a freshman may mean desirable grading leniency on the part of the instructor. To a College English instructor fairness would presumably mean fidelity to marking guidelines and answer keys supplied by the College English Test Committee. It may be for these reasons that the category *Assessment Skills* has produced the lowest evaluative score for all instructors in the study and interestingly both in the first and second evaluations. Hence the demonstrated possibility of colouring occurring in student testimonials of college teaching.

Yet, student evaluations, as measures of teacher competency, remain mandatory within the framework of Total Quality Control despite their significant problems principally dubious credibility. As the findings seem to demonstrate student ratings may be colored by several factors including the personal conditions of the students themselves. For instance, students in the first year in general seem to have unrealistic and unrevised self- assessments and wrong expectations of high grades-attributable to their high scores on the Ethiopian School Leaving Certificate Examination, which may eventually contribute to the dubious credibility of the information they may provide as evaluators of instructors. Nevertheless considerable North American literature supports the reliability of student evaluations, essentially the reproducibility of student ratings over a subsequent administration, which may not be replicated in non-western contexts.

In Ethiopia, owing to unaccustomed college grading realities and a refusal to ascribe low grades to self, students may give less than insincere evaluative information to institutions about the performance of instructors. Hence the need for a re-examination of the assumption that the North American evaluation practice will have the same consequence universally. For the Ethiopian College English lecturer at least there are clear career implications in the present study where in the repeat ratings in all sections there was a decline by one banding - which may have a serious administrative meaning.

A significant practical implication of the findings of the Ethiopian College English Language Instructor evaluation study is that instructors should not be reassigned to a section they previously taught since there are no guarantees that students do not come with evaluative biases against a reassigned instructor. Also, the study may have implications for writing and speaking courses where students may accumulate biases because tests are serially administered and students may guess the letter grades they will receive.

Apart from the negative impact on the professionalism of instructors, the ratings of retaught students may supply unreliable information to the institution on the instructors' quality. Such unreliable evaluative information may often lead institutional decisions to terminate contracts of professional staff. Despite the seriousness of the issue, teacher evaluation programs in ELT are often perceived to be of secondary importance (Murdoch 2000) and consequently remain poorly developed. Teacher evaluation in ELT needs to provide context-sensitive information about sustainability of particular approaches (Roberts and Roberts 1994).

Implications for Further Research

The present study considered the relationship between first year College English students' ratings given as class means at two psychologically significant times. However, the study did not consider the individual relationship between grades expected or earned and ratings given. However further research may be conducted to address the following research questions which the present study has not considered.

- A retrospective link between a College English instructor's grading history and his/her student ratings record viz. the number of A's and B's given and the student ratings received say over a period of four years.

- Whether there is a link between the number of College English A's and B's awarded in different sections and the student ratings in those sections in a given semester.

- Whether open-ended evaluation by College English students would correlate with their responses to close-ended items.

- How differently regular and part-time students rate their instructors of the same course.

- Whether ratings of the same instructor received from senior students would significantly differ from those received from first year students.

- Whether different sections taught by the same instructor would give significantly different ratings.

- Whether gender as a variable plays a significant part in Addis Ababa University student evaluations.

- Whether the reintroduction of teacher evaluation has liberalised grading standards in Addis Ababa University.

Notes

1. Present Ethiopian university students represent the best one percent of the Ethiopian student population which took the Ethiopian School Leaving Certificate Examination (ESLCE). The threshold score for AAU entrance for degree students in recent years has become as high as a GPA of 3.2 on a 4-point scale.

2. It is believed that past success often leads to a high self-esteem and to an expectation of success in other settings (Aronson and arlsmith, 1962). It may therefore be expected that Ethiopian students fresh from high school would expect similar levels of success in college as in high school. Also, research shows that people in several situations overrate their abilities which leads them to unrealistic expectations (Crowne and Marlowe, 1964).

3. If most students in North American universities expect an A or a B, they may be more realistic than their Ethiopian counterparts because the average North American grade today is a B, because of the inflationary consequences of student evaluations (Hocutt, ND). The general similarity between expected and received grades may explain the stability over time of student ratings reported in the literature.

4. Recent studies on Bahir Dar and Kotebe Colleges of Teacher Education students have also shown most students demonstrate a propensity to expect excellent grades. Low grades are attributed generally to external factors principally to biased instructors (Zeleke 1997; Tamire 1997). This may mean that such students harbour grudges against the instructor who graded them.

5. It is a shared knowledge in the college teaching profession that one of the -first questions first year Addis Ababa University students curiously ask about an assigned instructor is whether s/he is a tough or lenient grader which is a testimony to the overriding importance they attach to grades and their singular interest in the grading behaviour of teachers.

6. Instructors in the Department often carry on average an additional load of 10 evening credit hours, which may influence their marking responsibilities and feedback giving behaviour.

References

Addis Ababa University Revised Contract of Employment For Academic Personnel. 1994.AAU.

Aleamoni , L.M (1981), Student Ratings of Instruction In J. Millamn (ed) **Handbook of Teacher Evaluation**, Beverly Hills, (C:Sage)

Aronson, E. and Carlsmith, J.M. 1962. Performance Expectancy as a Determinant of Actual Performance. **Journal of Abnormal Social Psychology**, 65,178-182.

Belay Hagos (1993), Adolescents' Beliefs in the Internal - External Control of their Academic Outcomes and their Academic Standing. Unpublished Paper.

Crowne, D.P., and Marlowe , D. 1964. **The Approval Motive: Studies in Evaluative Dependence**. New York: Wiley.

Ducette, J, and Kenney, J. (1982). Do Grading Standards Affect Student Evaluation of Teaching? Some New Evidence on an Old Question. **Journal of Educational Psychology**. 74(3) 308-314.

Elmore P. and John P. (1974). Effects of Teacher, Student and Class Characteristics on the Evaluations of College Instructors, **Journal of Educational Psychology** 70(2), 187-f192.

Feldman, K. (1976). Grades and College Students' Evaluations of their Courses and Teachers, **Research in Education**, 4, 69-111

Firth, M.1979. Impact of Work Experience on the Validity of Student Evaluations Teaching Effectiveness, **Journal of Educational Psychology**, 71,726-730.

Howard, G.S. & Maxwell, S.F (1980). Correlation Between Student Satisfaction and Grades: A Case of Mistaken Causation? **Journal of Educational Psychology** 72(6),810-820.

Hand, D.J. and Taylor, C.C (1987): Multivariate Analysis of Variance and Repeated Measures: Chapman and Hall.

Helwig, J.T. and Council, K. (eds.) 1979. **SAS User's Guid***e*. SAS Institute, Releigh, North Carolina.

Hocutt, M.O. (No Date). De-grading Student Evaluations?: What's Wrong with Student Polls of Teaching: **Academic Questions***:* 55-64.

Holmes, D.S.1972. Effects of Grades and Disconfirmed Drade Expectancies on Students' Evaluations of their Instructor, **Journal of Educational Psychology**, 63, 130-133.

Howard, G.S. and Maxwell, S.F (1980). Correlation Between Student Satisfaction and Grades: A Case of Mistaken Causation? **Journal of Educational Psychology** 72(6),810-820.

Johnson, A. and Wichern, D. W(1992*):* **Applied Multivariate Statistical Analysis**, 3rd edition, New Jersey.

Kennedy, W.R. (1975). Grades Expected and Grades Received- their Relationship to Student Evaluations of Faculty Performance. **Journal of Educational Psychology**. 65, 109-115.

Marsh, H.W. (1984). Student Evaluation of University Teaching. Dimensionality, Reliability, Validity, Potential Biases and Utility. **Journal of Educational Psychology**, 76,707-754.

_____. (1987). Students' Evaluations of University Teaching: Research Findings, Methodological Issues and Directions for Future Researches. **International Journal of Educational Research**, 11, 253-388.

Murdoch, G. 2000. Introducing a Teacher Supportive Evaluation Program, **ELT Journal**, 54,1,54-64

Murray, H.G. (1980). Teacher Personality Traits and Students' Instructional Ratings in Six Types of University Course. **Journal of Educational Psychology**, 28(2),250-261.

Nurusis, M.J. 1990. **SPSS Base System.** User's *Guide*, Chicago, Illinois.

Overall, J.V. and Herbert W.M. (1980). Student Evaluations of Instruction: A Longitudinal Study of their Reliability. **Journal of Educational Psychology**. 72(3) 321-325.

Peterson, C. and Cooper, S. (1980). Teacher Evaluation by Graded and Ungraded Students **Journal of Educational Psychology**. 72(5) 682-685.

Pratt, M.& Pratt, T.1976. A Study of Student-teacher Grading Interaction Process. **Improving College and University Teaching** 24, 73-81.

Roberts, C. & Roberts, J. 1994. **Evaluation in ELT**. Oxford. Blackwells.

Stumpf, S. and Freedman R.D. (1979). Expected Grade Co-variation with Student Ratings of Instruction: Individual Vs Class Effects. **Journal of Educational Psychology**. 71,293-302.

Synder. C.R . and Clair, M. (1976). Effects of Expected and Obtained Grades on Teacher Evaluation and Attribution of Performance. **Journal of Educational Psychology**. 68,75-82.

Tamire Andualem.1997. Attributions and Academic Achievement of Education, Medicine and Polytechnic Freshman Students in Bahir Dar. **Ethiopian Journal of Education**, 27, 63-78.

Vasta, R. and Sacraminto R.F (1979) . Liberal Grading Improves Evaluation but not Performance. **Journal of Educational Psychology**. 711,207-211.

Webb, G. (1994). **Making the Most of Appraisal. Career and Professional Development Planning for Lecturers**. Kogan Page Limited. London.

Zeleke Demilew. 1997. Trainee Evaluations of Major Area Course Offerings. **Ethiopian Journal of Education**, 27, 42-62.

Annexes

VARIANCE COVARIANCE MATRIX

Table: 1.1 INSTRUCTOR A

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	0.4569014	0.4473093	0.5488036	1.4001886	0.5318879	0.4717857
[X2]	0.4473093	0.5944362	0.6101057	1.0809624	0.5951152	0.6773571
[X3]	0.5488036	0.6101057	1.3320114	2.8064714	0.7211121	0.8122143
[X4]	1.4001886	1.0809624	2.8064714	15.3108948	1.4861605	0.9177143
[X5]	0.5318879	0.5951152	0.7211121	1.4861605	1.2757290	0.6270714
[X6]	0.4717857	0.6773571	0.8122143	0.9177143	0.6270714	1.9571429

Table:1.2 INSTRUCTOR B

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	0.3174475	0.2530922	0.4106723	0.13582508	0.17795923	0.37518462
[X2]	0.2530922	0.4631226	0.4755172	0.35391031	0.22584292	0.35373846
[X3]	0.4106723	0.4755172	1.9472505	0.20892462	0.97023785	2.02067692
[X4]	0.1358251	0.3539103	0.2089246	0.84494415	0.06556846	-0.08423077
[X5]	0.1779592	0.2258429	0.9702378	0.06556846	1.02720138	0.70330769
[X6]	0.3751846	0.3537385	2.0206769	-0.08423077	0.70330769	3.46615385

Table: 1.3 INSTRUCTOR C

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	1.0078783	0.8556957	0.8524217	0.8529609	0.9476652	0.8117391
[X2]	0.8556957	0.9864172	0.9049906	0.9774806	0.9524926	0.6691667
[X3]	0.8524217	0.9049906	1.2400493	1.0662746	0.8073833	0.9523188
[X4]	0.8529609	0.9774806	1.0662746	1.5221520	0.7572922	0.7012681
[X5]	0.9476652	0.9524926	0.8073833	0.7572922	1.4946259	0.8958333
[X6]	0.8117391	0.6691667	0.9523188	0.7012681	0.8958333	1.557910

Table: 1.4 INSTRUCTOR D

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	0.3913433	0.3040638	0.2481920	0.3273647	0.2197614	0.2909058
[X2]	0.3040638	0.5169797	0.4166319	0.4478174	0.3925478	0.4159420
[X3]	0.2481920	0.4166319	0.7665536	0.4130239	0.3002848	0.3555072
[X4]	0.3273647	0.4478174	0.4130239	1.2481332	0.3038016	0.4370652
[X5]	0.2197614	0.3925478	0.3002848	0.3038016	0.6116810	0.3490217
[X6]	0.2909058	0.4159420	0.3555072	0.4370652	0.3490217	0.8623188

Table: 1.5 INSTRUCTORS (ALL)

		[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	[1,]	0.5898375	0.4913949	0.5727978	0.6620937	0.4883152	0.5050683
[X2]	[2,]	0.4913949	0.6552505	0.6256965	0.6707063	0.5364573	0.5349261
[X3]	[3,]	0.5727978	0.6256965	1.4126367	1.0676264	0.7595911	1.0754031
[X4]	[4,]	0.6620937	0.6707063	1.0676264	4.3335218	0.6331938	0.4294558
[X5]	[5,]	0.4883152	0.5364573	0.7595911	0.6331938	1.1133617	0.6486965
[X6]	[6,]	0.5050683	0.5349261	1.0754031	0.4294558	0.6486965	1.9625980

2. INVERSE OF VAR-COVARIANCE MATRIX

Table: 2.1 INSTRUCTOR A

	[X1]	[X2]	[X3]	[X4]	[X5]	[,6]
[X1]	11.4585697	-6.7110776	-0.2962100	-0.42740483	-1.09708847	0.23533347
[X2]	-6.7110776	9.1465268	-0.9974590	0.28332551	-0.72679087	-1.03384704
[X3]	-0.2962100	-0.9974590	2.1680581	-0.26429227	-0.18071650	-0.30129490
[X4]	-0.4274048	0.2833255	-0.2642923	0.12773846	0.02337952	0.04726525
[X5]	-1.0970885	-0.7267909	-0.1807165	0.02337952	1.62623398	0.05898782
[X6]	0.2353335	-1.0338470	-0.3012949	0.04726525	0.05898782	0.89600365

Table: 2.2 INSTRUCTOR B

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	6.28444733	-3.16878130	-0.8766708	0.5176302	0.3353296	0.09876099
[X2]	-3.16878130	5.85736574	-0.5648548	-1.7884777	-0.1305397	0.05754396
[X3]	-0.87667076	-0.56485478	3.1093919	-0.3881413	-1.7338550	-1.31777536
[X4]	0.51763018	-1.78847775	-0.3881413	1.9544634	0.3151251	0.33632400
[X5]	0.33532957	-0.13053975	-1.7338550	0.3151251	2.1834721	0.55243297
[X6]	0.09876099	0.05754396	-1.3177754	0.3363240	0.5524330	0.93625031

Table: 2.3 INSTRUCTOR C

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	4.8457211	-2.555907	-0.2093084	-0.18779379	-0.7743247	-0.76923487
[X2]	-2.5559071	9.207443	-2.5823903	-1.97615889	-2.6824328	1.38738432
[X3]	-0.2093084	-2.582390	4.1069538	-0.88162805	0.8288780	-1.37194790
[X4]	-0.1877938	-1.976159	-0.8816281	2.28941251	0.6438569	0.08480885
[X5]	-0.7743247	-2.682433	0.8288780	0.64385693	2.5032129	-0.68023717
[X6]	-0.7692349	1.387384	-1.3719479	0.08480885	-0.6802372	1.63832569

Table: 2.4 INSTRUCTOR D

	[X1]	[X2]	[X3]	[X4]	[X5]	[X6]
[X1]	4.89797574	-2.4425037	0.02862178	-0.3232214	0.1795423	-0.39484253
[X2]	-2.44250369	8.2144089	-1.81422606	-0.7375047	-2.5831144	-0.97099604
[X3]	0.02862178	-1.8142261	2.34517012	-0.1322247	0.1144340	-0.08069809
[X4]	-0.32322142	-0.7375047	-0.13222469	1.2139065	0.1380780	-0.15186442
[X5]	0.17954225	-2.5831144	0.11443400	0.1380780	3.2426439	-0.24421184
[X6]	-0.39484253	-0.9709960	-0.08069809	-0.1518644	-0.2442118	1.97031378

Table: 2.5 INSTRUCTORS (ALL)

[X1]	5.0475783	-2.80259374	-0.2702998	-0.18421160	-0.46107900	-0.19428280
[X2]	-2.8025937	4.79327760	-0.5352151	-0.08331626	-0.62960340	-0.06561393
[X3]	-0.2702998	-0.53521512	1.9859239	-0.23206533	-0.45523250	-0.67149712
[X4]	-0.1842116	-0.08331626	-0.2320653	0.31248456	0.03275538	0.11807037
[X5]	-0.4610790	-0.62960340	-0.4552325	0.03275538	1.71589548	-0.03461549
[X6]	-0.1942828	-0.06561393	-0.6714971	0.11807037	-0.03461549	0.93096180

Table 3: Mean Evaluation Scores Received By Instructors Before and After The Issue of Grades

Variable	Instructor A		Instructor B		Instructor C		Instructor D	
	Sem1	Sem 2	Sem1	Sem 2	Sem1	Sem 2	Sem1	Sem 2
A	4.46	4.10	4.25	3.91	4.48	3.87	4.53	3.56
B	4.30	3.95	4.36	3.86	4.47	3.72	4.56	3.72
C	4.39	3.88	4.21	3.59	4.67	4.15	4.60	3.26
D	4.48	4.14	4.05	3.74	4.35	3.78	4.57	3.43
E	3.86	3.42	3.75	3.25	3.95	3.65	4.19	3.28
F	4.24	3.67	4.27	3.25	4.63	3.71	4.83	3.75
Total	4.28	3.86	4.14	3.60	4.42	3.81	4.54	3.50

INSTRUCTOR PERFORMANCE EVALUATION QUESTIONNAIRE (ADDIS ABABA UNIVERSITY to be Completed by Students)

This questionnaire has been prepared to get your views regarding the teaching performance of your instructor. Please respond to the items on the questionnaire frankly and honestly. Do Not write your name on the questionnaire, but write the name of your instructor, your department and faculty, the title of the course number, the academic year, semester, and your college year in the spaces provided. After you have filled in these, read carefully each of the statements listed from 1-30 below. Then indicate how you evaluate your instructor on each statement by Circling one of following options against each statement:

VG= Very Good	F= Fair	VP= Very Poor
G= Good	P= Poor	DK= Do not Know

Instructor's name ____________________________ Course title________________

Course No. __________ Your Department __________________Faculty______

Academic Year 199__/199_______ Semester ________

Your year: Undergraduate program: I II III IV V VI (circle one)

Graduate program: I II III

1	Clarification of the statement of general objectives of course	1.	VG	G	F	P	VP	DK
2	Presentation and clarification of course plan and course outline	2.	VG	G	F	P	VP	DK
3	Clarification of the statement of specific objectives at the beginning of each chapter or unit	3.	VG	G	F	P	VP	DK
4	Knowledge of the subject matter	4.	VG	G	F	P	VP	DK
5	Preparation for classes	5.	VG	G	F	P	VP	DK
6	Presentation of subject matter clearly in the language of instruction	6.	VG	G	F	P	VP	DK
7	Presentation of subject matter	7.	VG	G	F	P	VP	DK
8	Willingness to encourage students to ask or answer questions in class	8.	VG	G	F	P	VP	DK
9	Willingness to let students express their opinions about the Course in the classroom	9.	VG	G	F	P	VP	DK
10	Availability during consultation hours	10.	VG	G	F	P	VP	DK
11	Punctuality for classes	11.	VG	G	F	P	VP	DK
12	Meeting classes regularly (non-absenteeism)	12.	VG	G	F	P	VP	DK
13	Ability to arouse students interest and provoke their thinking	13.	VG	G	F	P	VP	DK
14	Ability to encourage student participation in the classroom	14.	VG	G	F	P	VP	DK
15	Appropriate use of available and relevant instructional materials (blackboard, maps...)	15.	VG	G	F	P	VP	DK
16	Providing feedback on homework, tests and /or assignments on time	16.	VG	G	F	P	VP	DK
17	Usefulness of homework and /or assignments for course work	17.	VG	G	F	P	VP	DK
18	Presence of question in tests, exams or homework, that require reasoning	18.	VG	G	F	P	VP	DK

19	Amount of time allowed for tests, assignments, or mid semester exams	19.	VG	G	F	P	VP	DK
20	Coverage of course content in tests or mid-semester exams	20.	VG	G	F	P	VP	DK
21	Fairness in marking /grading	21.	VG	G	F	P	VP	DK
22	Clarification of the methods of assessing students	22.	VG	G	F	P	VP	DK
23	Coverage of content according to curse outline	23.	VG	G	F	P	VP	DK
24	Providing/giving a list of reference materials for the course	24.	VG	G	F	P	VP	DK
25	Use of class period for teaching or discussion of subject and related matters	25.	VG	G	F	P	VP	DK
26	Respect for students	26.	VG	G	F	P	VP	DK
27	Willingness to listen to a student's problems	27.	VG	G	F	P	VP	DK
28	Ability to maintain appropriate discipline in the class	28.	VG	G	F	P	VP	DK
29	Clarity of question in tests, and/or mid-semester exams	29.	VG	G	F	P	VP	DK
30	Overall assessment of instructor's teaching effectiveness	30.	VG	G	F	P	VP	DK