# A Multidimensional Analysis of Student Evaluations of College English Instruction: An Application of Nonparametric Methods[1]

Amanuel Gebru[2] and Mulugeta Gebreselassie[3]

**Abstract:** *A number of researchers give reason for an exercise of caution in interpretations of student-supplied evaluative data. Cross-culturally, student characteristics may be systematically extraneously related to student ratings. Arguably, EFL students with diverse backgrounds may have their own theories of good teaching which may be revealing more information about themselves than about their instructors. It may therefore be hypothesized that extraneous variables have a differential impact on English as a Foreign Language students' ratings. This study attempted to examine some of the most important extraneous variables which explain the variation in EFL student ratings of their instructors in Addis Ababa University. The data were obtained using a pre-tested adapted version of the Instructional Assessment Measure (IAM) developed by the University of Otago in New Zealand and adapted by a number of universities worldwide. The ten-item scale was rated on a 5 point Likert scale. Then the data were duly analyzed using descriptive and nonparametric inferential techniques. The results showed that evaluations are a function of gender, level of learning, program of learning and regional background.*

## Introduction

Student evaluations of college teaching, partly as a reflection of their genesis, have for a considerable period remained virtual North American phenomena. Also incontestably research in the area has shown a clear North American dominance (Husbands and Fosh, 1993). But in recent years the evaluation of college instruction has made inroads into the European university system and more recently

[1] This study was presented in the ELTNET founding conference.

[2] Amanuel Gebru, Assistant Professor, Department of Foreign Languages and Literature, Addis Ababa University

[3] Mulugeta Gebreselassie, Division of Biostatistics, University of Southern California.

into African universities where pressures for accountability have started to be felt. Expectedly, African research into the subject is in its infancy, but then it is also beset with circulatory difficulties, and therefore heavily regional/ national, much like research in many universities of the Third World (Hussain, et. al., cited in Tirusew,1998).

In Ethiopian public universities in general, and in Addis Ababa University in particular, the North American system of Student Evaluation of Instructional Effectiveness (SEIE) was reintroduced in earnest in 1996 in tune with the emergent spirit of free-marketism and total quality control. This recency offers a fertile field of inquiry for transdisciplinary researchers.

If evaluation of instruction is vital, it is probably no more so than it is in the area of English language instruction (itself a multidisciplinary field) where cries about Addis Ababa University students' communicative incompetence have led to methodological shifts, revision of syllabus, writing of new textbooks and retraining of teachers in more trendy communicative methodologies. Meetings with stakeholders have revealed considerable dissatisfaction with the incompetence of the generality of Addis Ababa University graduates in English as a Foreign Language (Department of Foreign Languages and Literature, 2000). Theoretical and methodological developments in general education have led to parallel developments in English Language Teaching and in consequence there is a growing recognition of the place of evaluation in English language instruction (Roberts and Roberts, 1994).

In light of the developments in the field of English instructional evaluation several important questions can be asked. What are the correlates and determinants of first year students' evaluations of First Year English Language Instruction? Do students recognize pedagogic excellence in College English or are they affected in their ratings by extraneous considerations? These are some of the

questions that need to be addressed by a study that investigates evaluation of ELT instruction which also require meticulous attention to theoretical, methodological and statistical considerations ( Gigliotti and Buchtel, 1990).

*Related Studies*

Studies into the correlates and determinants of college students' evaluations of instruction have often come up with mixed results. The most prominent support for the reliability and validity of student evaluations has come from Marsh (Marsh 1995; 1987; Marsh and Dunkin, 1992) who in numerous studies concludes that the contribution of biasing factors in ratings is nowhere close to significance. But other writers have come up with dissimilar findings. Some studies, for instance, indicate that male and female instructors are rated differently (Kaschak, 1978; Lombardo and Tocci, 1979). Bending (1952) found that female students rated male instructors more negatively than they did female instructors. In a more recent study, grade received differentially affected female instructors (Kierstead et. al, 1988). The studies that did not find sex differences include Elmore and Pohlmann (1978) and Freeman (1994).

In studies that manipulated other variables mixed findings were reported. In an investigation into the effects of grade point average, rank of instructor and grade expected/earned (Elmore and Pohlmann, 1978) found no significant levels of effect. Similarly in earlier studies Aleamoni and Graham (1974) and Aleamonie and Yimer (1973) did not find rank-based differences. Contrarily, Villano (in Elmore and Pohlmann, 1978) found that associate and full professors obtained more positive ratings than instructors and assistant professors.

In studies of the effect of expected grades on ratings, contradictory findings were reported. While Kennedy (1975) found no significant connection, Pohlmann (1975) obtained a covariation between grades obtained and ratings given. Similarly Beatty (in Elmore and Pohlmann

1978) found that there was no correlational relationship between grade point average and student ratings.

*Theoretical Perspectives*

As a transdiscipline, evaluation has been in the service of a wide range of subjects including ELT (Cross, 1989; Finocchiaro, 1989; Kwo 1989) despite the unabated theoretical controversies and unresolved issues related to its various uses. In a discussion of diverse theoretical orientations related to the investigative effort, Worth and Sanders (1987) mention that in the decades between 1967 – 1987 more than 50 evaluation models evolved. One simple example is Scriven's multiplist approach (1991), which views evaluation as "multidimensional" and "multiperspectival" (IEE, 1994). More relevant to this subject however is the accuracy-based approach (Patton, 1986) because summative evaluation requires psychometric accuracy. This is crucially important because Addis Ababa University tends to emphasize the management-oriented approach, which focuses on performance information for utility in personnel decisions.

The issue of the reliability of student-supplied data raises concomitant theoretical issues about the occurrence of biases in students' evaluative information. Despite the confidence inspired by North American research about the insignificance of the weight of biases (Marsh 1995; 1987; Marsh and Dunkin 1992; Elmore and Pohlmann, 1978) the ongoing theoretical debate and field research are far from confirmatory of the absence of discord. In contrast to the substantial evidence of the stability of student ratings, Amanuel (1999), for instance, demonstrated that students' ratings of first year English instruction were not reproducible upon a re-administration. Theoretically this may be explained by an extension of the self-esteem model (Giglioootti and Buchtel, 1990) which postulates that ratings are a function of grade outcomes. Also, Roberts and Roberts

(1994) indicate that ELT needs to provide context-sensitive information about the suitability of particular evaluative approaches.

A number of researchers (Hocutt Nd), Centra (1979); Cashin and Perrin (1983); Ducett and Kennedy (1982) give reason for an exercise of caution in interpretations of student-supplied evaluative data. Cross-culturally, student characteristics may be systematically related to student ratings. Arguably, students with diverse backgrounds may have their own theories of good teaching, which, according to Leventhal et. al. in Rotem and Glasman (1979), may be revealing more information about themselves than about their instructors. It may therefore be hypothesized that extraneous variables have a differential impact on students' ratings. Several conceptual questions may be raised. Does, for instance, the doctorate of an instructor affect the way students' process evaluative information any differently from the way a masters/bachelors does? Do male and female students have systematically differing evaluative frames which may be related to the homosocial theory (Bluemann, 1981) which posits students' preference for same-sex instructed classes? What is the effect of a success/failure experience as measured in terms of grade point average on the evaluative frame of EFL students, a condition which may relate to the expectation/ confirmation model which posits that ratings given to instructors are a function of students' academic success with failure biasing evaluations downward (Gigliotti and Buchtel, 1990)? Also does the provincial/metropolitan background of students systematically affect the way they judge their instructors' pedagogic effectiveness? In Gigliotti and Buchtel (1990), older students and students with more highly educated fathers who were also mostly cosmopolitan gave lower evaluations.

We also asked, supported by the hypothesis implied in the European literature (Husbands and Fosh, 1993), whether the time of the day in which classes are given and evaluations administered provides a systematic difference especially in tropical weather where afternoons

generally tend to be lazy. Language acquisition theory (Krashen, 1981) also supports the hypothesis that a relaxed mental/affective state would facilitate language acquisition and by implication a better reception and evaluation of instruction.

Assisted by the theoretical literature, we also wondered conceptually whether students' maturity levels and work experience affect the way they process evaluative information and rate instruction. Thus part-time students who are generally presumably older and more experienced than regular freshmen may be expected to have a differential evaluative orientation - a condition which may also be affected by whether they are fee-paying or non-fee-paying which all regular students in Ethiopian public colleges are.

We conceptualized that because of differences in experience, knowledge and perspective (Rotem and Glasman, 1979:498), geographic background (Addis Ababa versus regions), time of day, status (fee-paying, non-fee-paying), students may differentially rate instructional performance in the College English class.

*Research Questions*

With support from the literature we posed the following research questions:

- Do male and female EFL students rate a male instructor differentially?
- Do provincial and metropolitan students evaluate their College English instructors differently?
- Do regular and extension students demonstrate a systematic difference in their ratings?
- Do morning and afternoon students differentially rate their EFL instructors?

- Do degree and diploma students differ in their ratings of College English instruction?
- Does grade expected affect the way students rate their College English instructor?
- Does grade obtained from a previous College English instructor affect the way students rate their current College English instructor?
- Does Grade Point Average influence students' evaluation of College English instruction?

**Overview of Methodology**

*Subjects*

Subjects were first year College English students of AAU with contrastive characteristics. They were male and female, provincial and metropolitan, regular and extension and, degree and diploma level students. All were instructed consistently either in the morning or afternoon shifts by male full-time academics. A full catalogue of academic ranks was unavailable for a full treatment of academic rank as a variable.

*Instrument*

Subjects were administered a pre-tested adapted version of the Instructional Assessment Measure (IAM) developed by the University of Otago in New Zealand and adapted by a number of universities worldwide. The ten-item scale was rated on a 5 point Likert scale ranging from, for instance, very well organized (5) to very disorganized (1) or very helpful (5) to very unhelpful (1). The form also has an item requesting a global effectiveness evaluation.

In using the scale, we found some literature support for the applicability of western evaluative instruments in nonwestern contexts

(Watkins, 1994). In choosing a short form, which is useful for summative purposes, we also found research support that short forms help to save time and money (Cashin and Dowey, 1992; Frey, 1978). We also reasoned that Addis Ababa University's system of evaluation tends to be management driven rather than utility focused. Apart from helping to ward off respondent fatigue, the short form also proved to be more discipline-relevant than the long form developed by Addis Ababa University for use by all departments at all levels.

*Methods of data collection*

Primary data were collected using a questionnaire, which consists of 18 questions. The questions were related to the academic performance and the demographic aspects of the student, the academic program and the instructor. They were designed in such a way that they could be sufficiently simple to answer and include the most relevant elements of the official evaluation questionnaire of Addis Ababa University.

Before the administration of the questionnaire a sample design that included fixing the number of students to be administered and the sampling technique to be used were decided. The number of subjects was decided on a purposive basis (as one way of deciding the sample size is the purpose of the study) (Chochran, 1977). The sampling design was multistage sampling. It involved single stage cluster sampling after the university was stratified by faculties. That is, first, the university was divided into strata (faculties) and then from each faculty a certain number of freshman sections (clusters), with all units in each section to be involved in the study, were selected.

From a total of 44 sections in the Science Faculty and 66 sections in the Social Sciences College (10 of them extension) and 12 sections from the Faculty of Business and Economics (a total of about 4800 students), 11 sections were selected as a sample expecting 440 students to be studied. However, the total number of students studied

was 308. The expected number of students was not obtainable because some students were absent during the period/class time we administered the forms. Also other students returned partially filled questionnaires which we eliminated later. We feel that since the majority of non-respondents were those who did not attend during that particular period, the missingness is random (ignorable) and hence will not affect our results substantially (Lindsey, 1999).

*Method of Analysis*

Once data were collected the first primary step was cleaning and editing the data for possible inconsistencies and inaccuracies. Then the data were fed into a computer with an SPSSWIN version 9.0 Software. After the completion of the data entry a further cleanup was made. The second step of the analysis stage was a preliminary analysis of the data which involved tabulating the frequencies of each variable and summarizing relevant statistics. Here frequency tables, cross tables and summary statistics were computed. This step was followed by the test of hypotheses on the questions posed by the study. This was handled using nonparametric methods for the reason that our data was not normally distributed and in this kind of situation they are more efficient than parametric methods. An application of nonparametric methods assumes that the observed data set satisfy the following assumption: the two samples should be independently and identically distributed random samples from two mutually independent continuous populations (Hollander and Wolfe, 1999).

*Distribution Free Rank Sum Test*

Let $F_1$, $F_2$ and $F_3$ be distribution functions corresponding to random variables X1, X2 and X3 respectively. The hypothesis of interest is $F_1(t) = F_2(t) = F_3(t)$ for every t. This asserts that the random variables $X_1$, $X_2$ and $X_3$ have the same probability distribution which is not specified. The alternative to this is that they do not have the same probability distribution. That is, when described by the location shift

model, $F_i(t) = F_j(t-\Delta_i)$ where $\Delta_i = \sum (X_i) - \sum (X_j)$, for $i \neq j$. This can further be reduced to:

$H_0$: $\Delta_i = 0$

$H_1$: $\Delta_i \neq 0$

*Wilcoxon Signed Rank Test*

This procedure is important when the primary interest of the analysis is centered on the relative locations of two populations. Let a sample of $n_j$ observations be taken from 2 populations which are mutually independent. The procedure for computing the test statistic is as given below. Let W designate the Wilcoxon two sample rank sum statistic. To compute W, first of all rank the values of the random variables from least to greatest and denote them by $S_i$ (I= 1, 2...n). Let $W_1$ be the sum of ranks for the first random variable and $W_2$ be the sum of ranks for the second variable. Then take the minimum of $W_1$ and $W_2$ to be value of the test statistic, W. Reject $H_0$ if $W \geq W_{\alpha/2}$ or if $W \leq n(m+n+1)- W_{\alpha/2}$, otherwise do not reject $H_0$. Here $W_{\alpha/2}$ is the tabulated value of W. The large sample approximation of W can be used if n is sufficiently large.

*Mann Whitney Test*

This procedure is also important when the primary interest of the analysis is centered on the relative locations of two populations. Let a sample of $n_j$ observations be taken from 2 populations which are mutually independent. To compute the Mann Whitney test statistic (H), first combine all observations from the 2 samples and order them from least to greatest. Let $r_{ij}$ denote the rank of $x_{ij}$ in the joint ranking and let $R_j$ be the sum of $r_{ij}$'s for i =1 to $n_j$, and $R_{.j}$ be the arithmetic mean of $R_j$ then (j = 1, 2)

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{2} nj(R.j - \frac{N+1}{2})^2$$

The decision rule at $\alpha$ level of significance is to reject $H_0$ if $H > h_\alpha$, otherwise do not reject $H_0$. The large sample chi-square approximation of H can also be used if the sample size is significantly large.

**Results**

Descriptive results depicted in Annex 1 show that there were 77.7 percent males and 22.3 percent females among 308 students who were studied. These were 49 percent regular and 51 percent extension participants. When desegregated by shift of their class time, 42.6 percent were in the morning shift and the rest 57.4 were in the afternoon. Further classification by their level of learning showed that about 85.7 percent of the students were degree students while the rest were diploma-level students. With respect to their regional background about 35.3 percent of the students came from the regions. The majority were from Addis Ababa, may be because most of the extension students reported their region to be Addis Ababa.

The results further showed that, in terms of helping students to communicate, College English II was evaluated to be extremely or very valuable by most of the students. Only 2.9% of the students said it was not at all important. The reported considerable satisfaction with the course might be derived from the fact that almost all instructors were very well organized when they delivered lessons. This was further confirmed by the students' high ratings of the level of stimulation by their instructors. In fact the ratings given to the communication ability of the instructors was very good and above. This could suggest that that most of the time/always instructors came to class sufficiently prepared.

The responses given to the question asking freedom of discussing issues in groups was scattered between *always, frequently and sometimes*. About 2% reported that there was no possibility of making free discussions. Most students agreed that their instructors

were more than helpful to their students. This was explained by the tradition of fairness in marking of assignments and exams. About 79.2% said that assignments and exams were always returned promptly and about 87.3% responded that the marking of the assignments and exams was more than moderately fair.

Overall the effectiveness of instructors in teaching the course ranged between very effective and effective. About 47.4% rated their instructors as very effective, 37.3% as effective and about 2.2% as ineffective and very ineffective.

**Table 1: Test of Location by Sex**

| Grouping variable | Test variable | Mann Whitney | Wilcoxon | P value |
|---|---|---|---|---|
| Sex | Q1 | 5974 | 31625 | .017** |
| | Q2 | 6570.5 | 32221.5 | .150 |
| | Q3 | 6428 | 32079 | .086*** |
| | Q4 | 5724 | 31375 | .004* |
| | Q5 | 7262.5 | 32913.5 | .867 |
| | Q6 | 6806 | 32457 | .340 |
| | Q7 | 6113 | 31764 | .019** |
| | Q8 | 5988.5 | 31639.5 | .009* |
| | Q9 | 5455.5 | 31106.5 | .001* |
| | Q10 | 5972.5 | 31623.5 | .012** |
| | Q11 | 6585 | 8730 | .155 |
| | Q12 | 7320 | 32971 | .964 |
| | Q13 | 6332.5 | 31983.5 | .075*** |

(* < 1%; ** <5%; *** <10%)

The first semester College English grade of most (56.8%) of the students was C; and among the 308 subjects only 10% scored A and 2.2% scored D and below. However expected College English II grades were very high. About 34.4% expected an A grade, about 47.4% expected a B grade and 11.4% expected a C grade. Those who expected grades less than C were only 6.8%.

The overall performance of the study subjects as measured by their first semester GPA demonstrated the representatives of the sample.

About 41% of the students had a first semester GPA ranging between 2.0 and 3.0. while there were18.2% with GPA less than 2.0. There were also 17.9% with GPA greater than 3.0.

A further analysis of the data using nonparametric procedures on the location parameter of the different evaluation questions was done using the Mann Whitney and Wilcoxon tests. The values of the test statistic together with their P values are given in Tables 1 to 5.

Table 1 shows that evaluation of faculty by their students has a gender perspective. The rating of instructors by males was found significantly higher than females. In variables such as Q1, Q3, Q4, Q7, Q8, Q9, and Q10 the evaluation given by males was higher than the evaluation by females. The highest significance (at less than 1%) was observed in variables Q4, Q8 and Q9.

**Table 2: Test of Location by Program**

| Grouping variable | Test variable | Mann Whitney | Wilcoxon | P value |
|---|---|---|---|---|
| Program | Q1 | 10048.5 | 20488.5 | .282 |
| | Q2 | 7427 | 17867.0 | .000 |
| | Q3 | 9536.5 | 19976.5 | .052 |
| | Q4 | 10061.5 | 20501.5 | .282 |
| | Q5 | 8592.0 | 19032.0 | .000 |
| | Q6 | 8535.5 | 18975.5 | .001 |
| | Q7 | 9574.5 | 20014.5 | .054 |
| | Q8 | 6685.0 | 17125.0 | .000 |
| | Q9 | 9813.5 | 20253.5 | .155 |
| | Q10 | 9128.5 | 19568.5 | .012 |
| | Q11 | 10269.0 | 21594.0 | .414 |
| | Q12 | 9422.5 | 19862.5 | .040 |
| | Q13 | 8987.0 | 19427.0 | .009 |

(* < 1%; ** <5%; *** <10%)

Table 2 shows that program of learning, viz. extension or regular, also had a significant contribution to the variation in instructor evaluations by their students.

Table 3 shows the effect of attending classes in the morning or in the afternoon for regular students only. A statistically significant difference was observed in the variables Q1, Q2, Q3, Q4, Q5, Q6, Q7 and Q8. Except in variable Q6 afternoon students' rating was lower than the rating of morning students.

**Table 3: Test of Location by Class Time**

| Grouping Variable | Test variable | Mann Whitney | Wilcoxon | P value |
|---|---|---|---|---|
| Class time | Q1 | 2003 | 4281 | 0.054*** |
| | Q2 | 1983 | 4261 | 0.017** |
| | Q3 | 1791 | 4069 | 0.001* |
| | Q4 | 1616 | 3894 | 0.000* |
| | Q5 | 876.5 | 4154.5 | 0.001* |
| | Q6 | 1507 | 4208 | 0.000* |
| | Q7 | 1651 | 3929 | 0.000* |
| | Q8 | 2425 | 5126 | 0.904 |
| | Q9 | 2317.5 | 4595.5 | 0.575 |
| | Q10 | 1953 | 4231 | 0.020** |
| | Q11 | 2361 | 4639.5 | 0.688 |
| | Q12 | 2250 | 4528 | 0.377 |
| | Q13 | 2398 | 5099 | 0.831 |

(* < 1%; ** <5%; *** <10%)

Table 4 shows the contribution of level of learning to the differences in the ratings of instructors by their students. Degree students' rating of their instructors was found significantly higher than diploma students' in the variables Q2, Q3, and Q8. While in variables Q5, Q6 and Q7 the rating of diploma students of their instructors was significantly higher. Evaluative differences of global effectiveness were found insignificant.

Regional background was also found to be a contributor to the overall differences in students' ratings of their instructors. Q1, Q2, Q3, Q6, Q8 and Q10 were significant at less than 5% level of significance and Q4 and Q5 were significant at less than 10%. In all variables the ratings of Addis Ababa students were found to be lower than those of out-of-Addis Ababa students.

**Table 4: Test of Location by Level of Learning (Degree vs Diploma)**

| Grouping variable | Test variable | Mann Whitney | Wilcoxon | P-value |
|---|---|---|---|---|
| Level | Q1 | 5111.5 | 6014.5 | .744 |
| | Q2 | 4417.5 | 36043.5 | .062*** |
| | Q3 | 4306.0 | 35932.0 | .033** |
| | Q4 | 5141.5 | 36767.5 | .787 |
| | Q5 | 5138.5 | 6041.5 | .752 |
| | Q6 | 4936.5 | 5839.5 | .486 |
| | Q7 | 5069.0 | 5972.0 | .649 |
| | Q8 | 4415.5 | 36041.5 | .055** |
| | Q9 | 4861.5 | 36487.5 | .397 |
| | Q10 | 4610.0 | 36236.0 | .156 |
| | Q11 | 4135.0 | 35761.0 | .012** |
| | Q12 | 4011.0 | 35637.0 | .007* |
| | Q13 | 5185.0 | 36811.0 | .859 |

(* < 1%; ** <5%; *** <10%)

**Table 5: Test of Location by Region (Addis Ababa vs Others)**

| Grouping variable | Test variable | Mann Whitney | Wilcoxon | P value |
|---|---|---|---|---|
| Region | Q1 | 7509.5 | 12660.5 | .004* |
| | Q2 | 7438.0 | 12589.0 | .002* |
| | Q3 | 7746.5 | 12897.5 | .007* |
| | Q4 | 8038.5 | 13189.5 | .038** |
| | Q5 | 8411.0 | 13562.0 | .091*** |
| | Q6 | 7693.0 | 12844.0 | .009* |
| | Q7 | 8963.5 | 14114.5 | .514 |
| | Q8 | 7963.0 | 13114.0 | .018** |
| | Q9 | 8528.5 | 13679.5 | .200 |
| | Q10 | 7972.0 | 13123.0 | .025** |
| | Q11 | 9152.0 | 26357.0 | .749 |
| | Q12 | 8498.5 | 13649.5 | .169 |
| | Q13 | 8941.5 | 14092.5 | .528 |

(* < 1%; ** <5%; *** <10%)

A similar statistical analysis of the correlation between the evaluation items showed that most of the items were highly associated to each other (See Table 6). In fact the intercorrrelations were found to be mainly positive except those between Q11 and some items. A more than 50% positive correlation was observed between Q2 with Q3 and Q10, Q3 with Q10 and Q7 with Q10.

A test of independence between grouping variables such as sex, program of learning, class time, level of learning and region of origin and measures of performance such as first semester College English

**Table 6: Correlation (Spearman's correlation coefficient)**

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Q1** | | .3 | .25 | .29 | **.06** | .16 | .29 | .16 | .17 | .27 | **-.08** | .11 | .17 |
| **Q2** | | | .56 | .46 | .40 | .14 | .43 | .30 | .18 | .52 | **-.05** | **.08** | **.11** |
| **Q3** | | | | .52 | .33 | **.09** | .45 | .13 | .16 | .51 | **.10** | .14 | **.06** |
| **Q4** | | | | | .33 | .18 | .46 | .18 | .16 | .50 | **-.04** | **.06** | **11** |
| **Q5** | | | | | | **.09** | .36 | .23 | .16 | .29 | **-.06** | **.11** | **06** |
| **Q6** | | | | | | | .16 | .38 | .12 | .20 | **-.01** | **.06** | .13 |
| **Q7** | | | | | | | | .24 | .23 | .52 | **-.03** | .12 | **.03** |
| **Q8** | | | | | | | | | .24 | .22 | **-.09** | **.09** | **.11** |
| **Q9** | | | | | | | | | | .18 | **-.04** | .12 | **.03** |
| **Q10** | | | | | | | | | | | **.08** | .12 | **.09** |
| **Q11** | | | | | | | | | | | | .48 | **-.16** |
| **Q12** | | | | | | | | | | | | | **.04** |
| **Q13** | | | | | | | | | | | | | |

*Bold figures p-value>0.05.*

grade, expected College English grade, and overall first semester GPA showed that most of them were not significant showing independence. The association between first semester GPA with program of study was exceptionally high. And all performance measures were highly significantly associated with level of learning.

**Table 7: Test of Independence**

| Grouping Variable | Independent Variable | Chi Square | P-value |
|---|---|---|---|
| Sex | Q11 | 3.573 | 0.62 |
| | Q12 | 9.481 | 0.091*** |
| | Q13 | 7.111 | 0.130 |
| Program | Q11 | 9.45 | 0.092*** |
| | Q12 | 7.72 | 0.172 |
| | Q13 | 38.42 | 0.000* |
| Class time | Q11 | 2.596 | 0.762 |
| | Q12 | 4.487 | 0.482 |
| | Q13 | 6.119 | 0.190 |
| Level | Q11 | 10.943 | 0.050** |
| | Q12 | 11.316 | 0.045** |
| | Q13 | 11.790 | 0.019** |
| Region | Q11 | 4.511 | 0.478 |
| | Q12 | 5.730 | 0.333 |
| | Q13 | 7.290 | 0.121 |

(* < 1%; ** <5%; *** <10%)

## Discussion and Propositions

The study yielded results which both confirm and disconfirm findings reported in the extant literature. Interestingly, College English was rated as nearly excellent in terms of its communicational usefulness as indeed were the various dimensions of the evaluation scale viz. the organizational competence of the instructors, their communicative competence, motivational capacity, preparation for classes, democratic character, attitude to students, fairness and promptness in marking papers and their global effectiveness. The average rating a teacher received was 4.2 out of 5.

These high ratings are perhaps in agreement with Hocutt's (ND) assertion that students expect their instructors to be very effective and supply generous ratings of them. This may be particularly true in the first year where students have made a radical transition in the sense that they are now fresh in a university setting and instructed by more highly qualified academics under circumstances fundamentally different from school delivery systems. It may be argued that they have reverential attitudes to faculty whom they also probably take as role models.

Viewed from a gender perspective, the results seem to suggest that student gender was an important variable. The instructors in the study (all males) were more positively rated by male than female students. Particularly highly significant correlated were the items measuring marking responsibilities and stimulation of interest in the field (which may of course have been affected by pre-existing career plans). The relevant findings add to the absence of agreement in the vast volume of research in the interpretation of the role of gender as a variable. In substantial agreement with Basow and Silberg (1987), Kaschal (1978) and Basow and Howe (1982) the results seem to suggest that there is a preference for same sex instructors as well as a better rating thereof. Viewed in context the lower ratings given to male instructors by female students may suggest that male

instructors may often lack the warmth, cheerfulness and supportiveness that female instructors have been credited with ( Basow and Distenfeld 1985). It may also be assumed that female students more than male students value these expressive traits. Further, Freeman (1994) has postulated that an interaction of course type taught and instructor gender may influence in certain ways raters' perceptual and evaluative orientation. This argument may be tenable in consideration of the fact that modern languages are perceived as feminine fields and understood as requiring feminine teaching styles. However with contradictory evidence that student gender is inconsequential across disciplines (Marsh and Hau, 1997) further research seems to be warranted.

The effect of program of learning (regular versus extension) was notably significant. Regular students rated their instructors more positively than did extension students, except on "marking exams properly" and "overall instructional competence". This may suggest that extension students who are normally fee-paying may be more demanding clients than regulars who are scholarship students. The results may also suggest that extension students who may also be taken as older and more experienced (some already have diplomas) may have a more diagnostic perspective. The results confirm the hypothesis that ratings may be dependent on whether students are fee-paying (Husbands and Fosh, 1993). They are also consistent with Gigliotti and Buchtel's (1990) findings that older students gave lower ratings.

The effect of level of learning (degree versus diploma) on evaluations produced mixed results when univariate analysis was conducted. On items Q5, Q6 and Q7 diploma students rated their instructors more positively while the reverse was true for Q2, Q3, and Q8. Overall, level of learning was not a significant factor while it could be assumed that higher-level students could be more critical in their testimonials of English instruction. As the findings seem to suggest there is a small positive significant correlation between GPA and

ratings. Similarly in a correlation matrix for individual level analysis, Gigliotti and Buchtel (1990) found that GPA was not a significant predictor of ratings.

The ratings of students from Addis Ababa were significantly lower (at 99% level of confidence) suggesting that this group comprises most of the extension students who as fee-paying expect instructional quality worth their money. These cosmopolitan students may also be more diagnostic because of a probable higher level of English proficiency over regional students, which can moderate their levels of evaluative appreciation. All available evidence seems to indicate that students who come from Addis Ababa generally perform better in English language tests than those from the regions who look to be more disadvantaged when it comes to foreign language learning (Mulugeta, 1997). It also appears that cosmopolitan students come from more educated parents, a condition which may be contributory to the posited higher proficiency levels in this category of students.

The correlation of 48% between grade expected and ratings given supports some findings reported in the literature (Feldman 1976; Pohlman 1975) and disconfirm those that do not (Kennedy 1975). Amanuel (1999) did not find a covariation, but his findings need to be interpreted more cautiously because he did not pair individual expectations and ratings. The present research interestingly found a significant correlation (P-value<0.05) between gender and grade expected with males expecting higher grades. This strengthens the possible causal link between grades expected and ratings given in the sense that females gave lower ratings and had lower expectations.

Levels of ratings were also hypothesized as correlates of class times. We suspected that learning before noon would lead to higher ratings because arguably morning students would have a more positive frame of mind as a result of a finer state of weather than afternoon students who would experience greater lassitude as a result of the occasional sweltering heat of the metropolitan tropical weather. (In a

study of absenteeism, Darge (2000) found that there were more absences in the afternoon). Instructors may also be assumed to experience decreased instructional vitality in consequence of the strains of morning sessions which may lead to them receiving corresponding lower ratings. However our findings do not support the thesis that temporal differences can lead to differential ratings. Nevertheless these findings should be interpreted more cautiously. It should be understood that we did not have the requisite meteorological data to establish a valid temporal and evaluative covariation. It is possible given the meteorological variability in Addis Ababa that evaluations conducted in different afternoons may yield different results if indeed there is a causal relationship between a temporal state and an evaluative mood. We hope that a more rigorous investigation based on accurate meteorological data can lead to more confident conclusions.

## Propositions for Further Research

We believe the present research has identified some demographic and situational variables that can bias ratings of English instruction upwards or downwards. In view of the recency of ratings in our context we believe that further research is in place. We feel that research may be conducted which investigates extraneous factors such as length of period, class size, instructor sex, course type, course status (major, minor), temporal state and evaluative mood and course credit/ weight as well as the numeracy/literacy interests/competencies of raters. We believe that an understanding of the place of these factors can help to inform the evaluation of English language instruction in higher education ELT programs.

## References

Aleamoni, L.M. and Graham, M.H. (1974). *The Relationship Between CEQ Ratings and Instructor's Rank, Class Size, and Course Level.* **Journal of Educational Measurement**. 11, 189-202.

_____ and Yimer, M. (1973). *An Investigation of the Relationship between Colleague Rating, Student Rating, Research Productivity and Academic Rank in Rating Instructional Effectiveness*, **Journal of Educational Psychology**, 64, 274-277.

Amanuel Gebru. (1999). *Students' ratings of their College English instructors before and after the issue of grades.* **The Ethiopian Journal of Education**, 19,2, 47-76.

Basow, S.A. and Howe, K.G. (1982). Sex bias in Evaluations of College Professors. Paper presented at the meeting of the Eastern Psychological Association, Baltimore, MD.

Basow, S. And Distenfield, S. (1985). *Teacher Expressiveness: More Important for Male Teachers Than Female Teachers?* **Journal of Educational Psychology,** 77,1, 45-52.

Basow, S. And Silberg, N. (1987). *Are Male and Female Professors Rated Differently?* **Journal of Educational Psychology** 79,3: 308-314.

Bending, A.W. (1952). *A Preliminary Study of the Effect of Academic Level, Sex and Course Variables on Student Ratings of Psychology Instructors.* **Journal of Psychology**. 34, 21-26.

Bluemann, J. (1981). *Towards a Homo-social Theory of Sex Roles: An Explanation of Sex Segregation of Social Institutions.* In Blaxall Me Reagun Bb (eds).

Cashin, W.E. and Downey, R.G. (1992). *Using Global Rating Items for Summative Evaluation.* **Journal of Educational Psychology**. 84, 563-572.

Cashin, W.E. and Perrin, B.M. (1983). *Do College Teachers who Voluntarily have Courses Evaluated Receive Higher Student Ratings?* **Journal of Educational Psychology**. 75, 595-602.

Centra, J.A. (1979). **Determining Faculty Effectiveness**. Jossey – Bass, San Francisco, California.

Cochran, W.G. (1977).**Sampling Techniques**. Oxford:OUP.

Cross, D. (1989). *Observation and Teacher Evaluation.* In **Forum Anthology**, Vol. 4., 266-268, English Language Programs Division. Bureau of Educational and Cultural Affairs. United States Information Agency. Washington DC.

Darge Wole.(2000). *Patterns of Student Absenteesm in Addis Ababa Government High Schools and Considerations for Containment.* **The Ethiopian Journal of Educational Research**, 20, 1,59-90

Department of Foreign Languages and Literature. (2000). Revised Curriculum for Undergraduate and Postgraduate Programs.

Ducett, J. and Kennedy, 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 Pohlmann, J. (1978) *Effect of Teacher Student and Class Characteristics on the Evaluation of College Instructors.* **Journal of Educational Psychology**. 70, 2, 187-192.

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

Finochiaro, M. (1989). *Teacher Development : A Continuing Process.* In **Forum Anthology**, Vol.4., 269-273, English Language Programs Division. Bureau of Educational and Cultural Affairs . United States Information Agency. Washington DC.

Freeman, H. (1994). *Student Evaluations of College Instructors. Effect of Type of Course Taught, Instructor Gender and Gender Role, and Student Gender.* **Journal of Educational Psychology**. 86, 4, 627-630.

Frey, P.W. (1978). *A Two Dimensional Analysis of Student Ratings of Instruction.* **Research in Higher Education**. 9, 69-91.

Gigliotti, R and Buchtel, F. (1990). *Attributional Biases and Course Evaluations.* **Journal of Educational Psychology,** 82,2,341-351

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

Hollander,M. and Wolfe,D. (1999). **Nonparametric Statistical Methods**. John Wiley and Sons, New York, NY 2nd Edition.

Husbands, C. and Fosh, P. (1993). *Student Evaluation of Teaching in Higher Education: Experiences from Four European Countries and Some Policy*

*Implications of the Practice*. **Assessment and Evaluation in Higher Education**. 18, 2, 95-113.

International Encyclopedia of Education. (1994). Vol.4. London.

Kaschak, E. (1978). *Sex Bias in Student Evaluations of College Professors*. **Psychology of Women Quarterly**. 3, 235-243.

Kennedy, W.R. (1975). *Grades Expected and Grades Received – Their Relationship to Students' Evaluations of Faculty Performance*. **Journal of Educational Psychology**. 1975, 67,109-115.

Kierstead et. al. (1988). *Sex Role Stereotyping of College Professors: Bias in Student Ratings of Instructors*. **Journal of Educational Psychology**. 80, 3, 342-344.

Krashen, S. (1981). **Second Language Acquisition and Second Language Learning**. Pergamon.

Kwo, O. (1989). *Towards Self-evaluation: A Framework for Teacher Development*. In **Forum Anthology**, Vol.4, 274-282, English Language Programs Division. Bureau of Educational and Cultural Affairs, United States Information Agency. Washington DC.

Lindsey, J.K.(1999). **Models for Repeated Measurements**. Oxford: OUP.

Lombardo, J. and Tocci, M. (1979). *Attribution of Positive and Negative Characteristics of Instructors as a Function of Attractiveness and Sex of Instructor and Sex of Subject*, **Perceptual and Motor Skills**. 48, 491-494.

Marsh, H.W. (1987). *Students Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research* [Special Issue] **International Journal of Educational Research**. 11, 253-388.

_____. and Dunkin, M.J. (1992). *Student Evaluations of University Teaching: A Multi-Dimensional Perspective*. In J. Smart (ed.) **Higher Education: Handbook of Theory and Research**. (Vol. 8, pp 143-233) New York: Agathon.

_____. (1995). *Student Evaluation of Teaching*. In T.H. Husen and T.N. Postlethwaite (eds). **International Encyclopedia of Education**. Oxford: Pergamon Press.

Marsh, H., and Hau, K. (1997). *Student Evaluations of University Teaching: Chinese Version of Students' Evaluations of Educational Quality*. **Journal of Educational Psychology**. 89,3, 568-572.

Mulugeta Gebreselassie (1997). *Regional Differences Effect on the Performance of Freshman Students at Addis Ababa University*. **The Ethiopian Journal of Education**, Volume XVII, No. 1, June 1997, Addis Ababa.

Lindsey, J.K. (1999). *Models for Repeated Measurements*. **Oxford Statistical Science Series**. Oxford University Press.

Patton, MQ. (1986). **Utilization Focused Evaluation**, 2nd edition. Sage, Newbury Park.

Pohlmann, J.T. (1975). *A Multivariate Analysis of Selected Class Characteristics and Student Ratings of Instruction*. **Multivariate Behavioural Research**. 1975, 81-91.

Roberts, C. and Roberts, J. (1994). **Evaluation in ELT**. Oxford: Blackwells.

Rotem, A. and Glasman, N. (1979). *On the Effectiveness of Students' Evaluative Feedback to University Instructors*. **Review of Educational Research**. 49, 3, 497-511.

Scriven, M. (1991). **Evaluation Thesaurus**. 4th edition. Sage: Newbury Park, California.

Tirusew Tefera. (1998). *Issues Surrounding the Academic Efficiency of Addis Ababa University*. **Flambeau**, 5; 1, 19-35.

Watkins, D. (1994). *Student evaluations of Teaching Effectiveness: A cross-cultural Perspective*. **Research in Higher Education**. 35, 251-266.

Worth, R. and Sanders, R. (1987). **Educational Evaluation: Alternative Approaches and Practical Guidelines**. Longman, New York.

## Annex 1: Frequency Tables for each variable

| Variable | Values | Frequency | Percentage |
|---|---|---|---|
| Sex | Male | 226 | 77.7 |
| | Female | 65 | 22.3 |
| | Total | 291 | |
| Program | Regular | 144 | 49.0 |
| | Extension | 150 | 51.0 |
| | Total | 294 | |
| Class time | Morning | 107 | 42.6 |
| | Afternoon | 144 | 57.4 |
| | Total | 251 | |
| Level | Degree | 251 | 85.7 |
| | Diploma | 42 | 14.3 |
| | Total | 293 | |
| Region | Out of Addis Ababa | 101 | 35.3 |
| | Addis Ababa | 185 | 64.7 |
| | Total | 286 | |
| Q1 | Extremely valuable | 65 | 21.1 |
| | Very valuable | 111 | 36.0 |
| | Moderately valuable | 87 | 28.2 |
| | Slightly valuable | 34 | 11.0 |
| | Not at all valuable | 9 | 2.9 |
| | Unresponded | 2 | 0.6 |
| | Total | 306 | |
| Q2 | Very well organized | 166 | 53.9 |
| | Well organized | 99 | 32.1 |
| | Moderately organized | 36 | 11.7 |
| | Disorganized | 3 | 1.0 |
| | Very disorganized | 4 | 1.3 |
| | Total | 308 | |
| Q3 | Excellent | 175 | 56.8 |
| | Very good | 77 | 25.0 |
| | Good | 41 | 13.3 |
| | Fair | 8 | 2.6 |
| | Poor | 4 | 1.3 |
| | Unresponded | 3 | 1.0 |
| | Total | 305 | |
| Q4 | Very much | 137 | 44.5 |
| | Quite a lot | 74 | 24.0 |
| | Moderately | 70 | 22.7 |
| | A little | 17 | 5.5 |
| | Not at all | 8 | 2.6 |
| | Unresponded | 2 | 0.6 |
| | Total | 306 | |
| Q5 | Very much | 209 | 67.9 |
| | Quite a lot | 60 | 19.5 |
| | Moderately | 27 | 8.8 |
| | A little | 4 | 1.3 |
| | Not at all | 1 | 0.3 |
| | Unresponded | 7 | 2.3 |
| | Total | 308 | |
| Q6 | Always | 120 | 39.0 |
| | Often | 93 | 30.2 |
| | Sometimes | 86 | 27.9 |
| | Rarely | 4 | 1.3 |
| | Never | 3 | 1.0 |
| | Unresponded | 2 | 0.6 |
| | Total | 306 | |

| Variable | Values | Frequency | Percentage |
|---|---|---|---|
| Q7 | Very helpful | 184 | 59.7 |
| | Helpful | 84 | 27.3 |
| | Moderately helpful | 30 | 9.7 |
| | Rather unhelpful | 5 | 1.6 |
| | Very unhelpful | 1 | 0.3 |
| | Unresponded | 4 | 1.3 |
| | Total | 308 | |
| Q8 | Always | 186 | 60.4 |
| | Often | 58 | 18.8 |
| | Sometimes | 41 | 13.3 |
| | Rarely | 18 | 5.8 |
| | Never | 2 | 0.6 |
| | Unresponded | 3 | 1.0 |
| | Total | 308 | |
| Q9 | Very fair | 96 | 31.2 |
| | Fair | 120 | 39.0 |
| | Moderately fair | 53 | 17.2 |
| | Unfair | 23 | 7.5 |
| | Very unfair | 11 | 3.6 |
| | Unresponded | 5 | 1.6 |
| | Total | 308 | |
| Q10 | Very effective | 146 | 47.4 |
| | Effective | 115 | 37.3 |
| | Moderately effective | 33 | 10.7 |
| | Rather ineffective | 6 | 1.9 |
| | Very ineffective | 1 | 0.3 |
| | Unresponded | 7 | 2.3 |
| | Total | 308 | |
| Q11 | A | 31 | 10.1 |
| | B | 82 | 26.6 |
| | C | 175 | 56.8 |
| | D | 6 | 1.9 |
| | F | 1 | 0.3 |
| | Unresponded | 13 | 4.2 |
| | Total | 308 | |
| Q12 | A | 106 | 34.4 |
| | B | 146 | 47.4 |
| | C | 35 | 11.4 |
| | D | 2 | 0.6 |
| | F | 1 | 0.3 |
| | Unresponded | 18 | 5.8 |
| | Total | 308 | |
| Q13 | 0.0-1.0 | 1 | 0.3 |
| | 1.0-2.0 | 55 | 17.9 |
| | 2.0-3.0 | 126 | 40.9 |
| | 3.0-4.0 | 55 | 17.9 |
| | Unresponded | 71 | 23.1 |
| | Total | 308 | |

**Annex 2. Descriptive Statistics**

| VARIABLE | N | Mean | Std. Deviation |
|---|---|---|---|
| QA1 | 306 | 3.6176 | 1.0311 |
| QA2 | 308 | 4.3636 | .8256 |
| QA3 | 305 | 4.3475 | .9017 |
| QA4 | 306 | 4.0294 | 1.0662 |
| QA5 | 301 | 4.5681 | .7390 |
| QA6 | 306 | 4.0556 | .9018 |
| QA7 | 304 | 4.4638 | .7656 |
| QA8 | 305 | 4.3377 | .9635 |
| QA9 | 303 | 3.8812 | 1.0544 |
| QA10 | 301 | 4.3256 | .7748 |
| Grand Average | 308 | 4.1999 | .5459 |

**Annex 3**

Addis Ababa University
Department of Foreign Languages and Literature
College English II (FLEn 102)

Dear Student,

This form gives you an opportunity to indicate your reaction to this course and the way it has been taught. Student opinion is a valuable guide in course planning and in evaluating teaching.

In the questions below, the word 'course' refers to College English Two, which you have taken this semester. Please DO NOT write your name.

When considering the questions, please try not to let your overall reaction to the course prevent you from noting areas of strength or weakness. Circle the number which best indicates your reaction.

Section Sex Program Class Time
Level Place of Origin

Q1. Overall, how valuable do you think this course has been for you in terms of helping you to communicate effectively in the language?
1. Extremely valuable
2. Very valuable
3. Moderately valuable
4. Slightly valuable
5. Not at all valuable

Q2. How well organized have you found your instructor's contribution to this course?
1. Very well organized
2. Well organized
3. Moderately well organized
4. Disorganized
5. Very disorganized

Q3. How would you rate your instructor's ability to communicate ?
1. Excellent

2. Very good
3. Good
4. Fair
5. Poor

Q4. How much has the instructor stimulated your interest in the field?
1. Very much
2. Quite a lot
3. Moderately
4. A little
5. Not at all

Q5.Did the instructor look prepared before coming to class?
1. Always
2.Often
3. Sometimes
4. Rarely
5. Never

Q6. Students could discuss or debate with each other in groups or pairs freely ?
1. Always
2. Frequently
3. Sometimes
4. Rarely
5.Never

Q7. How would you describe your instructor's attitude towards students in this course?
1. Very helpful
2.Helpful
3.Moderately helpful
4.Rather unhelpful
5.Very unhelpful

Q8. Were tests/ assignments/exams marked and returned promptly ?
1.Always
2.Often
3.Sometimes
4.Rarely
5.Never

Q9. Was the instructor fair in marking assignments and tests?
1.Very fair

2. Fair
3. Moderately fair
4. Unfair
5. Very unfair

Q10. Overall, how effective have you found your instructor in teaching the course ?
1. Very Effective
2. Effective
3. Moderately effective
4. Rather ineffective
5. Very ineffective

Q11. Your college English grade in first semester
1.A 2.B 3.C 4.D 5.F
Q12. Grade you expect in college English 2
1.A 2.B 3. C 4. D 5.F
Q13. Your first semester GPA _____

Thank You and Good Luck