# A Report on the Training of Markers

Teshome Demisse

## ABSTRACT

This paper reports on the attempts made by the Department of Foreign Languages and Literature to promote the training of markers in the scoring of open-ended writing tasks.

A set of data from a training workshop is analysed according to a criterion to investigate the extent of consistency with which markers scored student scripts.

The overall picture gained from the anlaysis shows that there is a reasonably acceptable level of agreement: 67%, 70%, and 68% uniform marking was observed in sessions one, two, and three respectively. In addition. the examiners marked the scripts with slightly better consistency when using the analytic instead of the impressionistic marking scheme.

The results of the analysis also show cases of deviations from the group consensus and fluctuations between strict and lenient marking.

Given the level of agreement and the cases of deviations and fluctuations observed, it is noted that there is still room for improvement in uniform marking. In other words, the results of the analysis in this paper clearly show the need for continued concern in marker reliability.

## 1. Introduction

This paper reports on the attempt to standardise the marking of open-ended writing tasks.

The Department of Foreign Languages and Literature has organised several English Language Teaching (ELT) workshops with the aim of updating the quality of instructors teaching College English and achieving uniformity in marking open-ended writing tasks.

I was actively involved in an ELT workshop entitled "Testing the Skills" in Novemeber 1993, for example. As a member of the College English Testing Committee, I was also given the assignment to conduct a workshop on the training of markers in January 1997. This report is based on this latter workshop.

The makring of open-ended writing tasks involves subjective judgement; and for this reason, there is the likelihood of disparity in marking between and among markers.

Thus, the aim of the workship was to investigate the extent of this disparity as well as to provide markers with the necessary training.

## 2. Reivew of related literature

Test developers and test writers as well as test takers all make subjective decisions about tests. The test developer, based on the best information available to her/him, subjectively determines the content of the test whereas the test writer subjectively decides on the best way to construct the test items. The test takers, too, make subjective decisions to determine the best way to answer the question (Bachman, 1990:76).

Many authorities in the field, for example, Harrison (1983), Madsen (1983), Weir (1988), Baker (1989), Hughes (1989) and Heaton (1990), discuss the subjectivity involved in the assessment of the productive skills of speaking and writing. In view of the problem, the use of a rating scale (global and/or holistic scoring) is suggested (Harrison, 1983; Madsen, 1983; Heaton, 1990). Furthermore, methods of checking and achieving desirable levels of inter-and intra-rater or marker reliability are also recommended (Weir, 1988; Baker, 1989; Hughes, 1989).

More specifically, subjective tests differ from objective tests mainly in terms of the scoring procedure; i.e., the marker's judgement is involved in the scoring of the response to subjective test items.

> Tests such as the oral interview or the written composition that involve the use of rating scales are necessarily subjectively scored, since there is no feasible way to 'objectify' the scoring procedure (Bachman, 1990:76).

Focusing on written student scripts, the consistency with which markers award scores to written compositions is important to test reliability as well as the scores (as dependable estimates of the performance of the students). Weir expresses the concern thus:

> If we have different markers for a writing test will they arrive at the same results? What steps can we take to ensure that different markers will give the same picture of somebody's ability, so that they can maintain consistency in their own standards of marking from the first to the last piece of written work? The closer the agreement in these matters, the more reliable a test (1995:20).

The markers need to achieve consistency both in their own marking and with other markers. One way to minimise inconsistency (or improve marker reliability) is to conduct standardisation sessions in marking written scripts. According to Weir, "The purpose of standardisation procedures is to bring examiners into line, so that candidates' marks are affected as little as possible by the particular examiner who assesses them" (1995:26).

Much effort is normally put into the design and construction of tests: for example, the test specification needs to reflect the goals of the institution, the aims of the course (syllabus) and the items need to be pretested. The success in all this effort can only be complete when there is faith in the marks that the examiners give the candidates. Alderson, Clapham, and Wall state the importance of training examiners thus:

> The training of examiners is a crucial component of any testing programme, since if the marking of a test is not valid and reliable, then all of the other work undertaken earlier to construct a 'quality' instrument will have been a waste of time. (1995:105).

For Mathews, too, measurement "implies a standardised instrument of assessment and an operative who can consistently apply it" (1985:90).

The concern of the Department of Foreign Languages and Literature is, therefore, to standardise both the instrument of assessment and the application of it by our staff members (examiners).

## 3. Plan of work for the workshop

Three sets of actual student scripts were identified, photocopied and given codes to be used in three sessions of the training.

Staff members teaching College English were organised in groups of four to six participants. They, too, were given codes.

Three sets of score sheets were used for the three sessions of marking.

The markers were not allowed to write any mark (sign) on the scripts. Instead, they were required to read the script and marker codes on the score sheet.

For example, a group of six markers were given a set of six scripts for judging, i.e., each script would be judged by the six members.

It was also decided that the first session would be based on impressionistic marking, and the latter two sessions on analytic marking re markers would have to refer to criteria provided in the marking This latter method of marking is contained in the teachers' manual ollege English.

procedure in these sessions was that instructors first mark the ts, then discuss their marking, especially if there happend to be rences in scores for certain scripts.

### Discussions during the sessions

One issue discussed during the sessions was the limit of difference or the extent of uniformity. In this respect, it was suggested that the average score or the consensus score could be used as a criterion to determine the limit of difference. So, it was agreed that if a particular score of a marker for a script is away (below or above) only by one mark from the average score of the group for the same script, it should be considered tolerable. But if the difference is greater than one mark from the average score, then this should signal disparity in marking between and among the markers. The extreme scores should then be discussed to level out the difference.

At the end of each session, some sample group performances were analysed on the spot for the benefit of discussion amongst the whole workshop participants.

The workshop was then closed with the following observations:

- the disparity in marking was not as much as feared.
- the participants marked relatively more uniformly with the analytic marking key than when marking impressionistically.
- It was also agreed that more of such training sessions would be

    writing tasks.

## 5. Objective

The objective of this paper is to analyse the same data fully using a more stringent criterion.

One smaple of group performance from each session was analysed during the workshop whereas eleven groups' performances are treated in this paper. Moreover, it was agreed to tolerate a difference of one mark away from the average score in the workshop. I was not at ease with this consensus. This is because a one-mark difference from the average score could be a two-mark difference between two markers. For instance, a group of four markers could award 3/10, 4/10, 4/10 and 5/10 for the same script. The average score is clearly 4/10. Given the consensus of the workshop all four scores should be acceptable.

But according to this writer, only three of the scores are tolerable because there is a difference of two marks between the first (3/10) and the last (5/10) scores. Perhaps, there is also reason to accept the latter three scores and question the first (3/10) depending on the situation. For example, if it is a classroom (outside examination setting) task, a part of continuous assessment, the decision would probably have a motivational value for the students receiving feedback on the writing task.

## 6. The Performance of the groups

### 6.1 Session one

### 6.1.1. Group A

Table 1: Scores (S), Means (M), ranges (R), and Mean Deviations (MD)

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | XOY | S18 | X25 | U30 | U35 | A38 | M | R |
| 101 | S | 5 | 5.5 | 6 | 7 | 6 | 5 | 5.75 | 5-7 |
| | MD | -0.75 | -0.25 | 0.25 | 1.25 | 0.25 | -0.75 | | |
| 102 | S | 6 | 4 | 6 | 6 | 4 | 5 | 5.17 | 4-6 |
| | MD | 0.83 | 1.17 | 0.83 | 0.83 | 1.17 | -0.17 | | |
| 103 | S | 3 | 4 | 3 | 6 | 5 | 3 | 4 | 3-6 |
| | MD | -1 | 0 | -1 | 2 | 1 | -1 | | |
| 104 | S | 5 | 6 | 7 | 8 | 6 | 6 | 6.33 | 5-8 |
| | MD | -1.33 | -0.33 | 0.67 | 1.67 | -0.33 | -0.33 | | |
| 107 | S | 2 | 4" | 4 | 7 | 4 | 3 | 4 | 2-7 |
| | MD | -2 | 0 | 0 | 3 | 0 | -1 | | |
| 110 | S | 4 | 4 | 4 | 5 | 4 | 3 | 4 | 3-5 |
| | MD | 0 | 0 | 0 | 1 | 0 | -1 | | |
| | M | 4.17 | 4.58 | 5 | 6.5 | 4.8 | 4.17 | | |
| | R | 2-6 | 4-6 | 3-7 | 5-8 | 4-6 | 3-6 | | |

This group is composed of six members. Two of the markers, XOY and X25, hasve the widest range, ie, they were using a range of five points on the scale 0-10 during marking whereas two others, U35 and S18, have narrow ranges of three points on the same scale.

The marking of U30 shows wider difference with the average four times. In these cases we see that s(he) marks (awards scores) leniently. The worst instance is evident in the scores for script 107: marker U30 awards seven marks when the average score is only four marks, and when 50% of the markers actually award four marks for the same script.

The group performed at a good level of agreement 26 times out of 36 possible instances, and this shows about 72% uniform marking. This same group achieved less than perfect agreeement in marking two scripts, 101 and 110, out of the six. That is, five markers (out of six) agreed in the scores they awarded to each of the two scripts.

6.1.2. Group B

Table 2: Scores (S), Means (M), ranges (R), and Mean Deviations(MD)

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | N43 | I59 | U36 | XOP | H55 | H58 | M | R |
| 101 | S | 5 | 6 | 4 | 3 | 3.5 | 5 | 4.42 | 3-6 |
| | MD | 0.58 | 1.58 | -0.42 | -1.42 | -0.92 | 0.58 | | |
| 102 | S | 4 | 4 | 5 | 7 | 3.5 | 5 | 4.75 | 3.5-7 |
| | MD | -0.75 | -0.75 | 0.25 | 2.25 | -1.25 | 0.25 | | |
| 105 | S | 5 | 5 | 7 | 2 | 5.5 | 6 | 5.08 | 2-7 |
| | MD | -0.08 | -0.08 | 1.92 | -3.08 | 0.42 | 0.92 | | |
| 107 | S | 3 | 1 | 2 | 4 | 2.5 | 4 | 2.75 | 1-4 |
| | MD | 0.25 | -1.75 | -0.75 | 1.25 | -0.25 | 1.25 | | |
| 108 | S | 4 | 3 | 3 | 5 | 3.5 | 4 | 3.75 | 3-5 |
| | MD | 0.25 | -0.75 | -0.75 | 1.25 | -0.25 | 0.25 | | |
| 110 | S | 4 | 3 | 6 | 6 | 2 | 3 | 4 | 2-6 |
| | MD | 0 | -1 | 2 | 2 | -2 | -1 | | |
| | M | 4.17 | 3.67 | 4.5 | 4.5 | 3.42 | 4.5 | | |
| | R | 3-5 | 1-6 | 2-7 | 2-7 | 2-5.5 | 3-6 | | |

There are also six markers in this group. Three of the markers (I59, U36 and XOP) used the widest range (6 points) whereas one marker (N43) used the narrower range of three points.

The marking of XOP shows wider difference from the average six times. In these cases we notice that her/his marking fluctuates between being lenient (4 times) and strict (2 times).

The worst instance is evident in the scores for script 105: marker XOP awards two marks when the average score is 5.08. Needless to say, no student would like her/his work to be evaluated by this marker.

The group performed at a tolerable level of agreement 22 times out of 36 possible instances which shows about 61% uniform marking. This group also achieved less than perfect agreement in marking script 108.

### 6.1.3 Group C

Table 3: Scores (S), Means (M), ranges (R), and Mean Deviations (MD)

| | | **Marker** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Script** | | A29 | T39 | B03 | B50 | U23 | M | R |
| 101 | S | 6 | 5 | 4 | 6 | 3 | 4.8 | 3-6 |
| | MD | 1.2 | 0.2 | -0.8 | 1.2 | -1.8 | | |
| 102 | S | 6 | 5 | 5 | 6 | 4.5 | 5.3 | 4.5-6 |
| | MD | 0.7 | -0.3 | -0.3 | 0.7 | -0.8 | | |
| 103 | S | 4.5 | 2 | 3 | 4 | 3.5 | 3.4 | 2-4.5 |
| | MD | 1.1 | -1.4 | -0.,4 | 0.6 | 0.1 | | |
| 104 | S | 6.5 | 5.5 | 8 | 7.5 | 8 | 7.1 | 5.5-8 |
| | MD | -0.6 | -1.6 | 0.9 | 0.4 | 0.9 | | |
| 106 | S | 4 | 3.5 | 3 | 5 | 3 | 3.7 | 3-5 |
| | MD | 0.3 | -0.2 | -0.7 | 1.3 | -0.7 | | |
| 110 | S | 4 | 3 | 6 | 5.5 | 5 | 4.7 | 3-6 |
| | MD | -0.7 | -1.7 | 1.3 | 0.8 | 0.3 | | |
| | M | 5.17 | 4 | 4.8 | 5.67 | 4.5 | | |
| | R | 4-6.5 | 2-5.5 | 3-8 | 4-7.5 | 3-8 | | |

This group is composed of five markers, two of which (B03 and U23) were using a range of six points while one (A29) used a narrow range of four points on the ten-point scale.

The marking of T39 shows wider difference with the average three times, and in these case s/he was marking more strictly. In other words, the scores s/he awarded were below the average score of the group.

The group performed at an acceptable degree of agreement 18 tim of 30 possible instances, which amounts to a 60% uniform marking group achieved less than perfect agreement in two cases, ie, i marking of scripts 102 and 106.

6.1.4 Group D

Table 4: Scores (S), Means (M), ranges (R), and Mean Deviation

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | 001 | C04 | F08 | 009 | C37 | U22 | M | R |
| 101 | S | 6 | 4 | 4 | 5 | 7 | 2 | 4.67 | 2-7 |
| | MD | 1.33 | -0.67 | -0.67 | 0.33 | 2.33 | -2.67 | | |
| 102 | S | 6 | 6 | 6 | 7 | 6 | 5 | 6 | 5-7 |
| | MD | 0 | 0 | 0 | 1 | 0 | -1 | | |
| 103 | S | 5 | 4 | 4 | 4 | 5 | 2 | 4 | 2-5 |
| | MD | 1 | 0 | 0 | 0 | 1 | -2 | | |
| 105 | S | 5 | 7 | 4 | 6 | 7 | 7 | 6 | 4-7 |
| | MD | -1 | 1 | -2 | 0 | 1 | 1 | | |
| 109 | S | 6 | 5 | 5 | 5 | 7 | 3 | 5.17 | 3-7 |
| | MD | 0.83 | -0.17 | -0.17 | -0.17 | 1.83 | -2.17 | | |
| 110 | S | 5 | 5 | 5 | 4 | 5 | 3 | 4.5 | 3-5 |
| | MD | 0.5 | 0.5 | 0.5 | -0.5 | 0.5 | -1.5 | | |
| | M | 5.5 | 5.17 | 4.67 | 5.17 | 6.17 | 3.67 | | |
| | R | 5-6 | 4-7 | 4-6 | 4-7 | 5-7 | 2-7 | | |

Out of the six markers in this group, U22 marked with a range of six points while marker 001 judged the scripts with a two-point range.

It can be seen that there is a clear difference between the scores of marker U22 and the average score of the group four times. marker U22 under-scored scripts 101,103,109 and 110.

The group performed with a good level of agreement 26 times out of 36 possible instances, and this is a 72% uniform marking. It also achieved less than perfect agreement in three cases (scripts 102,103 and 110).

Generally, the analysis of the performance of the four groups during the first session shows that there was about 67% uniform marking of the scripts.

### 6.2. Session Two

### 6.2.1 Group A

Table 5: Scores (S), Means (M), ranges(R), and Mean Deviations (MD)

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | XOY | S18 | X25 | U30 | U35 | A38 | M | R |
| 201 | S | 5 | 5.8 | 5.5 | 5.5 | 7 | 8 | 6.13 | 5-8 |
| | MD | -1.13 | -0.33 | -0.63 | -0.63 | 0.87 | 1.87 | | |
| 202 | S | 5 | 5 | 4 | 4.5 | 5.5 | 5.5 | 4.92 | 4-5.5 |
| | MD | 0.08 | 0.08 | -0.92 | -0.42 | 0.58 | 0.58 | | |
| 203 | S | 4 | 6 | 6 | 6 | 5.5 | 6 | 5.58 | 4-6 |
| | MD | -1.58 | 0.42 | 0.42 | 0.42 | -0.08 | 0.42 | | |
| 204 | S | 6 | 6 | 4 | 5.5 | 7.5 | 4.5 | 5.58 | 4-7.5 |
| | MD | 0.42 | 0.42 | -1.58 | -0.08 | 1.92 | -1.08 | | |
| 205 | S | 6 | 5.5 | 5.5 | 4.5 | 6 | 5 | 5.42 | 4.5-6 |
| | MD | 0.58 | 0.08 | 0.08 | -0.92 | 0.58 | -0.42 | | |
| 206 | S | 7 | 5.6 | 7.5 | 5.5 | 7 | 7.5 | 6.68 | 5.5-7.5 |
| | MD | 0.32 | -1.08 | 0.82 | -1.18 | 0.32 | 0.82 | | |
| | M | 5.5 | 5.65 | 5.42 | 5.25 | 6.42 | 6.08 | | |
| | R | 4-7 | 5.5-6 | 4-7.5 | 4.5-6 | 5.5-7 | 4.5-8 | | |

This group is formed of six markers, two of which (A38 and X the widest range of five points on the ten-point scale, and S1 narrowest range of half a point in marking all the six scripts.

There is a wide deviation from the average score in the markin and A38 twice: while XOY tended to mark relatively stri fluctuated between marking leniently and strictly.

This group carried out the task with a good level of agreement 26 times out of 36 possible instances, which is a 72% uniform marking.

Group a achieved less than perfect agreement in three cases (202, 203, and 205) during this second session.

### 6.2.2. Group B

Table 6: Scores (S), Means (M), Ranges (R), and Mean Deviations (MD)

| | | Marker | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Script | | N43 | I59 | U36 | H55 | H58 | M | R |
| 201 | S | 5 | 7 | 6 | 5.5 | 4 | 5.5 | 4-7 |
| | MD | -0.5 | 1.5 | 0.5 | 0 | -1.5 | | |
| 202 | S | 5 | 4 | 4 | 6 | 5.5 | 4.9 | 4-6 |
| | MD | 0.1 | -0.9 | -0.9 | 1.1 | 0.6 | | |
| 203 | S | 4.5 | 5 | 6 | 5.5 | 5.5 | 5.3 | 4.5-6 |
| | MD | -0.8 | -0.3 | 0.7 | 0.2 | 0.2 | | |
| 204 | S | 5 | 5 | 6.5. | 4.5 | 3 | 4.8 | 3-6.5 |
| | MD | 0.2 | 0.2 | 1.7 | -0.3 | -1.8 | | |
| 205 | S | 6 | 7 | 4 | 4 | 4 | 5 | 4-7 |
| | MD | 1 | 2 | -1 | -1 | -1 | | |
| 206 | S | 6.5 | 7 | 7 | 6 | 6.5 | 6.6 | 6.5-7 |
| | MD | -0.1 | 0.4 | 0.4 | -0.6 | 0.1 | | |
| | M | 5.33 | 5.83 | 5.58 | 5.25 | 4.75 | | |
| | R | 4.5-6.5 | 4-7 | 4-7 | 4-6 | 3-6.5 | | |

From the five markers in this group, H58 used the widest range of five points whereas H55 Used a narrower range of three.

The marking of I59 and H58 deviates from the average score twice: while the former tended to be lenient, the latter tended to be strict.

The group carried out the task with a good level of agreement 21 times out of 30 possible instances, which indicates a 70% uniform marking. There is perfect agreement in the marking of script 206 for this group. That is, all the (five) markers agreed in the scores they awarded to this script.

### 6.2.3. Group C

Table 7: Scores (S), Means (M), ranges (R), and mean Deviations (MD)

| | | Marker | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Script | | A29 | T39 | B03 | B50 | U23 | M | R |
| 201 | S | 5.5 | 5 | 5.5 | 5 | 5 | 5.2 | 5-5.5 |
| | MD | 0.3 | -0.2 | 0.3 | -0.2 | -0.2 | | |
| 202 | S | 5 | 4.5 | 5.5 | 3.5 | 6 | 4.9 | 3.5-6 |
| | MD | 0.1 | -0.4 | 0.6 | -1.4 | 1.1 | | |
| 204 | S | 5.5 | 5.5 | 6.5 | 6 | 6 | 5.9 | 5.5-6.5 |
| | MD | -0.4 | -0.4 | 0.6 | 0.1 | 0.1 | | |
| 205 | S | 6.5 | 7 | 6 | 6.5 | 5.5 | 6.3 | 5.5-7 |
| | MD | 0.2 | 0.7 | -0.3 | 0.2 | -0.8 | | |
| | M | 5.63 | 5.5 | 5.88 | 5.25 | 5.63 | | |
| | R | 5-6.5 | 5-7 | 5.5-6.5 | 3.5-6.5 | 5-6 | | |

Of the five markers in this group, marker B50 used the widest range of five points, but markers B03 and U23 used the narrowest range of one point in marking four scripts.

Markers B50 and U23 deviate once from the average score of the group in awarding marks. This deviation indicates that B50 was strict·whereas U23 was lenient.

The group performed at a very good level of agreement 17 tir 20 possible instances, and this amounts to 85% uniform Moreover, the group achieved perfect agreement in marking sc and 204. The level of consistency this group achieved is the be groups in session two.

### 6.2.4. Group D

Table 8: Scores (S), Means (M), Ranges (R), Mean Deviations (MD)

| | | Marker | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Script | | 001 | F08 | 009 | C37 | U22 | M | R |
| 201 | S | 5.5 | 7.5 | 8.5 | 6.5 | 6 | 6.8 | 6-8.5 |
| | MD | -1.3 | 0.7 | 1.7 | -0.3 | -0.8 | | |
| 202 | S | 5.5 | 4 | 8 | 3.5 | 5.5 | 5.3 | 3.5-8 |
| | MD | 0.2 | -1.3 | 2.7 | -1.8 | 0.2 | | |
| 204 | S | 7 | 6.5 | 6 | 5 | 5 | 5.9 | 5-7 |
| | MD | 1.1 | 0.6 | 0.1 | -0.9 | -0.9 | | |
| 205 | S | 7 | 7 | 5.5 | 5.5 | 6 | 6.2 | 5.5-7 |
| | MD | 0.8 | 0.8 | -0.7 | -0.7 | -0.2 | | |
| 206 | S | 6 | 7 | 9 | 7.5 | 6.5 | 7.2 | 6-9 |
| | MD | -1.2 | -0.2 | 1.8 | 0.3 | -0.7 | | |
| | M | 6.2 | 6.4 | 7.4 | 5.6 | 5.8 | | |
| | R | 5.5-7 | 4-7 | 5.5-9 | 3.5-7.5 | 5-6.5 | | |

Five markers rated five scripts in this group. the widest range of six points was used by marker C37, but markers 001 and U22 used the narrower range of three points.

Markers 001 and 009 deviate three times from the average score of the group. this deviation shows that marker 009 was consistently lenient, but that 001 fluctuated between being strict and lenient.

The performance of the group shows only 56% uniform marking, ie, 14 times out of 25 possible instances. while this level of agreement may just be acceptable, we can see that it is the least level of agreement achieved.

Generally, the anlysis of the performance of the four groups during the ession shows that there was about 70% uniform marking of the

## Session Three

### 5.3.1. Group A

Table 9: scores (S), Means (M), ranges (R), and Mean Deviations (MD)

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | 228 | M35 | H46 | N44 | T32 | U36 | M | R |
| 301 | S | 6 | 6.5 | 5 | 5.5 | 6.5 | 3 | 5.42 | 3-6.5 |
| | MD | 0.58 | 1.08 | -0.42 | 0.08 | 1.08 | -2.42 | | |
| 302 | S | 6 | 5 | 5 | 6.5 | 6 | 3 | 5.25 | 3-6.5 |
| | MD | 0.75 | -0.25 | -0.25 | 1.25 | 0.75 | -2.25 | | |
| 303 | S | 5.5 | 5 | 7 | 5.5 | 8 | 4 | 5.83 | 4-8 |
| | MD | -0.33 | -0.83 | 1.17 | -0.33 | 2.17 | -1.83 | | |
| 304 | S | 9 | 7 | 8 | 7.5 | 4.5 | 6 | 7 | 4.5-9 |
| | MD | 2.0 | 0 | 1 | 0.5 | -2.5 | -1.0 | | |
| 306 | S | 4.5 | 5.5 | 4 | 4.5 | 5 | 3 | 4.42 | 3-5.5 |
| | MD | 0.08 | 1.08 | -0.42 | 0.08 | 0.58 | -1.42 | | |
| 307 | S | 4.5 | 6 | 5 | 5.5 | 5.5 | 6 | 5.42 | 4.5-6 |
| | MD | -0.92 | 0.58 | -0.42 | 0.08 | 0.08 | 0.58 | | |
| | M | 5.92 | 5.83 | 5.67 | 5.83 | 5.92 | 4.17 | | |
| | R | 4.5-9 | 5.5-7 | 4-8 | 4.5-7.5 | 4.5-8 | 3-6 | | |

Six instructors participated in marking six scripts in this group. While marker 228 used the widest range of six points, marker U36 used a narrower range of four points.

A very high deviation from the mean score for the group is observed four times in the marking of one instructor (U36). In all the cases, s/he was stricter in awarding scores to the scripts, and this clearly suggests the greater likelihood of students failing unfairly when assessrd by this marker.

The group performed at an accpetable level (22 times out of 36 instances) of uniformity in marking (61%) and the group also achieved less than perfect agreement in marking script 307.

### 6.3.2. Group B

Table 10: Scores (S), Means (M), ranges (R), and Mean Deviations (MD)

| | | Marker | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Script | | A38 | C04 | F08 | 009 | C37 | U22 | M | R |
| 301 | S | 4.5 | 4 | 5.5 | 5.5 | 5.5 | 4.5 | 4.92 | 4-5.5 |
| | MD | -0.42 | -0.92 | 0.58 | 0.58 | 0.58 | -0.42 | | |
| 302 | S | 3.5 | 3.5 | 4 | 4 | 3.5 | 4 | 3.75 | 3.5-4 |
| | MD | -0.25 | -0.25 | 0.25 | 0.25 | -0.25 | 0.25 | | |
| 303 | S | 4 | 5.5 | 7 | 5.5 | 5 | 7 | 5.67 | 4-7 |
| | MD | -1.67 | -0.17 | 1.33 | -0.17 | -0.67 | 1.33 | | |
| 304 | S | 5.5 | 6.5 | 6 | 7 | 8 | 8.5 | 6.92 | 5.5-8.5 |
| | MD | -1.42 | -0.42 | -0.92 | 0.08 | 1.08 | 1.58 | | |
| 305 | S | 2.8 | 4 | 3.5 | 3 | 4 | 1.5 | 3.13 | 1.5-4 |
| | MD | -0.33 | 0.87 | 0.37 | -0.13 | 0.87 | -1.63 | | |
| 306 | S | 3.5 | 6 | 3 | 4 | 5 | 1 | 3.75 | 1-6 |
| | MD | -0.25 | 2.25 | -0.75 | 0.25 | 1.25 | -2.75 | | |
| | M | 3.97 | 4.92 | 4.83 | 4.83 | 5.17 | 4.42 | | |
| | R | 2.8-5.5 | 3.5-6.5 | 3-7 | 3-7 | 3.5-8 | 1-8.5 | | |

In this group, too, six markers were involved in judging six scripts. Marker U22 used the widest range of nine whereas marker A38 used a narrower range of four points.

A high deviation from the mean score is observed four times in the marking of one instructor (U22). S(he) was twice lenient and twice strict, an instance that shows an erratic fluctuation of the marker in awarding scores.

In this light, the use of nine points from the available ten-point scale cannot be taken as a positive quality of marker U22.

The group performed with a relatively good level (24 times out of 36 instances) of uniform marking (67%) and it achieved perfect and less than perfect agreement in marking scripts 302 and 301 respectively.

### 6.3.3. Group C

Table 11: Scores (S), Means (M), Ranges (R), and Mean Deviations (MD)

| | | Marker | | | | | |
|---|---|---|---|---|---|---|---|
| **Script** | | 227 | N13 | P17 | T19 | M | R |
| 301 | S | 6 | 4.5 | 5 | 5 | 5.13 | 4.5-6 |
| | MD | 0.87 | -0.63 | -0.13 | -0.13 | | |
| 302 | S | 6.5 | 5 | 6 | 6.5 | 6 | 5-6.5 |
| | MD | 0.50 | -1 | 0 | 0.5 | | |
| 303 | S | 6.5 | 6.5 | 6.5 | 7.5 | 6.75 | 6.5-7.5 |
| | MD | -0.25 | -0.25 | -0.25 | 0.75 | | |
| 304 | S | 8 | 7.5 | 7.5 | 7 | 7.5 | 7-8 |
| | MD | 0.5 | 0 | 0 | -0.5 | | |
| | M | 6.75 | 5.88 | 6.25 | 6.5 | | |
| | R | 6-8 | 4.5-7.5 | 5-7.5 | 5-7.5 | | |

Four markers were involved in the marking of four scripts in this group. Marker N13 used a wider range of five points, and marker 227 used a narrower range of three points.

There is no serious deviation from the mean score, i.e., the deviations from the mean score did not exceed one mark.

The group performed at the best level of agreement 14 times out of 16 possible instances, and this is 88% uniform marking - the best uniformity observed in the marking exercise of that day. In addition, the group achieved perfect agreement in the marking of two scripts (303 and 304).

Generally, the analysis of the performance of the three groups during the third session shows that there was about 68% uniform marking of the scripts.

## 7. Summary and Conclusion

The combined picture of the groups in each session shows some level of agreement: 67%, 70% and 68% uniform marking was observed in sessions one, two and three respectively. The markers performed better in the last two sessions than in the first, i.e., the percentages suggest that the markers performed with slightly better consistency when using the analytic marking scheme than the impressionistic marking scheme.

On a different count, no perfect agreement was observed during the first session, but there were three such instances in each of the last two sessions. This is also evidence of the likelihood of achieving maximum consistency in marking when using the analytic marking scheme.

It is clear from the foregoing discussion that there is still room for improvement in uniform marking. We need a greater degree of agreement, than has been observed, between and among markers in awarding scores to the same scripts to ascertain reliability in our marking (rating) of open-ended test items.

We have also seen clear cases of deviations from group consensus (average scores) in awarding scores, and serious cases of fluctuation between strict and lenient marking, both of which need to be moderated through such sessions (workshops) of marking written scripts. These deviations and fluctuations signal the need for a rigorous analysis (e.g. rank corelation) of the data in the investigation of consistency in the marking of the scripts.

Finally, although disparity in marking was not as feared - a consensus arrived at during the workshop - the results of the analyses in this paper clearly show the need for continued concern in the marking of open-ended writing tasks.

## References

Alderson, J.C., C. Clapham and D.Wall. 1995. *Languag Construction and Evaluation*. Cambridge: Cambridge Uni Press.

Bachman, L.f. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Baker, D. 1989. *Language Testing* . London: Edward Arnold.

Harrison, a. 1983. *A Language Testing Handbook*. London: Macmillan Publishers.

Heaton, J.B. 1990. *Classroom Testing*. London: Longman.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Madsen, H.S. 1983. *Techniques in Testing*. Oxford: Oxford University Press.

Mathews, J.C. 1985. *Examinations: A Commentary*. London: George Allen and Unwin.

Weir, C.J. 1988. *Communicative Language Testing*. Exeter: University of Exeter.

________ 1995. *Understanding and Developing Language Tests*. New York: Phonix ELT.

**Teshome Demisse**
**"A Small-Scale Evaluation of College English Examination (First Semester Final, 1996/97)"**
**(Issue No. 7, pp. 82-95)**

p. 83, paragraph 1, line 5:

'conscious assessment' should read 'continuous assessment'.

p. 85, paragraph 1 (after table), line 3:

'mixture' should read 'a mixture'.

p. 85, paragraph 1 (after table), line 9:

'(46.69)' should read '(46-69)'

p. 94. Part III, Reading Comprehension, Sec. A:

between items 8 & 9 insert 5.2; 0.83; and 0.33 for Exam Part/Section, Facility Value, and Discrimination Index, respectively.

## NOTES ON CONTRIBUTORS*

***Fekade AZeze*** *(PhD) is Associate Professor in the Department of Ethiopian Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Hussein Ahmed*** *(PhD) is Associate Professor in the Department of History, College of Social Sciences, Addis Ababa University.*

***Askale Lemma*** *( MA) is Lecturer in the Oromo Unit of the Department of Ethiopian Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Hirut Wolde Mariam*** *(MA) is Lecturer in the Department of Linguistics, Institute of Language Studies, Addis Ababa University.*

***Teshome Demisse*** *(PhD) is Assistant Professor in the Department of Foreign Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Daniel Abera*** *(MA) is Lecturer in the Department of Linguistics, Institute of Language Studies, Addis Ababa University*

***[*Contributors are listed in the order in which their respective articles have appeared. We would also like to take this opportunity to remind future contributors (particularly those outside the University) to send us a short bio-data describing their qualifications, rank, and institutional and/or departmental affiliations.]***

## NOTES ON CONTRIBUTORS*

***Fekade AZeze*** *(PhD) is Associate Professor in the Department of Ethiopian Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Hussein Ahmed*** *(PhD) is Associate Professor in the Department of History, College of Social Sciences, Addis Ababa University.*

***Askale Lemma*** *( MA) is Lecturer in the Oromo Unit of the Department of Ethiopian Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Hirut Wolde Mariam*** *(MA) is Lecturer in the Department of Linguistics, Institute of Language Studies, Addis Ababa University.*

***Teshome Demisse*** *(PhD) is Assistant Professor in the Department of Foreign Languages and Literature, Institute of Language Studies, Addis Ababa University.*

***Daniel Abera*** *(MA) is Lecturer in the Department of Linguistics, Institute of Language Studies, Addis Ababa University*

***[*Contributors are listed in the order in which their respective articles have appeared. We would also like to take this opportunity to remind future contributors (particularly those outside the University) to send us a short bio-data describing their qualifications, rank, and institutional and/or departmental affiliations.]***