Performance on a Classroom Test: Interrogating Scores, Time and Grades

Teshome Demisse*

Abstract: The objective of this paper[1] is to report on an analysis of performance on a classroom test, administered at the beginning and end of the semester. A descriptive analysis was made for the purpose of reflection: average performance and variation in scores and the time taken to complete the test at initial and final administrations are worked out. Performance on the test (initial and final scores) is checked for relationship with final course grades. Rank correlations between scores and time taken to complete the test (twice) are computed. The test correlates positively with the final course grades of the students: initial test, r = 0.32; final test, r = 0.68. In about the same average time taken (43 minutes each) for the initial and final test, the average gain in performance was about nine points (8.65). Less variation in scores and time taken was also observed in the final test. The evidence from rank correlations between scores and time taken tend to suggest that they are inversely related: bordering on no relationship (R = 0.082) for the initial and entering negative relationship (R = -0.063) for the final. The positive gain in performance and the relatively moderate correlation of the test with the final course grades is found to be reassuring, while the inverse relationship between the scores and the time taken calls for a thorough investigation on a larger sample.

*Former Associate Professor, Department of Foreign Languages and Literature, Addis Ababa University.

[1] An earlier version of this paper was first presented at the 18th Annual Conference of the Institute of Language Studies held in November 2006 in Addis Ababa.

Introduction

This paper reports on an analysis of data generated from an actual classroom activity conducted along with other requirements such as term papers, classroom presentations, assignments and projects for the course Language Testing (TEFL 622). The course is offered to graduate students (MA in TEFL) in the Department of Foreign Languages and Literature. The test described in this paper was one such classroom activity.

The test was intended to serve as an evaluation, to check on the general awareness of students as developed from reading, input from lectures, and other requirements. Other advantages of giving classroom tests were also entertained: for example, classroom tests "… could be used for the purpose of increasing motivation" (Heaton 1990:10); and more specifically, tests "… serve both to make a rough check on students' progress and to keep students on their toes" (Hughes 1989:13).

In terms of the initial administration, two further advantages were anticipated:

- First, that the test is given during the first class on the very first day of the semester, it was expected to ensure regular attendance, and in the long run improve or reinforce the starting of class on day-one-class-one (Eberly Center 2008).
- Second, since the test attempted to sample the course contents (*or to take samples from the)*, it was expected that it would whet the students' appetite about the course, and that their performance on it would provide them with useful feedback on expectations – their own expectations from the course and what is expected of them for the course (Berkeley Center for Teaching and Learning 2014).

The test covered test purposes, quality of tests, and continuous assessment as contents. It had forty items in four sections. The test formats were such that the responses of the students could be scored objectively. In the first section, the candidates were asked to match the type of test with the purpose of testing from a given list of purposes. In the second section, the candidates were required to complete gaps in four paragraph contexts on issues of reliability and validity. In the third section, they were asked to identify language skills that are better measured by continuous assessment rather than by formal examinations from a list of language skills and/or tasks. In the last section, a list of statements on the definitions and features of test types are given. The candidates were required to write one of the two named (in the instructions) test types against the statements or leave the statement blank if it did not match the two named test types.

This test was then administered as an initial test during the first period of the semester. Next period, the scored scripts were returned to the students followed by a statistical description of the test and discussion of their performance on it. This pedagogic exercise is supported by Madsen (1983:178) :

> For one thing, good evaluation of our tests can help us measure student skills more accurately. It also shows that we are concerned about those we teach. … Students appreciate an extra effort like this, which shows that we are concerned about the quality of our exams. And a better feeling toward our tests can improve class attitude, motivation, and even student performance.

Then at the end of the description and discussion, the scripts were collected from the students. The same test, however, was administered again during the last period of the semester. This was motivated by the desire to learn, among other things, whether test performance would increase while the time it takes to complete it would decrease.

The information from the scored test (including the initial scores on it) was consulted or referred to in the event of doubts and uncertainty when deciding on the final course grade. It was not part of the value for the course work. The total course value was composed of work submitted and evaluated as per the agreement between the two instructors assigned to handle the course as a team.

Apart from the initial analysis for the purpose of classroom description and discussion for the benefit of the students, additional analysis has been conducted for the purpose of reflection.

Objectives

The overall objectives of the study were to assess whether students had developed general awareness in language testing and to explore the relationship between performance (as expressed by scores) and the time it took to complete the test. The objectives were addressed by answering the following questions:

1. Is there a difference between the scores of the students on the test for the two administrations?
2. Is there difference between the times taken to complete the test during the two administrations?
3. What is the correlation between the two sets of scores on the test and the final course grades of the students?
4. What is the relationship between time taken and the scores?

Procedure

In the classroom, the time each student took to complete (do) the test was recorded as the students handed in their work both at the initial and final administration. The test was scored and added up each time. For the purpose of this report, the names were given codes; and the scores, the time taken, and the final grades were tabulated. Then, central tendencies and dispersions of the two sets of scores and times were worked out. Correlation coefficients of the two sets of scores with the final grades were computed. Furthermore, the students were rank ordered according to the time they took to complete the test and according to the scores they earned on both administrations. Then rank correlation coefficients were calculated. The results of these computations are discussed in the next section.

Results and discussion

The test was administered as an initial test during the first period of the semester. It was intended that this test would facilitate comprehension of the objectives of the course as it samples the course description, that it would signal to the students that some serious course work has begun, and that it would also yield information about the students' (prior) knowledge in language testing. When the test was administered again as a final test during the last period of the semester, it was anticipated that there would be some gain in scores. If less or no gain was observed, it was felt that there was need for troubleshooting; and that reflection on the results of the troubleshooting would guide any future course of action. So in this section, the attempt is to answer the questions raised under the objectives.

> *1. Is there any difference between the scores of the students on the test for the two administrations?*

Table 1 displays initial scores, final scores and scores gained thereof in order to answer this question.

As observed in the score columns in Table 1, only two individuals (codes 42 & 45) show no gain in score. All the others show positive gains – ranging from a low of 3 (7.5%) to a high of 17 (42.5%) points. The sum of the scores rises to 541 in the final from 394 in the initial test; and this is a twenty-two percent rise overall at the end of the semester. The average score for the final (31.8) is higher than the average score for the initial test (23.2). This increase in the average score shows an average gain of about nine points, or again about twenty-two percent, by the end of the semester. The standard deviation of the scores for the final test (3.9) is less than that for the initial test (5.6). This difference suggests that the students' responses were relatively stable in the final test. The variation observed in the average scores and the standard deviations is a welcome finding as it indicates a desirable and positive difference.

Table 1: Scores on Initial and Final Administrations (N=17)

Code	**Initial Scores**	**Final Scores**	**Score Gains**
	(40 pts)	**(40 pts)**	
55	31	36	05
45	30	30	00
56	29	36	07
44	28	31	03
48	28	34	06
43	27	40	13
57	27	30	03
41	26	34	08
42	24	24	00
49	24	36	12
51	20	31	11
52	19	28	09
53	18	29	11
50	17	32	15
46	16	27	11
54	16	33	17
47	14	30	16
Total= 17	394	541	147
Average	23.176	31.824	8.647
S. Deviation	5.626	3.941	

2. *Was there a difference between the times taken to complete the test during the two administrations?*

It was considered proper to reflect on the time required to complete the test on both occasions. For instance, it would be good to confirm whether candidates had adequate time for the task, and it would also be interesting to find out whether candidates need different timing for the test administered at two different points during the semester.

Table 2 shows the scores and the time taken to complete the test on the two administrations. The maximum time for the test was sixty minutes, i.e. the regular fifty minutes of the period plus, if found necessary, the ten minutes' break between periods.

Table 2: Scores and Time Taken on Test (N=17)

Code	**Initial Scores (40 pts)**	**Initial Time (60 mins)**	**Final Scores (40 pts)**	**Final Time (60 mins)**
55	31	50	36	39
45	30	40	30	39
56	29	40	36	38
44	28	35	31	38
48	28	46	34	46
43	27	50	40	45
57	27	34	30	42
41	26	46	34	44
42	24	41	24	40
49	24	45	36	43
51	20	50	31	39
52	19	40	28	43
53	18	47	29	52
50	17	45	32	44
46	16	45	27	45
54	16	36	33	45
47	14	40	30	45
Tot = 17	394	730	541	727
Average	23.176	42.941	31.824	42.765
St. Deviation	5.626	5.166	3.941	3.666

Studying Table 2, it can be seen that the longest time taken during the initial administration is fifty minutes and for the final administration it is fifty-two minutes. The shortest time is thirty-four and thirty-eight minutes for the initial and final administrations of the test, respectively. While forty minutes is the most frequent time taken (followed by fifty and forty-five minutes) for the initial test, forty-five minutes is the most frequent (followed by thirty-nine minutes) for the final administration. Here it is interesting to note that the longest, the shortest, and the most frequent times are greater for the second administration. However, the average time taken shows that the candidates took approximately the same amount of time for the two administrations, i.e. 42.941 for the initial and 42.765 for the final administration. Again, the standard deviation of the time taken to complete the final test (3.666) is less than that of the initial test (5.166). The variation in the time taken to complete the tasks levelled off suggesting that the candidates behaved more uniformly during the second administration. The little variation observed in the average time taken and the direction of the variation in the standard deviations of the same is again a welcome finding as it tends to reassure the timing of the test.

3. What is the correlation between the two sets of scores on the test and the final course grades of the students?

The test, when administered at any one or more times during the semester, should show some (positive) relationship with other events and/or activities in the course. When there is evidence of relationship, it can be said that the test is part and parcel of the course, and that it contributes to the objectives of the course. Thus, the two sets of scores on the test are compared with the final course grades as laid out in the following table. It is also useful to note at this point that the scores from the test did not form part of the total course value as set out in the method of course evaluation.

Table 3: Scores on Test and Final Grades (N=17)

Code	Initial Scores (40 pts)	Final Scores (40 pts)	Final Grades (04 pts)
55	31	36	A+
45	30	30	B+
56	29	36	B+
44	28	31	B
48	28	34	A
43	27	40	A
57	27	30	A$^-$
41	26	34	B+
42	24	24	B$^-$
49	24	36	A
51	20	31	B
52	19	28	B
53	18	29	B+
50	17	32	A$^-$
46	16	27	B
54	16	33	B+
47	14	30	B+
Tot = 17	394	541	

Correlation Coefficients

Initial Scores and Final Grades, r = 0.320
Final Scores and Final Grades, r = 0.675

The relationship was examined using a scientific calculator (fx-570c) with a built-in linear correlation formula.

The results of the computation show positive relationships between scores on the test and the final course grades. More specifically, the relationship between the final scores on the test and the final course grades (r = 0.675) is stronger than that between the initial scores on the same test and the final course grades(r = 0.320). The coefficients also suggest that the relationship has improved (increased) as the amount of course events and activities increased. This situation parallels that of concurrent validity, and “most concurrent validity coefficients range from +.5 to +.7” (Alderson et al. 1995:178; Downie and Heath 1974: 244). Thus, the relationship between the final scores on the test and the final course grades (r = 0.675) clearly shows that the test has acceptable concurrent validity.

The relationship between the initial scores on the test and the final course grades (r = 0.320), taking into account the course duration, can be regarded as predictive validity, and the coefficient (r = 0.320) respectable. Regarding the coefficients for this, we can only expect a moderate one - something around +0.4 is generally considered satisfactory (Hughes 1989:25; Kline 1986:5; Downie and Heath 1974:244). The claim, herein, that the test has a respectable predictive validity is further supported by (Alderson, Clapham, and Wall 1995:182) when they write, "In fact, in predictive validity studies, it is common for test developers and researchers to be satisfied when they have achieved a coefficient as low as +0.3!". This, i.e. the examination of relationships, is also a welcome finding as it clearly confirms the anticipated outcome of the plan of work for the course.

4. What is the relationship between time taken and the scores?

It is understood that there is variation of knowledge and abilities among the students at all times. It can also be assumed that there is variation of awareness about language testing before and after the course. Simplistically put, the level of the students’ awareness in language testing is different at the beginning and at the end of the semester. Given some success in accomplishing the course objectives, the students are expected to demonstrate better awareness in language testing than when they first arrived. Thus they are better disposed to respond to the test at the end of the semester than at the beginning. It was considered interesting to interrogate the data set to see the relationship between knowledge/awareness (as expressed in scores) and time taken to complete the test. So the rank ordering of the students in terms of scores and time taken is examined.

Table 4: Rank Orders by Score and Time Taken (N=17**)**

Initial Administration					**Final Administration**			
Code	**Score**	**Rank**	**Time**	**Rank**	**Score**	**Rank**	**Time**	**Rank**
55	31	1	50	2	36	3	39	14
45	30	2	40	12.5	30	12	39	14
56	29	3	40	12.5	36	3	38	16.5
44	28	4.5	35	16	31	9.5	38	16.5
48	28	4.5	46	5.5	34	5.5	46	2
43	27	6.5	50	2	40	1	45	4.5
57	27	6.5	34	17	30	12	42	11
41	26	8	46	5.5	34	5.5	44	7.5
42	24	9.5	41	10	24	17	40	12
49	24	9.5	45	8	36	3	43	9.5
51	20	11	50	2	31	9.5	39	14
52	19	12	40	12.5	28	15	43	9.5
53	18	13	47	4	29	14	52	1
50	17	14	45	8	32	8	44	7.5
46	16	15.5	45	8	27	16	45	4.5
54	16	15.5	36	15	33	7	45	4.5
47	14	17	40	12.5	30	12	45	4.5
Total = 17	394		730		541		727	

Rank Correlation Coefficients
Ranking in Initial Scores and Time Taken, R = 0.082
Ranking in Final Scores and Time Taken, R = ─ 0.063

The rank correlation computation was done by hand using the formula,

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

Where R = coefficient of rank correlation,
D = the difference between paired ranks,
N = the number of pairs(Gupta and Gupta 1995:477).

According to the rank correlation computation, the relationship between the initial scores and the time taken is positive though small (R = 0.082), and that between the final scores and the time taken is negative (R = ─ 0.063). Broadly speaking, the evidence from rank correlations between scores and time taken tend to suggest that they are inversely related: bordering no relationship (R = 0.082) for the initial and entering negative relationship (R = ─ 0.063) for the final. The overall situation of relationship here is either very small or in opposite direction; that is, high scores are not associated with long time taken to complete the tasks. This finding carries the suggestion that the more able and knowledgeable students tend to take less time to complete the test. However, a larger data set is needed to test the truth of the suggestion.

Summary

This study reports on a kind of self-reflective practice given that action research "… is a practical way of looking at your practice in order to check whether it is as you feel it should be." (McNiff 2002:15).

The main motivation in this study was thus to answer questions regarding classroom practice, i.e., whether the practice was proper and relevant in the context of the course. The questions needed to be answered to confirm anticipated outcomes and thereby, among other things, develop confidence in all the stakeholders.

The study showed that students improved their performance as reflected in the increase in the sum of scores and the average scores, and that the students' responses were relatively stable. They also took more or less the same time to complete the test during the two administrations, but they behaved more uniformly in the final administration. This carries the suggestion that the more able and knowledgeable the candidates are the more likely it is to observe homogeneity of task completion behaviour.

Furthermore, the study revealed evidence of association, notably the relationship between the scores on the final test and the final course grades is strong; and that between scores on the initial test and final course grades is satisfactory. This finding is interesting in that it is indicative of the suggestion that the test, if need be, can serve (be used) as an alternative test or, slightly remotely, as an equivalent test.

On the other hand, the very small or inverse relationship between the scores and the time taken to complete the tasks offers some evidence that high achievers tend to take less time to complete tasks. This last suggestion, however, needs to be explored further more thoroughly on a much larger sample than was used for this study.

Finally, it can be said that this study is a step forward in the direction advocated by McNiff (2002:146) that "Practitioners are required to account for their practice by producing reports to show that they can explain how their work has improved in terms of enhancing the quality of learning and experience for themselves and others."

Reference

Alderson, J. C., C. Clapham, and D Wall. 1995. . *1995 Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Berkeley Center for Teaching and Learning. 2014. "What to Do on the First Day of Class." http://teaching.berkeley.edu/what-do-first-day-class.

Downie, N. M., and R. W. Heath. 1974. *Basic Statistical Methods*. 4the Editi. New York: Harper & Row Publishers.

Eberly Center. 2008. "First Day of Class - Teaching Excellence and Educational Innovation - Carnegie Mellon University." http://www.cmu.edu/teaching.

Gupta, C.B., and V. Gupta. 1995. *An Introduction to Statistical Methods*. 19th Revis. New Delhi: Vikas Publishing House.

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

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge and New York:

Cambridge University Press.

Kline, P. 1986. *A Handbook of Test Construction*. New York: Methuen.

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

McNiff, J. 2002. *Action Research: Principles and Practice*. 2nd editio. London: Routledge Falmer.