A REPORT OF A PILOT STUDY OF A SERIES OF ENGLISH LANGUAGE TESTS

Teshome Demisse

1. Introduction

This report is part of a study project aimed at the construction and validation of a suitable English Language test for Freshman students at Addis Ababa University. The interest to pursue such a study arises from an awareness of the gap between the existing practice in Addis Ababa University and current trends in language testing. In this regard, Alderson and Clapharu (1992:149) write:

> Since language tests inevitably embody a view of language and indirectly a notion of language learning, it is important that test developer stake account of generally accepted views of language proficiency and language use when designing their tests. ... As theories of language knowledge become more refined, language tests which were formerly thought to be satisfactory start to loose their appeal, and are replaced by ones which reflect more closely the beliefs of the time.

The design, construction and validation of a new test requires the trailling and piloting of the test on a sample before the final administration to the target population. At the trialling stage, opinions of colleagues

and students have been taken into account to moderate the tests.

Hughes (1989:52) states the need for piloting more explicitly:

> Even after careful moderation, there are likely to be some problems with every test. It is obviously better if these problems can be identified before the test is administered to the group for which it is intended. The aim should be to administer it first to a group as similar as possible to the one for which it is really intended.

Broadly speaking, the results of the analysis of the initial administration on a sample (ie, the piloting of the tests) yields information about the behaviour of the tests (Baker, 1989:46). More specifically, the results gained from pretesting could provide useful information regarding the performance of the students (as individuals and as a group), and the performance of each of the items that make up the test (Heaton, 1975:174; Madsen, 1983:180).

Therefore, the particular aim of this report is to highlight the results of the analysis of the initial administration of a series of tests for Freshman students which were used for further moderation of the tests. In other words, the results of item analysis, and an investigation into the reliability and validity of the tests are reported.

2. Method of the study

Initially it was proposed that the pilot administration would be conducted on 120 students taking the Freshman English Course. Accordingly, students taking Freshman English 101A were invited to volunteer

to sit for the tests. However, only 39 candidates positively responded to this call. These students were required to take five tests, and immediately afterwards to fill in questionnaires about three of the tests. Seven English Language teachers were also involved in the examining and marking of them. These language teachers were also required to comment on the tests by filling in questionnaires. The questionnaires, both for the students and teachers, were designed in three parts to elicit information on the background of the respondents, and the face and contnet validities of the three new tests.

Considering the candidates' grades in ESLCE English and Freshman English 101 A, it should be noted that they were high achievers in English examinations. We should note that 76, 21 and 3 per cent of them had A's, B's and C's respectively in ESLCE English; and 28,36 and another 36 per cent of them had A's, B's and C's respectively in Freshman English 101A.

Furthermore, scores on two external (international) tests, ESLCE English and Freshman English 101A grades, and ESLCE and semester Grade Point Averages were also used for validation purposes, ie, constuct, concurrent and predictives validities.

Item analysis, and statistical investigation into the reliability and validity of the tests were carried out.

3. Description of the tests

The Written Test (WT1): This test has 14 sections. It is composed of reading comprehension, vocabulary, grammar, transitionals, reference skills, writing and cloze.

Generally, it integrates the receptive and productive categories of the reading and writing macro-skills.

The objective of this test is to assess the candidates' performance in the enabling skills necessary to read instructions, textbooks, handouts, reference sources, and to take notes from lectures or books, to answer short and/or long examinations in writing, as well as to assess their awareness of structural accuracy and understanding of closely related ideas.

The Listening Test (LT1): This test has five sections. It is composed of listening and labelling, listening for gist and details, transitionals, and partial dictation.

Generally, it integrates the receptive and productive categories of the listening and writing macro-skills.

The objective of this test is to assess the candidates' performance in the enabling skills necessary to listen to lectures, instructions or explanations, teachers' questions and discussions as well as their understanding of closely related ideas.

The Oral Test (OT1): This test has three sections. It is composed of a single dialogue (for reading aloud), a double dialogue, and a source with an information gap.

Generally, it integrates the receptive and productive categories of the listening, reading and speaking macro-skills.

The objective of this test is to assess the candidates' ability to ask and answer questions as well as the appropriacy, clarity and fluency of their speech.

These three tests are designed with the overall aim of assessing a candidate's ability to function effectively in English for academic purposes both in the receptive and productive skills.

External Test A(WT2): This test has two sections. It is composed of "structure and written expression, and vocabulary and reading comprehension". The candidates are required to read sentences and short passages and to answer the questions by choosing from the four options given.

While it appears to integrate grammar, reading comprehension and vocabulary, the writing is extremely controlled, especially when recording answers.

External Test B(LT2): This test has four parts. It is composed of a radio discussion, a radio news bulletin, a job interview, and a person talking about his job - all on recording. The candidates are required to listen to the recordings and answer the questions by putting a tick ( ) in two or four box choices. Only four questions of part three require the candidates to write a one - or two - word answer.

Generally, this test appears to be the least integrated because the other skills are quite controlled in its design.

Both of these external tests (WT2 & LT2) are used to assess the English Language mastery of students wishing to pursue their education in the U.S.A and Britain from all over the world.

One reason for their inclusion in the series is that they are approximately the same in purpose as the newly designed tests. That is, the purpose of the two external

(international) tests is to screen candidates wishing to enter educational institutions. Besides, some Ethiopian students who have completed high school and aspire for further education abroad are quite likely to take these tests. One can safely assume that English for academic purposes is assessed to some extent togehter with English for social survival, for example. But, unfortunately, there is no statistical information available concerning these tests.

4. Descriptive statistics and Test reliability

Table 1: Descriptive statistics and Reliability coefficients

Test	Total value	Average	Standard deviation	Coefficient of Discrimination	Test Reliability
WT1	187	130.85	22.53	0.121	0.915
LT1	43	36.07	5.08	0.118	0.793
OT1	5	3.62	0.75	0.150	0.530
WT2	100	65.25	12.58	0.126	0.865
LT2	30	12.48	6.24	0.208	0.797

Considering the average scores (in Table 1) of the tests, it can be seen that the candidates have found the three new tests and the first external test (WT2) rather easy, unlike the second external test of listening (LT2). Perhaps, this is not surprising given the fact that the sample population is rather homogeneous in language achievement levels as evidenced in their grades for the ESLCE English and Freshman English 101A.

It can also be seen, considering the standard deviations, that most of the tests do not spread out the candidates across the score range for each test very broadly. The coefficients of discrimination, which are expressions of the standard deviations as proportions of

the total marks for the different tests, reveal that the second external test of listening (LT2), followed by the oral test (OT1), spread out the candidates more effectively than the others.

The particularly minimal difference between the coefficients of the new written test (WT1) and the first external test (WT2), and to some extent that between the new listening test (LT1) and the second external test (LT2) is clear evidence of the fact that the group is fairly homogeneous with a narrow range of proficiency levels.

Perhaps one could reasonably assume that, at least, the two external tests would have spread the candidates rather more widely if it had not been for the homogeneity of the group. Note that the fact that the three new tests should behave more or less the same as the two external tests is quite encouraging.

Another criterion for judging a test is its reliability. For Weir (1988:34) "The concern here is with how far can we depend on the results that a test produces or" ... could the results be produced consistently.", and for Bachman (1990:160) it "... is concerned with answering the questions, 'How much of an individual's test performance is due to measurement error, or to factors other than the language ability we want to measure?'"

Generally, "Reliability is thus a measure of accuracy, consistency, dependability, or fairness of scores resulting from administration of a particular examination," (Henning, 1987:74).

While 0.9 or above is often mentioned as an appropriate coefficient of reliability for well made standardized tests, a coefficient of 0.7 (Baker, p.61; Kline, 1986:3, for example) is hinted at as the minimum value. Since "Internal consistency coefficients are very suitable for use in computing the reliability of academic tests," (Downie and Heath, 1974:239), the Kuder-Richardson Formula 21 (KR21) was used for this purpose.

Thus, a quick glance down the reliability column in Table 1 shows that the tests had quite satisfactory coefficients, except the oral test (OT1) which was mainly subjective in nature.

5. Item analysis of the tests

Item analysis is a useful procedure for revealing information about the performance of the test items comprising a test. It allows us to examine all the items in terms of their level of difficulty, level of discrimination (Heaton, p.173), and contribution to the total test (Hughes, p.160).

The scripts of the candidates who took the series of the tests in the pilot study were rank ordered according to their total score from the highest to the lowest. Applying suggestions for small samples (Harrison, Heaton, Madsen, Downie and Heath, for example) the top 1/3 and the bottom 1/3 of the group were used for the analysis.

Regarding facility values, any value falling between 0.4 and 0.6 is generally acceptable, 0.5 (50%) being the most desirable value (Harrison, pp. 128, & 131; Heaton, p.173). But other ranges are also suggested: for example, Kline (p.143) 0.2 to 0.8, Baker (p.54) 0.25 to 0.75, Heaton (p.173) 0.3 to 0.7, and Madsen (p.182) 0.3 to 0.9.

Discrimination indices, cited by Ddjenie (1990:72), ranging from 0.2 to 0.8 are considered acceptable, with 0.67 as the most desirable value. But, while a value of 0.3 or higher is satisfactory for Baker (p.54) and Harrison (p.131), it is 0.15 or higher for Madsen (p.183).

For item-test/-total correlation, the satisfactory levels are set at 0.3 or above by Hughes (p.160) and beyond 0.2 by Kline (p.143).

Generally, given the nature of the group, the more relaxed levels are kept at close range during selection.

Each item of the 200-item written test (WT1), with the exception of one task of writing a paragraph, was scrutinized in the light of these three criteria. Especially in the cloze section and at other places in the written test, some attempt to change the items (questions) was made as well as rejecting those items that failed to satisfy the requirements of the criteria. First, items that met the three criteria were retained without any change, followed by the acceptance of those items that fulfiled any two of the three criteria with some changes made to most of them.

Overall, the items included in the revised written paper had a range of facility values of 0.18 to 0.91 (with no more than three items at the extremes), and discrimination values of 0.18 to 0.75. Thus, the moderated written test (WT1) has 124 items and/or tasks for the value of 110 marks for further analysis at the final administration.

Similarly, the 43 items of the listening test (LT1) were examined. Accordingly, Section D, which was on transitionals, was wholly rejected. Quite a bit of change was made particularly to the partial dictation, and slight changes elsewhere in the paper.

Generally, the items included in the revised listening test had a rnage of facility values of 0.39 to 0.89, and discrimination values of 0.22 to 0.78 (and one item of 1.00). Thus, the moderated version of the listening test has 30 items for further analysis at the final administration.

The oral test (OT1) is not found amenable to item analysis, and the whole set was retained for the final administration with only slight changes made to the instructions.

According to the item analysis, it can be said that the two newly designed tests (ie, WT1 & LT1) are rather easy, and this may be due to the nature of the sample. The written test had an overall mean difficulty of 0.71 and the listening test had 0.82. The overall mean discrimination for the former is 0.21 and for the latter 0.28. Notice, here, that the listening test is easier than the written test, but it also discriminates better.

Given the small size of the sample and which, by coincidence, happens to be rather motivated with high language achievement levels, it was thought best to retain the easiness of the tests in many of the cases. This is done because the target population may not be as highly motivated as the pilot group.

A quick analysis of the two external tests was also made to see how they have functioned, though not for revision. Accordingly, the first external test (WT2) had facility values ranging from 0.06 to 1.00, and discrimination levels ranging from - 0.25 to 0.75. The second external test (LT2) had facility values ranging from 0.17 to 0.89, and discrimination levels ranging from- 0.11 to 1.00. In terms of these two criteria, 53 per cent of the items in the first external test and 78 per cent of the items in the second external test are found acceptable. The rest of the items in both tests were candidates either

for rejection or for some revision. Comparing the two, it can be observed that the second external test of listening behaved much better than the other.

6. Test validity

According to Henning (p.89), "validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is supposed to measure."

In this work, we proceed from the non-emprical to the empirical kinds of validity. First, the face and content validities, which require no use of formulae and have no coefficients or mathematical computations involved (Henning, p.94) are presented. This is followed by construct validity, which is empirical in nature though it does not have any one particular validity coefficient (Henning, p.98). It deserves this middle position to reflect the extent of empirical investigation it requires, and the comprehensive nature of the concept, ie, its overlap with content validity, for instance. In this connection, Weir (p.22) states that "The most helpful exegesis regards construct validity as the superordinate concept embracing all other forms of validity."

Finally, criterion-related validities, ie, concurrent and predictive validities are treated. These are empirical in that they involve the use of mathematical formulae for the computation of validity coefficients (Henning, p.94).

Face and Content validity: Candidates who sat the tests and language teachers who were involved in the examining and marking of the tests were invited to give their views on them in questionnaires. Information obtained from these were used to judge the face and content validities of the tests. In other words, they were asked if the tests actually met their expectations of language tests.

In reporting the responses to the questionnaires, the five-point scale is reduced to three categories: that is, disagree, neutral and agree, or bad, neutral, good in the students' case for content validity. When reporting in percentages, any value greater than 33 per cent is considered significantly meaningful-both for the students and teachers.

Table 2: *Students' response in frequencies and percentages*

SQ1WP	SD	%	D	%	N	%	A	%	SA	%	NR	%	Tot	%
FVIS	5	1.4	20	5.5	40	11.0	168	46.3	124	34.2	6	1.6	363	100
CVIS	12	1.7	26	3.8	108	15.6	327	47.2	185	26.7	35	5.0	693	100
OVAL	17	1.6	46	4.4	148	14.0	495	46.9	309	29.3	41	3.8	1056	100
SQ2L														
FVIS	0	0	9	3.3	39	14.4	133	49.3	84	31.1	5	1.9	270	100
CVIS	0	0	4	2.4	17	10.5	91	56.2	44	27.2	6	3.7	162	100
OVAL	0	0	13	3.0	56	13.0	224	51.9	128	29.6	11	2.5	432	100
SQ3S														
FVIS	1	0.4	8	3.5	29	12.6	102	44.4	88	38.3	2	0.8	230	100
CVIS	0	0	0	0	6	6.5	38	41.3	40	43.3	8	8.7	92	100
OVAL	1	0.3	8	2.5	35	10.9	140	43.5	128	39.8	10	3.0	322	100

Abbreviations: - SQ1WP, SQ2L, SQ3s = Student Questionnaire on the written, Listening, Speaking tests respectively.

- SD= Strongly Disagree; D= Disagree; N=Neutral;A=Agree SA = Strongly Agree; NR= No Response.

- FVIS = Face Validity Items; CVIS = Contnet Validity Items OVal = Overall, ie, combination of the two.

Overall, while about 76 per cent of the students have expressed their positive views regarding the appropriacy of the written test (WT1), about 29 per cent showed their strong agreement. More specifically, 81 per cent and 74 per cent thought that the test had good face and content

appearance, respectively. 34 per cent expressed their strong agreement about the face validity of the test whereas 27 per cent thought the content of the test was very good.

About 82 per cent expressed a positive view about the quality of the lestening test, and 30 per cent strongly agreed. More specifically, 80 per cent held positive views about the face validity whereas 83 per cent liked the contnet of the test. While 31 per cent strongly agreed with the face validity, 27 per cent thought the content was very good.

About 83 per cent of the candidates felt the oral test had good face and content validity, and 40 per cent expressed strong positive views. 83 per cent believed the test had good face validity, and 85 per cent thought the content was good. While 38 per cent strongly agreed to the former, 44 per cent believe the latter was very good.

Table 3: Teachers' response infrequencies and percentages

LTQ1WP	SD	%	D	%	N	%	A	%	SA	%	NR	%	Tot	%
FVIS	1	1.3	6	7.8	10	12.9	33	42.9	25	32.5	2	2.6	77	100
CVIS	0	0	0	0	10	13.0	30	39.0	37	48.0	0	0	77	100
oval	1	0.6	6	3.9	20	13.0	63	40.9	62	40.3	2	1.3	154	100
LTQ2L														
FVIS	0	0	0	0	8	11.43	36	51.43	26	37.14	0	0	70	100
CVIS	0	0	0	0	6	17.1	14	40.0	14	40.0	1	2.9	35	100
OVAL	0	0	0	0	14	13.0	50	48.0	40	38.0	1	1.0	105	100
LTQ3S														
FVIS	0	0	2	3.3	16	26.67	33	55	9	15.0	0	0	60	100
CVIS	0	0	0	0	4	22.22	9	50	5	27.78	0	0	18	100
OVAL	0	0	2	2.56	20	25.64	42	53.85	14	17.95	0	0	78	100

Abbreviations: The same as for Table 2, p.16.

Overall, 81 per cent of the language teachers agreed that the written test met their expectations, and 40 per cent expressed their positive views strongly. Specifically, 75 and 87 per cent agreed with the appearance and content of the test, respectively. And, 33 and 48 per cent held strong positive views about the face and content of the test, respectively.

86 per cent of language teachers agreed that the listening test had face and content validity; and 38 per cent showed strong agreement. Specifically, 89 and 80 per cent agreed with the claimed appearance and content of the test, respectively. 37 and 40 per cent respectively expressed strong positive opinions about the face and content of the test.

On the whole, 72 per cent of the respondents agreed with the claimed appropriateness of the oral test, and 18 per cent strongly agreed. While 70 per cent had positive views about the face of the test, 78 per cent agreed with the content of the test. 15 and 28 per cent of respondents expressed strong agreement regarding the face and content of the test.

Therefore, given the information in Tables 2 and 3, one can say that the tests have quite acceptable face and content validities as measures of language ability.

Construct validity: Hughes (p.26) defines it thus:

> A test, part of a test, or a testing technique is said to have construct validity if it can bedemonstrated that it measures just the ability which it is supposed to measure.The word 'construct' refers to any underlying ability (or trait) which is hypothesised in a theory of language ability.

Furthermore, he goes on to state how this research activity can be demonstrated by saying:

> If the coefficients between scores on the same construct are consistently higher than those between scores on different constructs, then we have evidence that we are indeed measuring separate and identifiable constructs (p.27).

In this work, very little attempt is made to test separate language abilities; rather, an attempt to integrate the different skills is sought. An example is the written test (WT1) which involves reading and writing; and the other two (LT1 & OT1) were designed to involve more than just listening and/or speaking. The three new tests and the two external tests are compared between each other to provide some idea of the construct validity of each of the tests.

Table 4: Correlation coefficients between tests

WT1	LT1	LT2	WT2	
–	–	–	–	WT1
0.60+	–	–	–	LT1
0.29	0.50+	–	–	LT2
0.72+	0.75+	0.37	–	WT2
0.72+	0.55+	0.45+	0.64+	OT1

+ = Significant at the 5% level

Comparing the new written test (WT1) with the others, we observe that there is a meaningful overlap and a significant relation with one of the external tests (ie, WT2), and least overlap with the other external test of listening (LT2). This is an evidence that the test is doing the job it is designed for.

There is also evidence that the new listening test is doing the right job given the reasonably meaningful overlap and significant relation with the external test of listening. That the new listening test should show higher correlation with the other tests is a reflection of the fact that other skills are highly controlled in the external test of listening.

The new oral test, too, bears more meaningful overlaps and significant relationships with the other tests than with the external test of listening (LT2). The meaningful overlaps and significant relationships between the oral test and the written tests could be due to the amount of reading involved in both whereas, though to a lesser extent, that between the oral test and the listening tests may be due to the amount of listening involved in both - listening to partners in the oral test, for example. Again, that the oral test should correlate the least with the external test of listening is evidence that the test is doing its job.

Concurrent validity: This is concerned with the validation of tests against some criterion measure of performance. "Another approach to test validity is to see how far results on the test agree with those provided by some independent and highly dependable assessment of the candidates ability," writes Hughes (p.23). And for Henning (p.96), "It is criterion-related in the sense that the validity coefficient derived represents the strength of relationship with some external criterion measure." Thus, when a strong relationship or a high level of agreement between tests and criterion measures is observed, we can consider this as indicative of the validity of the new tests.

In this study, the new tests are compared with two external tests, Freshman English grades and ESLCE English grades. The comparison is done pair wise and in combinations to examine the extent of agreement in what they yield.

Table 5: Correlation coefficients between the new tests and the criterion tests and grades

WT1	LT1	OT1	WLT1	WOT1	LOT1	WLOT1	
0.72+	0.75+	0.64+	0.78+	0.73+	0.78+	0.79+	WT2
0.29	0.50+	0.45+	0.36	0.22	0.53+	0.30	LT2
0.53+	0.71+	0.53+	0.64+	0.59+	0.69+	0.65+	FLEG
0.11	0.28	0.14	0.22	0.21	0.42	0.25	FSLEG
0.46+	0.68+	0.45+	0.60+	0.55+	0.73+	0.61+	ESLFLG

+ = Significant at the 5% level

In table 5, we notice the highest level of agreement between the new listening test (LT1) and the first external test (WT2), and between this latter one and the new written test (WT1), followed by that between the new listening test and Freshman English grades (FLEG). The least agreement is observed between the new written test and ESLCE English grades (ESLEG).

Generally, a combination of the new tests (in two's and the three in one) bears the highest level of agreement with the first external test (WT2), followed by Freshman English grades (FLEG).

Given the nature and the status of the criterion measures against which the new tests are compared, the hierarchical level of agreement observed is interesting. It is interesting because it seems to suggest that the candidates' level of maturity is matched. The external test

(WT2), which is a proficiency test, is for undergraduates and above. The Freshman English, which is an achievement test, is for undergraduates. And the ESLCE English grade is used both to certify high school completion and university entrance. That the new tests, as proficiency tests, should agree best with the external proficiency test is also encouraging.

Predictive validity: This "--- is usually reported in the form of a correlation coefficient with some measure of success in the field or subject of interest," says Henning (p.97). In this procedure, test scores are correlated with some future criterion of performance to find out to what extent the test(s) can predict candidates' future performance (Weir, p.28; Hughes,p.25).

In terms of the coefficients derived, Hughes (p.25) and Kline (p.5) point out that we can only expect a moderate one - something around 0.4 is generally considered satisfactory.

In this study, ESLCE Grade Point Averages, which candidates already have, and university Semester Grade Point Averages, which would be obtained at the end of the semester, are used as criterion measures of performance.

Table 6: Correlation coefficients between the new test and the criterion grade point averages.

WT1	LT1	OT1	WLT1	WOT1	LOT1	WLOT1	
0.42+	0.33	- 0.12	0.35	0.15	0.30	0.23	ESLGPA
0.46+	0.03	0.36	0.30	0.28	0.22	0.32	SGPA
0.55+	0.11	0.30	0.38	0.32	0.29	0.38	ESLSGPA

+ = Significant at the 5% level

In Table 6, we notice that the new written test (WT1) has respectable and significant correlation coefficients. Thus, this test satisfactorily predicts the candidates' overall academic performance as expressed in the form of university (SGPA) and ESLCE (ESLGPA) grade point averages. The new oral test (OT1) also predicts the university grade point averages (SGPA) with a close to 0.4 coefficient, but at a relaxed level of significance (ie, at 10%).

7. Summary

Given the evidence herein, the behaviour of the new tests is quite satisfactory; especially, the new written test behaved the best. The tests have acceptable reliability coefficients, and modest construct, concurrent and predictive validity. Above all, quite a reasonable proportion of respondents (teachers and students) agree that the tests are valid, especially in terms of their face and content, for assessing Freshman students' English language ability.

Finally, the tests have been moderated for the final administration at the end of which they will be subjected to further analysis.

REFERENCES

Alderson, C.and Carolyn Clapham 1992. "Applied Linguistics and Lnaguage Testing: A case study of the ELTS test, " Applied Linguistics, 13, 149-167.

Bachman, Lyle F 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Baker, David 1989. Language Testing. London: Edward Amold,

Dejenie Leta 1990. "Achievement, Washback, And Proficiency In School Leaving Examination: A case of Innovation In An Ethiopian Setting, "Ph.D. Thesis. Lancaster: University of Lancaster, Department of Linguistics and Modern English Language, (Unpublished)

Downie, N.M. and R.W 1974. Heath. Basic Statistical Methods, Fourth edition. New York: Harper & Row, Publishers,

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

Heaton, Brian 1975. Writing English Language Tests. London: Longman,

Henning, Grant 1987. A Guide to Language Testing. Cambridge: Newbury House Publishers,

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

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

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

Weir, Cyril 1988. Communicative Language Testing. Exeter: University of Exeter,