# A Small-Scale Evaluation of College English Examination (First Semester Final, 1996/97)

**Teshome Demisse**

## 1. Introduction

The Department of Foreign Languages and Literature has compiled a new textbook for the English course it offers to first year university students. Unlike the preceding textbooks for this course, the new textbook is based on a different understanding of "what it means to 'know' a language" (Baker, 1989:7). It is compiled in such a way that it puts ". . . emphasis on what is done with language" (Ibid.).

The textbook focuses on using English for academic purposes. The aim of the course is to help students improve their use of English -- their language skills and study practices. More specifically, while the development of language skills includes the students' abilities in listening, speaking, reading, writing, learning vocabulary, grammar, etc., the development of study practices includes reading, understanding and criticising real academic texts, taking lecture notes, writing academic essays, etc. (*College English*, 1996:3).

Given that the assessment of the outcome of the course is as important a concern as the teaching of the language, the Department has exerted some effort to bring about changes in the testing practice. Among some of the events this writer has actively participated in are the introduction of 'conscious assessment' as part of the evaluation mechanism for the course, and the several induction and awareness raising workshops that have been conducted: "Testing the Skills (November 1993), Freshman English Testing and Assessment (March 1996), and Training Markers (January 1997)."

Among other things, analysis of test results, involving graduate students in conducting research on the tests, and the establishment of a research group are on agenda, one which evolved from the March 1996 workshop held in the presence of an external testing expert.*

The aim herein is, therefore, to report an analysis, in terms of the level of difficulty and discrimination, of the examination designed to measure the progress and/or achievement of Freshman students who attended the course in the first semester of 1996/97.

## 2. Procedure

Examination papers of ten regular Freshman sections (a mixture of degree and diploma students) were collected. Each section was then alphabetized, and seven papers from eight sections and eight from two sections were pulled out systematically. This produced 72 [(7x8) + (8x2)] papers/scripts which were combined into one pile in rank order of the raw scores, i.e., from the highest to the lowest score.

* Professor Charles Alderson, visitor from the University of Lancaster.

Then, the upper one third (24 scripts) and the lower one third (24 scripts) of the pile were taken for analysis. Thus, a total of 48 students' responses, which is 1.5% of the enrollment, were used to analyse the examination. Note that according to the University Registrar's Office, 3111 students were admitted to the University in 1996/97 (AAU *News and Views*, 1996:7). The scores of the 48 students for the full course was also collected for analysis.

## 3. Description of Examination

Before the final examination, speaking was assessed twice; two listening tests, one writing test and a mid-semester examination were designed and administered centrally by the Freshman English Testing Committee. This took up 55% of the value for the course: speaking, listening and writing were 10% each and the mid-semester was 25%. The value of the final examination, therefore, was 45%. The sum of these would then determine the profile of the students in English.

The final examination contained two unmutilated and three mutilated passages based on which there were tasks of guided summary, grammar, reading comprehension, and vocabulary. It was organized in four parts, three of which had a total of eight sections. The examination was thirteen pages long (with 90 items) to be attempted in two and a half hours.

## 4. Analysis and Discussion

### 4.1. Descriptive Statistics

**Table 1: Information on Freshman English Results out of 100 marks for the course (N = 48)**

| Mode | Median | Average | Standard Deviation | Range |
|---|---|---|---|---|
| 67 | 34 | 48.4 | 22.94 | 69 |

The information in this table is indicative of the difference between and among the groups used for this analysis, i.e., that the sample was mixture of students in the degree and diploma programmes. The high variation, as observed in the standard deviation and range, is evidence of this, i.e., that the group is heterogeneous in its academic level. When the raw scores (out of 100%) are ranked, there is a wide gap between the lowest of the upper 24 scores and the highest of the lower 24 scores. The rather low average of 48.4 falls within this gap. This average is also within the actual range of the cut-off points (46.69) for the grade of 'C' as determined for the target population. The most frequent score (Mode) is found towards the lower end of the upper group whereas the middle score (Median) is found at the top of the lower group.

## 4.2. Item Analysis of the Final Examination by Parts And Sections

Item analysis is a systematic procedure carried out to see how good or appropriate a test or an examination is. It produces information on how each question, item or task functions in the whole examination. More specifically, it tells us how difficult each item is, and whether the item discriminates between high- and low-achieving students. Item analysis "is used with any important exam - for example, . . . tests given at the end of . . . a term or course" (Madsen, 1983:180) as a useful source of feedback for all concerned in the design and use of the test. In this regard, Madsen comments that "while many teachers are too busy to evaluate each item in every test that they give, at least major class tests should be carefully evaluated" (1983:179).

Considering level of difficulty, any value falling between .4 and .6 is generally acceptable, .5 being the most desirable value (Harrison, 1983:128 & 131; Heaton, 1975:173). Discrimination indices of .4 or above are satisfactory for Alderson *et al.* (1995:82), and Dejene (1990:72) cites .67 as the most desirable value.

Another value in carrying out item analysis is to gain an indication of the reliability and validity of the examination. In the words of Davies (1990:5-6):

> Item analysis is part reliability, part validity, the assumption being that items with modestly high discrimination are likely also to be replicable, and that that very discrimination is itself an augur of satisfactory test construct, that is, of the items belonging together.

He goes on to explain further when he writes (Ibid., p. 6):

> The purpose of item analysis is to determine test homogeneity: the more similar to one another (without being identical) test items are, the more likely it is that they are measuring in the same area and therfore that they are doing something useful (validity) and doing it consistently (reliability).

Alderson et al. (1995:80) also suggest that the discrimination index gives an indication of the validity of the test in the sense of the domain it claims to measure in the definition they offer: " . . . the discrimination index measures the extent to which the results of an individual item correlate with results from the whole test."

Some of the items (see appendix) are clearly candidates for revision or rejection/replacement if the item analysis was carried out on a sample of the target population before the actual administration of the examination. The inspection begins with items with low (high) values for difficulty and low indices for discrimination. For example, thefollowing items need to be scrutinized during moderation of the examination. Items 1 and 2 in part one, item 9 in section B of part two, and items 1 and 5.5 in section A, item 12 in section C, items 17 and 18 in section D of part three are difficult and poorly discriminating. Items 7, 8 and 9 in section A, items 2 and 3 in section B of part two, and item 5.6 in section A of part three are easy and poorly discriminating.

On the other hand, items 6 and 13 in part one are effective. Item 6 has the best level of difficulty and perfect discrimination, i.e., all of the upper group, but none of the lower group, responded correctly.

Item 13, too, approaches the best level of difficulty and discriminates very well.

The table below is intended to show the average level of difficulty and discrimination of each section.

**Table 2: Level of Difficulty and Discrimination in Averages: Reliability (KR-20) = 0.97**

| | Facility Value (F.v) | Discrimination (D) |
|---|---|---|
| Part One: Guided Summary | 0.45 | 0.72 |
| Part Two: Grammar | 0.73 | 0.40 |
| Section A | 0.79 | 0.35 |
| Section B | 0.66 | 0.44 |
| Part Three: Reading Comprehension | 0.53 | 0.41 |
| Section A | 0.60 | 0.41 |
| Section B | 0.49 | 0.46 |
| Section C | 0.39 | 0.31 |
| Section D | 0.47 | 0.42 |
| Part Four: Vocabulary | 0.51 | 0.52 |
| Section A | 0.66 | 0.55 |
| Section B | 0.41 | 0.50 |
| Full Examination | 0.55 | 0.49 |

From Table 2 we can see that the students found section C of part three the most difficult (.39), and the level of discrimination is the lowest (.31). In this section, as part of reading comprehension, the

students were required to decide whether the given statements were true/false, supported by evidence from the text (passage). What probably makes this section difficult is providing evidence as well as the fact that the allotted mark is awarded only if both conditions are satisfied.

Section A of part two was the easiest (.79) for the students with a moderate level of discrimination (.35). In this section, as part of grammar, the students were provided with a two-paragraph context with alternative words or phrases in brackets in the text. They were required to determine the tense of the verbs according to the context. The fact that the method (format) and the content of the test are familiar to the students probably accounts for the easiness of this section.

The average level of difficulty for the vocabulary part (.51), followed by the reading comprehension part (.53), comes closest to the most desirable value of .5 (50%). The grammar part (.73) was easy with moderate discrimination (.4) whereas the guided summary tended to be difficult (.45) but with a modestly high discrimination (.72).

The overall average difficulty (.55) for the full examination is quite encouraging, and the level of discrimination (.49) is also adequate for an achievement test.

However, the moderate values of discrimination for the different sections and parts, except the guided summary, and the overall average discrimination (.49) for the full examination do not clearly suggest that the items are strongly pulling together in terms of measuring in the same area (construct); and this casts a shadow of

doubt on the validity of the test. Still, the fact that there are no negative values in discrimination and the high reliability coefficient (r=.97) are encouraging.

Note that the reliability of a test depends on the type and length of the test. For instance, an objective test of 100 items might have a reliability index of .95 (Alderson *et al.*, 1995:88). Furthermore, the inclusion of several passages with differrent content increases the reliability of a test because the bias due to passage content could be minimized (Bachman, 1990:220). The examination, item analysed herein, is objective and has 90 items in the contexts of five passages, and the high reliability observed can be attributed to these features of the examination.

## 5. Conclusion

The 1996/97 College English (first semester) final examination was at about the right level of difficulty (.55) with a moderate discrimination index (.49). The examination is a reliable (r=.97) measure of the English language achievement of College English students.

The level of difficulty and discrimination of the individual items (see appendix), the separate sections and parts are reasonably satisfactory - although, undeniably, there are few items that are found to be outliers.

Such analysis and evaluation of our tests and examinations, when exercised over many semesters and years, are bound to yield reliable information on the quality of testing and assessment in College English.

## Acknowledgements

This item analysis was part of the plan of activities for the graduate course "Language Teaching Methodology II, TEFL 506.

I would like to acknowledge the contributions of the 1996/97 first year graduate students for the scripts and scores they brought to our class, the computations of indices of difficulty and discrimination of the individual items, alphabetizing, sampling and ranking of the scripts, and for calculating the average and standard deviation of the scores.

## References

Alderson, J. C. *et al.* 1995. *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Atkins, J. *et al.* 1996. *College English* (1 & 2). Addis Ababa: AAU Printing Press.

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

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

Davies, A. 1990. *Principles of Language Testing.* Oxford: Basil Blackwell.

Dejene Leta. 1990. "Achievement, Washback, and Proficiency in School Leaving Examination: A Case of Innovation in an Ethiopian Setting." PhD Thesis. Lancaster: University of Lancaster, Dept. of Linguistics and Modern English Language (Unpublished).

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

Heaton, B. 1975. *Writing English Language Tests.* London: Longman.

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

*News and Views.* Nov/Dec. 1996 (No.7). Addis Ababa University.

## Appendix
## Facility Values and Discrimination Indices of Individual Items

| Exam Part/Section | Facility Value | Discrimination Index |
|---|---|---|
| **Part I: Guided Summary** | | |
| 1 | 0.23 | 0.21 |
| 2 | 0.15 | 0.21 |
| 3 | 0.48 | 0.88 |
| 4 | 0.56 | 0.88 |
| 5 | 0.38 | 0.58 |
| 6 | 0.50 | 1.00 |
| 7 | 0.50 | 0.75 |
| 8 | 0.39 | 0.79 |
| 9 | 0.44 | 0.88 |
| 10 | 0.52 | 0.71 |
| 11 | 0.48 | 0.79 |
| 12 | 0.54 | 0.67 |
| 13 | 0.46 | 0.92 |
| 14 | 0.46 | 0.83 |
| 15 | 0.48 | 0.88 |
| 16 | 0.48 | 0.71 |
| 17 | 0.60 | 0.79 |
| 18 | 0.54 | 0.67 |
| 19 | 0.58 | 0.83 |
| 20 | 0.29 | 0.50 |
| **Part II: Grammar** | | |
| **Section A:** | | |
| 1 | 0.60 | 0.54 |
| 2 | 0.77 | 0.38 |
| 3 | 0.81 | 0.38 |
| 4 | 0.81 | 0.38 |
| 5 | 0.81 | 0.29 |
| 6 | 0.73 | 0.54 |
| 7 | 0.85 | 0.21 |

| Exam Part/Section | Facility Value | Discrimination Index |
|---|---|---|
| **Grammar: Sec. A cont'd.** | | |
| 8 | 0.88 | 0.17 |
| 9 | 0.83 | 0.25 |
| 10 | 0.81 | 0.38 |
| **Section B:** | | |
| 1 | 0.81 | 0.38 |
| 2 | 0.85 | 0.29 |
| 3 | 0.85 | 0.29 |
| 4 | 0.56 | 0.63 |
| 5 | 0.46 | 0.25 |
| 6 | 0.65 | 0.63 |
| 7 | 0.54 | 0.75 |
| 8 | 0.73 | 0.46 |
| 9 | 0.44 | 0.21 |
| 10 | 0.75 | 0.50 |
| **Part III: Reading Comprehension** | | |
| **Section A:** | | |
| 1 | 0.44 | 0.12 |
| 2 | 0.39 | 0.29 |
| 3.1 | 0.33 | 0.33 |
| 3.2 | 0.75 | 0.25 |
| 3.3 | 0.50 | 0.42 |
| 4.1 | 0.48 | 0.29 |
| 4.2 | 0.44 | 0.46 |
| 5.1 | 0.85 | 0.29 |
| 5.3 | 0.75 | 0.50 |
| 5.4 | 0.69 | 0.54 |
| 5.5 | 0.35 | 0.21 |
| 5.6 | 0.88 | 0.25 |
| 6 | 0.52 | 0.63 |
| 7.1 | 0.65 | 0.71 |
| 7.2 | 0.65 | 0.71 |
| 7.3 | 0.65 | 0.71 |
| **Section B:** | | |
| 8 | 0.33 | 0.42 |
| 9 | 0.75 | 0.33 |
| 10 | 0.39 | 0.63 |

| Exam Part/Section | Facility Value | Discrimination Index |
|---|---|---|
| **Reading Compr. cont'dSection C:** | | |
| 11 | 0.50 | 0.42 |
| 12 | 0.27 | 0.21 |
| 13 | 0.39 | 0.29 |
| 14 | 0.38 | 0.33 |
| **Section D:** | | |
| 15 | 0.71 | 0.42 |
| 16 | 0.58 | 0.67 |
| 17 | 0.27 | 0.13 |
| 18 | 0.25 | 0.25 |
| 19 | 0.33 | 0.33 |
| 20 | 0.65 | 0.71 |
| **Part IV: Vocabulary** | | |
| **Section A:** | | |
| 1 | 0.65 | 0.46 |
| 2 | 0.77 | 0.46 |
| 3 | 0.50 | 0.83 |
| 4 | 0.69 | 0.54 |
| 5 | 0.69 | 0.54 |
| 6 | 0.69 | 0.54 |
| 7 | 0.56 | 0.54 |
| 8 | 0.73 | 0.46 |
| Section B: | | |
| 1 | 0.33 | 0.58 |
| 2 | 0.38 | 0.58 |
| 3 | 0.44 | 0.54 |
| 4 | 0.27 | 0.38 |
| 5 | 0.65 | 0.54 |
| 6 | 0.29 | 0.33 |
| 7 | 0.54 | 0.50 |
| 8 | 0.46 | 0.50 |
| 9 | 0.35 | 0.46 |
| 10 | 0.38 | 0.33 |
| 11 | 0.52 | 0.79 |
| 12 | 0.33 | 0.42 |