STATISTICAL ANALYSIS OF THE 1968 ESLC ENGLISH LANGUAGE EXAMINATION

Harold S. Madsen

At present in Ethiopia, an elaborate statistical analysis is being made of the 1968 and 1969 school-leaving English language examinations. Such a project may seem curious, particularly to those in the humanities. In fact, to a great many English teachers throughout the world, the use of statistical or mathematical analyses of English seems absurd. These persons often regard such investigations as attempts to apply the methods of science to the humanities in order to borrow some of the prestige of the former. The more charitable consider statistical analysis of English as well-intentioned but of little practical value. They feel more comfortable making subjective evaluations of their "subjective" subject.

But in recent decades, particularly as linguistics, TESL, and TOEFL have come into their own, statistical investigations of English language teaching have gained respectability. Despite demonstrated soundness, however, widespread acceptance of these techniques has been slow. John C. Gerber, a nationally prominent English specialist, has complained that "we in English have been far too slow in employing appropriate scientific techniques in studying our problems."

The value of statistical analysis. — Two examples demonstrate the usefulness of the statistical approach.

Julias Boraas, aware of the perrenial claims that traditional grammar instruction is the most satisfactory means to develop proficiency in English (such as composition or speaking), conducted a study in which he determined the relationship or statistical correlation between traditional grammar and various subjects in the curriculum. He discovered, ironically, that results of grammar instruction "correlated **highest with ability in arithmetic, and least of all (.28) with ability in composition."**

Earlier this year, a seasoned teacher with a fair amount of experience in objective testing identified with the help of associates what he felt were the weak questions among the 240 structure items on the English language ESLC paper. Then he selected the eleven **weakest**—questions which presumably were very ambiguous, especially for bright students. These were questions 1, 32, 36, 38, 65, 66, 77, 80, 83, 109, 120. Again a statistical analysis proved the subjective impression to be incorrect. A sample of 600 ESLC papers was taken. An item analysis of this sample demonstrated that six of the eleven questions listed as very weak (numbers 32, 55, 66, 77, 80, 83) were not even remotely ambiguous—even for the brightest students; four might possibly be ambiguous; and one (number 109) was unmistakably ambiguous. For instance, question number 80 reads:

```
            A                    B                      C
Some of the / students seems to think / that memorizing notes /
                      D                      E
   is all they have to do in school.      (Correct)
```

The "correct" response was listed as "B", but some teachers felt "A" might be a choice brighter students would make, by mentally changing part "A" to read "A section of the." In actual fact, **none** of the 200 brighter stulents selected "A"; 187 selected "B" (the correct response); 2 selected "C"; 5 selected "D"; 5 selected "E"; and 1 failed to answer the question.

Earlier ESLC studies.—Very few statistical studies have ever been made of the ESLC and fewer still of the English language paper. In 1964 David Korten and Irene Bayorke did some research on the ESLC as a predictor of academic success in the College of Business at Haile Sellassie I University. The next year Lane Tracy from the same faculty conducted a broader study. He discovered that the correlation been ESLC passes and University grade point average was rather low (.36). Turning to subject areas, Tracy felt that "a good subject examination probably should have a correlation coefficient of at least .50 [I would suggest the .60's or .70's] when correlated with grades in the same subject in college." But when comparing students' ESLC English grades with first-year English grades among various faculties at the University, Tracy found that correlations ranged from a low of .20 to a high of .46. Yet another study, begun in 1967 by John Rogers, Grover Hudson, and other English teachers at the University, is still in progress. Drawing on a sample of 230 essays from the 1967 ESLC, these researchers are attempting to classify every error, as well as to indicate the frequency of these errors.

The 1969 ESLC Study

Sources of data.—In April 1968, the author developed the research plan for evaluating this year's ESLC English language paper. Supported initially by ESLC funds, the project included a language questionnaire which was administered to all 6,500 candidates who sat for the English examination. After the initial phase was completed, the research plan was expanded.

The following data is to be collected: Out of the 6,500 candidates who sat for the examination, a sample of 600 "regular" candidate papers are to be selected at random. In addition, test papers from the schools (averaging no more than 100 per school) are to be extracted, together with the twelfth grade, second-semester English performance of these students as well as their first- and second-semester English grades during their first year at the University. Finally, a tabulation is needed for all 6,500 questionnaires; a random selection of 100 candidates whose native tongue is Amharic and an identical number whose native tongue is not Amharic—both groups of whom have similar educational background, as determined by the questionnaire; and a random selection of 100 native Amharic speakers who were raised in a large city and 100 who were raised in a village.

Areas of analysis.—Nine areas of analysis will be included in the overall study.

First an **item analysis** of the 600-paper sample is to be completed, including each of the structure items and each of the comprehension questions. This involves grouping the top 200, middle 200, and bottom 200 papers; then tabulating the number of responses for each of the five choices, and computing the discrimination level. The purpose of this analysis is to determine which questions are very easy, which are very difficult, which have negative discrimination, which demonstrate ambiguity, and which demonstrate satisfactory discrimination among poor, average, and superior students. The primary aim will be to assist English examiners in the preparation of an improved test for the following year.

Second is the **subject matter area analysis.** Using the 600-paper sample, researchers will correlate each subject area with the total section score minus the area score. For example, the verb score will be correlated with the strucutre score from which the verb score has been deducted. Next, using the ten-school sample, a correlation of the subject area with the secondary school English grade will be computed. Then another series of correlations will be worked out using section scores (structure, comprehension, controlled essay) minus each of the area results (particles, vocabulary, etc.). This study will help determine which areas of the examination are the most significant indicators of English proficiency, which areas — if any — have a negative effect in such measurement, and which are redundant.

Third, a **question-type analysis** will be carried out using the 600-paper sample. Similar to the split-half reliability study, this consists of computing the statistical relationship between "slashed-sentence" structure questions and multiple-choice structure questions, an equal number of each to be compiled from the subject areas. Then using the ten-school sample, a correlation will be computed between the slashed sentences and the twelfth-grade results, followed by correlation between the multiple-choice sentences and the twelfth-grade results. Statistically significant differences will be determined. The obvious purpose of this investigation is to determine whether the form of the question has a significant bearing on the measurement of specific skills — that is, whether the slashed sentence or multiple-choice question does the better job in measuring language proficiency, whether one is more difficult than the other, and whether each is superior in certain areas.

The 600-paper sample will also be used in the **reliability analysis.** A split-half reliability computation will utilize right-wrong responses of paired questions from each subject area and each question type. Structure, comprehension, and combined structure-comprehension computations will be made. The purpose of this analysis is to determine how consistent the examination is in measuring language skills. A comparison will be made with the previous more subjective ESLC test. Lindquist has said that reliability is "an important characteristic of any test, a charactedistic which is essential to but not a guarantee of validity."

The **validity analysis** will make use of the ten-school sample by providing a series of correlations with each school (total ESLC score, structure score, comprehension score, and controlled essay score) followed by a composite correlation combining the scores from all ten schools. In addition, a correlation will be worked out between University grades and ESLC grades. Then too, 1967 ESLC English language papers will be subjected to a similar evaluation. Afterwards, 1967-1968 results will be studied to determine whether or not the differences are statistically significant. This portion of the research is designed to determine whether or not the examination actually measures what it is intended to measure. As many evaluations as possible will be used (comparisons with standardized verbal tests, correlation with actual writing etc.). The subject matter area analysis can assist in determining what areas need to be strengthened, modified, or eliminated in order to make the examination more valid.

The guessing penalty is to be analysed by employing the ten-school sample. A correlation will be worked out relating ten-school secondary-level performance with corrected and uncorrected ESLC results — (structure, comprehension, and combined scores). This should determine whether the test is strengthened or weakened by the "correction" penalty for guessing. Currently, the penalty consists of R-W/4 (right minus the number wrong divided by four).

The **comparative linguistic analysis** will be based on the questionnaire and ESLC performance. One hundred ESLC papers of native Amharic speakers will be compared with those of a predominant non-Amharic tribal group, both in terms of total performance and proficiency in subject areas of the English language paper, The level of significant difference will be computed. If funds permit, the total ESLC performance of at least five major tribal groups will be tabulated and significant differences established. Finally, comparisons will be made between 100 city and 100 village candidates; total ESLC English performance as well as subject area performance will be compared, and the statistical level of significant difference will be determined. For one thing, it will be learned whether or not native Amharic speakers perform more satisfactorily than non-native Amharic tribal groups; whether there are certain areas of difficulty peculiar to various tribal groups; and whether those raised in big cities have a significant advantage over those raised in villages.

The **readability level of the ESLC English test** will be compared with first-year vocabulary proficiency at Haile Sellassie I University. The Dale-Chall readability formula will be employed as well as Langmuir-Bowers-Lee studies on the vocabulary proficiency of Ethiopian University students. This investigation should determine whether the examination is suitable for the candidates as far as readability and vocabulary are concerned.

The ninth and final analysis consists of a non-statistical **grammatical** structure comparison between those structures taught in grades 7 to 12 and those tested in the ESLC English examination. This will determine whether or not the examination is covering adequately the various areas of English that the student was exposed to during his secondary schooling.

Statistical findings. — At the time of writing, only a fraction of the projected statistical analyses have been completed. Nevertheless, some significant findings can be reported.

A very detailed item analysis of the entire structure paper has been worked out by Mr. Joe Wendel, with the assistance of the University Testing Center, the Central Statistics Office, and student assistants. The vast majority of structure questions were found to discriminate in the right direction. That is, better candidates selected the "right" answer more frequently than did average candidates; average candidates selected it more frequently that did weak candidates. Ambiguous questions and questions of negative discrimination were rare (the latter is a question the correct answer to which is supplied more frequently by weak candidates than by strong candidates). Test expert Charles Langmuir, Head of the University Testing Center, concluded that unsatisfactory questions were so few that considering the length of the examination ESLC English language candidates were definitely not penalized by faulty questions.

Greatest boon of the item analysis is the vast amount of valuable information provided English examiners. Knowing precisely which kinds of questions are easy, difficult, and mildly challenging, they can give even better balance to the test. Analysing results question by question, they can alter choices and eliminate unsatisfactory question forms until a very strong instrument has been devised.

The item analysis also served a diagnostic function: it disclosed which structures students had mastered rather well and which they still have difficulty with. For instance, students had relatively little difficulty with tense items, tag questions, and connectives (coordinating, correlative, and adverbial). On the other hand even the brighter students had difficulty with comparatives, articles, mechanics, and structures such as: so . . . that, few/ a few, little/ a little. Of moderate difficulty were questions testing relative pronouns, negatives, uncountable nouns, repeated direct objects, and fragments. Some areas contained questions with a wide range of difficulty; these included agreement (students did rather well here), verb particles, proper word order, and case. Questions on case ranged from a low of 15 per cent to a high of 88 per cent correct among the top third of the students.

It was also found that slashed sentences are more difficult than multiple-choice questions, probably because the former do not focus on the potential problem area. No evidence has so far been extracted to justify the elimination of either form of question. The completion-type question was least satisfactory, but this seems to be the result of faulty rubric rather than inherent weaknesses in the question type.

To estimate the consistency (or reliability) of the structure examination (Part II) Dr. Michael King of the Testing Center used Scott's Homogeneity Ratio (HR). Reporting on his findings, Dr. King says:

> For a test, such as the English ESLCE, that already has a great many items, a measure of internal consistency can serve as a good estimate of test reliability. For if there exists even a moderate degree of internal consistency for such a lengthy test, one can be certain that a split-half reliability coefficient would be quite high.
>
> The HR obtained for the structure section was 154 . . . On the basis of experience with tests of similar internal consistency and length I would predict a quite high split-half reliability coefficient (between .85 and .95).

Both King and Langmuir of the Testing Center consider the 1968 English language ESLC to be very strong as far as reliability is concerned.

While the validity study of the 1968 English ESLCE is just in its initial stage, a tentative evaluation can be proffered. Dr. King recently summarized the validity studies in which he and his wife, Dr. Johanna King, have been involved:

> The manner in which the objective English ESLCE was constructed, systematically analyzing errors made on a previous essay type English ESLCE, provides a strong argument that the test is in fact reflective of students' English language skills. Our effort has been directed at demonstrating that scores resulting from the objective test correspond with English language skill information coming from other sources. Specifically, we have attempted to show that students judged to have good or poor English language skills on the basis of the objective exam were judged similarly by secondary school English teachers and by scores on another general test of English language verbal ability. We compared the 1968 English ESLCE letter grades with 2nd semester final English letter grades of 172 Laboratory School students. A product-moment correlation between these two sets of letter grades was .69. When raw scores rather than grades were considered, this correlation was higher due to less restriction in range of the numbers entering

into the computation (.75). Such a high correlation strongly demonstrates that the objective test made the same kinds of discriminations among students as did secondary school teachers' judgments. Students scoring high on the objective exam had higher English grades in secondary schools. One hundred and seventy eight of these 179 Laboratory School students had also taken a test of verbal reasoning developed by the University Testing Center. This test, FORM 68, has a section that can be said to reflect general ability to use the English language. Students' objective ESLCE letter grades correlated .64 with their scores on the verbal section of FORM 68.

There is, therefore, a strong relation between the objective English ESLCE scores and information on English language ability from two other sources.

This "first look" at the objective English ESLCE is very tentative but strongly supports the assertion that the exam is a reliable instrument for assessing English language skills.

Since the above statement was prepared, a correlation between the twelfth grade performance of Menelik Secondary School students and their performance on the English ESLCE was computed at .72 — a strong correlation and very close to that at the Laboratory School. However, since objective tests comprised part of the evaluation at these schools, additional means are being sought to esablish still more convincingly the validity of the 1968 English ESLC.

Similar correlations to those mentioned above have also been computed for individual sections of the test. Whereas the entire test provides a correlation with Lab School grades at .75, the structure section by itself is only .66, and the comprehension .55. The controlled essay did not discriminate particularly well. It is encouraging to note that the various parts of the examination complement each other. Each contributes additional information regarding the student's language proficiency.

Since the language questionnaire does not have top research priority, there is not a great deal to report at present. One interesting fact deserves attention, however. Even in the provincial secondary schools outside Addis Ababa, it appears that very close to 50 per cent of all candidates are native Amharic speakers. If Dean Abraham Demos' estimate is correct, that only five to seven million out the twenty million inhabitants of Ethiopia are native speakers of Amharic, then it would appear that the present school system offers native Amharic speakers a greater chance of success than it does those who do not speak the national language as a mother tongue. However, projected research may well demonstrate that rural academic disadvantages are even more significant than tribal.

Non-statistical findings. — Despite the value of statistical data, a report on the ESLC English language test would not be complete without a summary of non-statistical findings.

The briefing of twelfth-grade English teachers throughout the Empire regarding the new exam seemed to be quite effective. First, they received word of the January 1967 ESLC-Curriculum which outlined the proposed changes more than a year prior to the April 1968 examination. Additional information was provided by means of the **Journal of the Teachers of English in Ethiopia.** Then during the 1967-1968 school year, the Ministry of Education disseminated to all secondary schools two sets of detailed information regarding the examination as well as sample questions provided by the English examiners.

Setting the examination was extremely time-consuming, requiring many weeks of intensive labor plus days of intensive scrutiny of each question by four English and American specialists in order to avoid troublesome expressions peculiar to one nationality. It is expected that subsequent tests will be considerably less time-consuming.

Production of the test greatly taxed existing facilities. Printing, collating and stapling the 36-page exam required regular and volunteer workers to work almost round the clock for many days in order to meet the deadline. The bulk of the examination was far greater than any other of the ESLC papers; transporting and storing the exam were both troublesome.

Administration of the test went smoothly. Directions appeared to be clear; students seemed to be well prepared for the kind of questions set; there was very little if any evidence of confusion. Feedback from proctors indicated that the time allotment was adequate. The major difficulty was mechanical: missing pages, duplicate pages, and unstapled sheets; complaints regarding such difficulties were sufficiently numerous to indicate that a solution of this difficulty for next year was imperative. There was very little evidence of attempted cheating.

Test security was erratic. English examiners, for three reasons, desired that test papers not be left in any center: first, to enable the present test to be perfected (as in standarized examinations abroad) by merely modifying existing questions; second, to save countless hours of time in preparing a new examination; and third, to help prevent poor teaching on the part of weaker teachers who might simply use the test as the ' text" for this year's twelfth-grade students. But while in some centers all test papers were returned, in other centers proctors were very lax, permitting the loss of tests.

Reaction to the new examination was generally quite favorable. Ato Tesfa, ESLC Administrator, reported he was astonished to receive no negative reaction considering that the test represented such a departure from the usual format. Some teachers were irate at not receiving copies of the test. Student comments ranged from, "It was too long" or "It was easy" to "The English paper was fair to the student." As mentioned earlier, the time allotted was adequate for the vast majority of students. Scores followed a perfect bell-shaped curve, with equal numbers of highs and lows. Letter grades followed basically the same pattern as those in previous years.

The reading and grading of the test took longer than expected, considering that one of the reputed advantages of the objective test is its efficiency. First, having to turn through 36 pages of a booklet to score the test was time-consuming. Second, the guessing penalty slowed down the grading. Third, reading the controlled paragraph took more time than expected. Fourth, the greatly increased number of candidates (nearly 6,500 total) compensated for the time saved in grading the objective section.

In short, the examination was well received but there was a demonstrated need for improvement in production, security, and grading.

Implications for the 1969 examination. — Initial findings indicate that the 1968 examination was both reliable and valid, thus justifying the continuation of the objective test together with a controlled essay. Since the structure section and the comprehension paper were found to complement each other, they have both been retained. There was no evidence to suggest that either the slashed sentence or the multiple-choice question should be discontinued. However, there were findings which strongly suggested a modification in the weight assigned to various subject matter areas. Because of the backwash effect and diagnostic value, it was decided that no areas tested in 1968 would be eliminated in 1969; but with the reduction of emphasis in some areas, additional structures were tested in 1969. Weak security on the last test dictated that new questions had to be formulated throughout. And the item analysis pointed out the need for several hundred minor modifications of questions which were in turn incorporated in the 1969 English language ESLC test. It was not deemed feasible to introduce an oral section in this year's ESLC.

Burgeoning ranks of ESLC candidates suggested the need for a separate answer sheet; therefore this was introduced in the 1969 test. For one thing, accurate scoring was assured by the new electronic "reading" of test papers. In addition correction time was reduced considerably; research time and cost were likewise reduced, and storage space was trimmed enormously. The 1969 test was printed (and numbered) abroad, to assure a perfect copy for each candidate plus maximum security. Then too, the bulk of the papers was greatly reduced while maintaining approximately the same length of test.

Finally, a 400 paper sample of the 1969 test was graded in advance; from this an item analysis was prepared. Next the weak questions were eliminated from the answer key, and at length all 8,400 papers were scored from this altered key which ignored the questions with demonstrated defects.

In brief, this massive ESLC English language research project is designed to insure continued improvement of the school-leaving examination and hopefully of secondary school English instruction itself.