# A Comparison of the Abilities Measured by the Cambridge and Educational Testing Service EFL Test Batteries[1]

**Lyle F. Bachman**
*University of California, Los Angeles*

**Fred Davidson**
*University of Illinois, Urbana-Champaign*

**John Foulkes**
*University of Cambridge Local Examinations Syndicate*

*This paper compares the abilities measured by the First Certificate of English (FCE) administered by the Cambridge Local Examinations Syndicate and the Test of English as a Foreign Language (TOEFL) administered by Educational Testing Service (ETS). An investigation of the factor structure of the two test batteries, both within each test battery as well as across test batteries, is presented. The analyses suggest that the FCE and ETS test batteries administered in this study appear to measure, to a large degree, the same common aspect of language proficiency of the subjects in the sample. The contribution of this study to language testing is also discussed.*

## INTRODUCTION

The First Certificate in English (FCE), administered by the University of Cambridge Local Examinations Syndicate (Cambridge) and the Test of English as a Foreign Language (TOEFL), administered by Educational Testing Service (ETS), are widely used as measures of proficiency for English as a foreign language (EFL) throughout the world. Hundreds of thousands of

Issues in Applied Linguistics
ISSN 1050-4273

Vol. 1 No. 1 1990 30-55

individuals take these tests each year, and it is likely that most of these individuals make some sort of personal decision, such as seeking employment, advancing in a career or beginning an educational program, that is determined partly by their scores on these tests. Furthermore, since many of these individuals will submit job or educational program applications to more than one employer or institution, it is probably safe to say that the number of individual career decisions affected in some degree by the results of these tests is well over one million annually.

While the EFL proficiency test batteries developed by Cambridge and ETS are designed to measure many of the same abilities, they nevertheless represent radically different approaches to language test development. The TOEFL is perhaps the prototypical "psychometric/structuralist" language test (Spolsky 1978), representing the best qualities of this approach, which emphasizes reliability and item analysis. Its complements, the Test of Spoken English (TSE) and Test of Written English (TWE), while still developed in the psychometric tradition, represent an expansion of the structuralist linguistic framework and incorporate features associated with a broader range of language abilities and test methods. The FCE, on the other hand, has been designed and developed largely in the tradition of the British examinations system which places more emphasis on expert judgment and institutional experience in the production, scoring and interpretation of test results.

In short, there are important differences in the approaches to educational measurement that characterize the FCE and the ETS tests of EFL. In conducting this study, we became increasingly aware of these differences, since they affected the way we processed and analyzed the data. But rather than causing us to prejudge one approach or the other, our awareness of these differences has, we believe, enriched our treatment of the data and, we hope, enlightened our interpretation of the results.

## Objectives

The results reported in this paper were obtained as part of a larger study, the Cambridge-TOEFL Comparability Study (CTCS), commissioned by Cambridge to examine the comparability of the Cambridge EFL tests (both the FCE and the Certificate of Proficiency in English) and the ETS tests of EFL (Bachman et al., 1989a). The first goal of the CTCS was short-term: to examine the comparability of abilities measured by these two EFL proficiency



test batteries. This involved two different but complementary approaches: 1) the qualitative content analysis of the two tests, including the specific language abilities and the types of test tasks employed, and 2) the quantitative investigation of patterns of relationships in examinee performance among the different tests. The results of the content analysis have been reported elsewhere (Bachman et al., 1988a; Bachman et al., 1988b). This paper will thus focus on the results of the quantitative analyses of performance on the FCE and the ETS tests of EFL.

The second goal of the study is long-term -- to initiate a program of research and development with two aims: 1) improving the content and measurement characteristics of the Cambridge EFL examinations by understanding better what language abilities they measure, and 2) investigating the nature of communicative language ability and its measurement. This long-term goal will be accomplished in a number of ways, including, but not limited to the following:

1) routine monitoring of the reliability of Cambridge EFL examinations and refinement of procedures for maximizing reliability;
2) research into the patterns of performance across the various parts of the Cambridge EFL examinations;
3) research into the relationships between test-taker characteristics and test performance;
4) research into the analysis of test content and the subsequent refinement of the models upon which this study based and of the procedures for operationalizing these models in test design and specifications;
5) research into the relationships between aspects of test content and test performance.

These last two areas represent a particularly important concern for both test developers and test users, since recent work has indicated a poor relationship between what "experts" believe a given test item measures and test takers' performance (Alderson & Lukmani, 1986; Alderson et al., 1987; Perkins & Linnville, 1987; Bachman et al., 1989b; Alderson, 1990).

# PROCEDURES

## Subjects

### *Characteristics of Typical FCE Candidates*

Candidates for the FCE represent a widely varied population and have many reasons for seeking FCE certification. For example, if a candidate is working in a clerical or support position where English proficiency is required in a business or in the government service , an FCE certificate might be one criterion for promotion or salary increase. In addition to these adult candidates, large numbers of FCE takers are of school age; in many parts of the world, particularly in countries without formal institutional examinations for school-leaving, the FCE is taken as a *de facto* school-leaving examination. It is definitely not the case that candidates take the FCE for purposes of university entrance, a function served by the Certificate of Proficiency in English.

Although systematic data on native language background, educational status, age, sex and amount of non-FCE prior English study are not available, it is probably safe to assume that among the FCE population there is a wide variation in background variables. It is also likely that candidates are familiar with the FCE format, since the vast majority of FCE candidates have undergone a prior course of instruction covering the syllabus on which the FCE is based.

### *Characteristics of Typical TOEFL Takers*

Extensive information about TOEFL takers is available in several published ETS reports (ETS, 1987; Wilson, 1982; Wilson, 1987). According to these reports, typical TOEFL takers, in contrast to typical FCE candidates, are largely "degree planners" -- individuals planning to enter a college or university degree program in the U. S. or Canada -- (80%), male (72%) and have median ages of 20.6 and 25.4 for undergraduate and graduate level degree planners, respectively. Furthermore, 28% of the individuals who took the TOEFL for the first time in 1977 and 1978 had taken the test two or more times by 1982, with much higher percentages of test repeaters (40-50%) being reported for three CTCS sites: Hong Kong, Thailand and Japan.



### *Sampling Procedures*

The sampling procedures followed in the study were intended to insure that sites and subjects would constitute a representative sample of the operational worldwide populations of the TOEFL and FCE. The first consideration in selecting sites was the need to represent different geographic regions and native languages in the sample. Based on the geographic distributions of the Cambridge and TOEFL operational populations, we decided to include sites in the Far East, Middle East, Europe and South America. Within each of these regions we attempted to obtain representative samples of different language groups: Chinese, Japanese and Thai for the Far East; Arabic for the Middle East; Spanish, French and German for Europe; Spanish and Portuguese for South America.

The second consideration in selecting sites for the study was the need to obtain a representative sample both of typical FCE candidates and of typical TOEFL takers. We identified two types of sites: 1) "Cambridge-dominant," which were identified from Cambridge records as having large numbers of FCE candidates, and 2) "TOEFL-dominant," which generally had small numbers of FCE candidates and, according to published ETS reports (e.g., ETS, 1987), relatively large numbers of TOEFL test takers.

Finally, the availability of local persons to participate as Site Coordinators, who could assure that they were able to administer both sets of tests under operational conditions, and the availability of adequate numbers of subjects obviously had to be taken into account. After lengthy discussions revolving around these areas of concern, and considerable negotiation, the following eight sites were agreed upon: (a) TOEFL-dominant sites: Bangkok, Cairo, Hong Kong, Osaka; and (b) Cambridge-dominant sites: Madrid, São Paulo, Toulouse, Zurich.

At the Cambridge-dominant sites, the Site Coordinator was the examinations officer in charge of Cambridge EFL tests, while the Site Coordinator at the TOEFL-dominant sites was a staff member at an institution of higher education who was not only familiar with the local population of "typical" TOEFL-takers, but who also had access to adequate numbers of these individuals from which to draw a sample. This person generally worked closely with the local Cambridge examinations officer in arranging the schedule for test administration.

### *Characteristics of the Sample Subjects*

***Test performance*** Summary descriptive statistics for test or paper scores for the sample subjects are given in Table 1. Inspection of these score distributions indicated that all the measures were reasonably normally distributed and that the distributional assumptions for parametric statistical analyses were warranted. In order to examine the extent to which the sample subjects were typical of their respective operational populations, the sample means were compared with relevant norm groups. For the FCE, the norm group was all individuals who took the FCE in December, 1988 (University of Cambridge Local Examinations Syndicate 1988), while for the ETS tests, the norm groups were those reported in the most recent editions of the TOEFL Test and Score Manual (ETS, 1987), the Test of Spoken English Manual for Score Users (ETS, 1982) and the Test of Written English Guide (ETS, 1989).

The sample means and standard deviations on the FCE and the ETS tests as well as those of the FCE and ETS norm groups are presented in Tables 2 and 3, respectively. While virtually all the differences between the sample and norm group means were statistically significant, this is primarily a function of the large sample sizes, and thus it makes more sense to consider the practical importance of the differences. Looking at Table 2, we see that the largest difference between the sample and FCE norm group means is for Paper 2: 1.77 points on a scale of 40. The differences between the sample and ETS norm group means, presented in Table 3, are also small, relative to the standard deviations of the ETS norm group. The largest differences here was for the TSE/SPEAK Comprehensibility rating, with a mean difference of 19.4 on a scale of 300. Thus, although the sample means tend to be slightly lower than those of their relevant norm groups, these differences are too small to be of practical importance, and it can be concluded that the sample subjects constitute a representative sample of the FCE and ETS operational populations in terms of their test performance.

***Test takers' characteristics*** Information on subjects' age and sex was obtained from responses to questions on their TOEFL answer sheets, while information on their current educational status was obtained from responses to questions on a background questionnaire. The majority of the CTCS subjects were enrolled as students, whether at the secondary school level (21.3%), the college level (full-time, 27.6%; part-time, 10.4%), or in a language institute or other English course (17%), while 23.7% indicated that they



were not enrolled as students. The median age was 21, the youngest test taker being 14 years of age and the oldest 58. Slightly over half (59.4%) were female.

Information on the characteristics of TOEFL examinees from 1977-1979, provided by Wilson (1982), was used as a basis for comparing the sample subjects' characteristics with those of the operational TOEFL population. Since no such information is available for the operational FCE population, no comparisons could be made. Wilson did not include current educational status in his study, but the mean age for his population was 21.4 for individuals intending to apply to an undergraduate degree program ("undergraduate level degree planners") and 26.3 for graduate level degree planners, as compared to 22.7 for the sample group. A larger difference between the operational TOEFL population and the sample can be seen in the sex of the test takers, with Wilson reporting that 72% of his group was male, compared to about 41% for the sample subjects.

TABLE 1
Score Distributions, All Measures

| VARIABLE | MEAN | STD DEV | MIN | MAX | N |
|---|---|---|---|---|---|
| TOEFL 1 | 49.619 | 6.665 | 29 | 68 | 1448 |
| TOEFL 2 | 51.118 | 6.900 | 25 | 68 | 1448 |
| TOEFL 3 | 51.489 | 6.696 | 28 | 66 | 1448 |
| TOEFL TOTAL | 507.427 | 58.863 | 310 | 647 | 1448 |
| TEW | 3.929 | .891 | 1 | 6 | 1398 |
| SPEAK GRAM. | 1.934 | .454 | 0 | 3 | 1304 |
| SPEAK PRON. | 2.134 | .380 | 0 | 3 | 1314 |
| SPEAK FLUENCY | 1.945 | .440 | 0 | 3 | 1304 |
| SPEAK COMP. | 201.572 | 40.906 | 50 | 300 | 1304 |
| FCE PAPER 1 | 25.945 | 4.896 | 10 | 40 | 1359 |
| FCE PAPER 2 | 24.303 | 6.043 | 0 | 40 | 1357 |
| FCE PAPER 3 | 24.861 | 5.706 | 1 | 40 | 1353 |
| FCE PAPER 4 | 13.600 | 3.175 | 4 | 20 | 1344 |
| FCE PAPER 5 | 27.203 | 5.951 | 1 | 40 | 1381 |

TABLE 2
Differences between CTCS Group Means, Standard Deviations and FCE Norms

| | N | | MEAN | | STD DEV | |
|---|---|---|---|---|---|---|
| Test | CTCS | Norm | CTCS | Norm | CTCS | Norm |
| Paper 1 | 1,359 | 30,816 | 25.95 | 27.19 | 4.90 | 5.19 |
| Paper 2 | 1,357 | 30,818 | 24.30 | 26.07 | 6.04 | 5.22 |
| Paper 3 | 1,353 | 30,805 | 24.86 | 26.30 | 5.71 | 5.26 |
| Paper 4 | 1,344 | 30,936 | 13.60 | 14.47 | 3.18 | 3.25 |
| Paper 5 | 1,381 | 31,040 | 27.20 | 28.04 | 5.95 | 5.72 |

TABLE 3
Differences between CTCS Group Means, Standard Deviations and ETS Norms

| | N | | MEAN | | STD DEV | |
|---|---|---|---|---|---|---|
| Test | CTCS | Norm | CTCS | Norm | CTCS | Norm |
| TOEFL 1 | 1,448 | 714,731 | 49.60 | 51.20 | 6.70 | 6.90 |
| TOEFL 2 | 1,448 | 714,731 | 51.10 | 51.30 | 6.90 | 7.70 |
| TOEFL 3 | 1,448 | 714,731 | 51.50 | 51.10 | 6.70 | 7.30 |
| TOEFL Tot | 1,448 | 714,731 | 507.40 | 512.00 | 58.90 | 66.00 |
| TWE/TEW | 1,398 | 230,921 | 3.93 | 3.64 | 0.89 | 0.99 |
| TSE/SPEAK | | | | | | |
| GRAM | 1,304 | 3,500 | 1.93 | 2.43 | 0.45 | 0.39 |
| PRON | 1,314 | 3,500 | 2.13 | 2.10 | 0.38 | 0.49 |
| FLCY | 1,304 | 3,500 | 1.95 | 2.15 | 0.44 | 0.45 |
| COMP | 1,304 | 3,500 | 201.57 | 221.00 | 40.91 | 45.00 |

In summary, the sample subjects appeared to be quite similar to the operational populations of both the TOEFL and the FCE in terms of their test performance. With respect to test-taker characteristics, the sample was quite close in age to TOEFL undergraduate degree planners, but had a higher proportion of females than is typical of TOEFL test takers.

## Test Instruments

### *First Certificate in English (FCE)*

FCE Paper 1, entitled "Reading Comprehension," includes two sections of 4-choice multiple-choice items: 25 items which appear to test use or usage and ten items based on reading passages.



FCE Paper 2, entitled "Composition," consists of five prompts from which the candidate chooses two, writing 120-180 words in response to each. FCE Paper 3, entitled "Use of English," includes items that appear to test various aspects of lexicon, register and other elements of English usage. Paper 4 is a tape-recording-plus-booklet test, entitled "Listening Comprehension," for which candidates listen to several passages and respond to items per passage. Paper 5 consists of a face-to-face oral interview conducted as either a "one-on-one," with one examiner and one candidate, or as a "group" interview, with more than one candidate and two examiners.

### *Test of English as a Foreign Language (TOEFL)*

Early in the planning of the study it was obvious that we would not be able to synchronize the operational administrations of the two test batteries; we therefore decided to use the institutional versions of the TOEFL and the TWE as well as a composition test similar to the TWE. The institutional TOEFL consists of those official international forms of the TOEFL that have been retired from operational use. While ETS does not report scores on the institutional TOEFL to other institutions, it does guarantee content and statistical equivalence of the institutional and international forms of the TOEFL. There are three sections to the test: Section 1 - Listening Comprehension; Section 2 - Structure and Written Expression; Section 3 - Vocabulary and Reading Comprehension. Item types vary somewhat, but all follow a four-option multiple-choice format.

### *Speaking Proficiency in English Assessment Kit (SPEAK)*

The SPEAK is a semi-direct test of oral performance and is the institutional counterpart of the Test of Spoken English (TSE), consisting of retired forms of the operational TSE. It consists of a complete kit, including materials for training raters in the scoring procedure and is administered entirely by tape recorder, with candidates listening to a number of prompts from a cassette source tape, looking at verbal and graphic stimulus material in an accompanying booklet and responding on a target cassette tape which also records the prompts from the source tape.

### *Test of English Writing (TEW)*

ETS produces a composition test called the Test of Written English (TWE) which, because it was still considered experimental by ETS at the time the study began, had no institutional counterpart. An experienced TWE rater was therefore asked to produce a prompt similar to example prompts that ETS makes available in its information to prospective TWE takers. We called this test the "Test of English Writing" (TEW). The TEW test booklet contained two printed sheets: on the first were the instructions, while the second contained the single prompt, including both verbal and graphic information, to be answered by all candidates in the study.

### Test Administration

Because it was not possible to synchronize operational administrations of the FCE and the ETS tests, we decided to administer all tests within the schedule of an operational FCE administration (December, 1988). This meant that candidates generally took the FCE and ETS pencil-and-paper sections (i.e., FCE Papers 1, 2 and 3, TOEFL and TEW) on adjacent days. FCE Papers 4 and 5 were administered within the five weeks devoted to Cambridge Papers 1, 2 and 3, while at most sites the SPEAK was given in the same two-day period as the pencil-and-paper tests. Operational procedures and time allocations prescribed by all tests were strictly adhered to.

### Scoring Procedures

All tests were scored according to operational procedures prescribed by Cambridge and ETS. FCE Paper 1 was scored by optical scanner at Cambridge, while answers to Papers 3 and 4 were hand scored using scoring keys or "marking schemes" prepared by examiners. Papers 2 and 5 were rated subjectively by trained examiners, using rating scales developed by Cambridge. The TOEFL was scored by optical scanner at Illinois, while the SPEAK and TEW were rated subjectively by trained raters in North America, according to rating scales developed by ETS.



### Data Preparation and Analyses

Data from all the FCE papers were prepared in Cambridge. While Paper 1 answer sheets were optically scanned, the majority of the data from the other papers were manually entered into computer files. The machine-scorable answer sheets for the TOEFL and SPEAK were optically scanned, while the TEW ratings were manually entered into computer files at the University of Illinois at Urbana-Champaign. Subsequent data merging and data file assembly were performed using SAS or PC-SAS (SAS, 1988). Statistical analyses were performed using SPSS-X Version 3.0 (SPSS, 1988a), SPSS/PC+ Version 2.1 (SPSS, 1988b; SPSS, 1988c), GENOVA (Crick & Brennan, 1983) and factor analysis programs written for the PC by John B. Carroll (Carroll, 1989).

## MEASUREMENT CHARACTERISTICS OF TWO TEST BATTERIES

### Reliability

Classical internal consistency estimates (coefficient alpha) were calculated for the discrete-item tests (FCE Papers 1, 3, 4 and TOEFL). Since two forms of FCE Paper 1 and ten forms of Paper 4 were administered in the study, the reliabilities reported here for these papers are the weighted averages (using Fisher's Z transformation) of the coefficient alphas for the different forms of each paper. Inter-rater reliabilities for the TWE and SPEAK ratings were estimated using generalizability theory. The values reported here were obtained from a single facet G-study design with raters as the facet. Because multiple independent ratings are not done as part of the operational FCE, inter- and intra-rater reliabilities could not be estimated for FCE Papers 2 (composition) and 5 (oral interview).

Reliability estimates for the sample, along with their respective population norms, are reported in Table 4. The norm for FCE Paper 1 is the reported KR-20 operational examinees who took the FCE between December 1988 and December 1989, while the norms for FCE Papers 3 and 4 consist of average KR-20s (using Fisher's transformation to Z) across samples of operational examinees who took the FCE in December, 1989. Norms for the ETS tests are those reported in the ETS score and interpretative manuals cited above. In general, the FCE reliabilities for our sample are slightly lower than those generally obtained operationally by Cambridge and somewhat below the norms reported for the ETS

tests. While the reliability estimate obtained for Paper 3 is within acceptable limits, those obtained for Papers 1 and 4 are lower than normally acceptable for standardized tests.

## Comparability of Abilities Measured

In order to determine whether and to what extent the two test batteries (FCE and ETS) might be comparable, it was first necessary to investigate the extent to which patterns of performance support interpretations of similar abilities. This was done primarily by examining the patterns of correlations within and across the two test batteries through exploratory factor analysis.

Three correlation matrices were analyzed: 1) intercorrelations among the scaled scores for the five FCE papers, 2) intercorrelations among the eight ETS standard scores, and 3) intercorrelations among all 13 of these measures. These correlation matrices are given in the Appendix. The matrix of product-moment correlations among the various test scores to be analyzed was examined for appropriateness of the common factor model in several ways. Principal axes were extracted with squared multiple correlations on the diagonal as initial communality estimates. The eigenvalues obtained from the initial extraction were plotted on a scree plot, an examination of which, along with differences between successive eigenvalues, led to an initial decision about the appropriate number of factors to extract. Specified numbers of principal axes, generally including one fewer and one more than the number initially decided upon for a given correlation matrix, were then successively extracted and rotated. Two rotated factor structure matrices were obtained for each number of principal axes extracted: an orthogonal solution with the normal varimax procedure and an oblique solution with Tucker & Finkbeiner's (1981) least-squares hyperplane fitting ("DAPPFR"). The final determination of the number of factors and the "best" solution was made on the basis of simplicity and interpretability, these qualities being judged, of course, subjectively. Simplicity was evaluated by examining both the patterns of salient loadings for the orthogonal and oblique solutions and the scatter plots of loadings on the rotated axes. Interpretability was evaluated with reference to the extent to which the salient factor loadings and factor correlations reflected the nature of the tasks and the abilities thought to be operationalized in the different measures.



TABLE 4
Reliability Estimates

| Test | k | N | a | Norm |
|---|---|---|---|---|
| FCE Paper 1[1] | 40 | 1,394 | .791 | .901[1] |
| FCE Paper 2 | (Not available) | | | |
| FCE Paper 3 | 52 | 995 | .847 | .870[1] |
| FCE Paper 4[1] | 27 | 759 | .616 | .705[1] |
| FCE Paper 5 | (Not available) | | | |
| TOEFL 1 | 50 | 1,467 | .889 | .900 |
| TOEFL 2 | 34 | 1,467 | .834 | .860 |
| TOEFL 3 | 58 | 1,467 | .874 | .900 |
| TEW[2] | | 1,399 | .896 | .860 |
| Speak Comp[2] | | 1,318 | .970 | .880 |

1 Weighted average coefficient alphas across more than one test form
2 Generalizability coefficients

### *Within-Paper/Test Factor Structures*

The results of the exploratory factor analysis for the FCE paper scores are given in Table 5, while those for the ETS test scores appear in Table 6. The scree plots and the parallel analyses suggested that two factors underlay both sets of test scores, while the oblique rotations yielded factors that were highly correlated for both sets of tests. Since both of these factor solutions were highly oblique, with correlations between factors of .826 for the FCE and .601 for the ETS tests, a Schmid-Leiman transformation to orthogonal primary factors with a second-order general factor was performed on each (Schmid & Leiman, 1957).

All the FCE papers loaded most heavily on a higher-order factor, which accounted for 50.3% of the common variance. The first primary factor was characterized by high loadings on Paper 1, 2 and 3, while Papers 4 and 5 had high loadings on the second primary factor. This suggests that the FCE Papers all tend to measure a common component of the subjects' English language ability, with two specific ability factors, "reading, structure and writing" and "speaking and listening," being identified.

This pattern of loadings was repeated for the ETS tests, with all tests loading most heavily on a second-order general factor that accounted for 43.1% of the common variance. The ETS tests, except perhaps for TOEFL Section 1, also had high loadings on the two primary factors, with TOEFL Section 1 and the SPEAK ratings

loading most heavily on the first, and with TOEFL Sections 2 and 3 and the TEW loading on the second.

TABLE 5
Exploratory Factor Analysis of FCE Papers

| VARIABLE | COMMUNALITY |
|---|---|
| Paper 1 | .54835 |
| Paper 2 | .48888 |
| Paper 3 | .62272 |
| Paper 4 | .41468 |
| Paper 5 | .32595 |

| FACTOR | EIGENVALUE | % OF VAR | CUM % |
|---|---|---|---|
| 1 | 3.18529 | 63.7 | 63.7 |
| 2 | .63769 | 12.8 | 76.5 |
| 3 | .48866 | 9.8 | 86.2 |
| 4 | .41719 | 8.3 | 94.6 |
| 5 | .27117 | 5.4 | 100.0 |

Orthogonalized Factor Matrix with Second-Order General Factor

| | GENERAL FACTOR | FACTOR 1 | FACTOR 2 | $h^2$ |
|---|---|---|---|---|
| Paper 1 | .733 | .275 | .062 | .617 |
| Paper 2 | .689 | .260 | .057 | .546 |
| Paper 3 | .809 | .433 | .061 | .846 |
| Paper 4 | .679 | .071 | .241 | .524 |
| Paper 5 | .622 | .024 | .310 | .484 |
| Eigenvalue | 2.515 | .336 | .165 | 3.016 |
| % of $h^2$ | 50.300 | 6.700 | 3.300 | 60.300 |

These results suggest that the ETS tests also tend to measure a common component of the subjects' language ability, with specific factors associated with listening and speaking, on the one hand, and reading, structure and writing, on the other, being identified.

These two sets of test scores show remarkable similarities in their factor structures, with higher-order general factors accounting for large portions of the common variances in the two test batteries. But whereas relatively little common variance in the FCE papers is accounted for by first-order factors (10%), in the ETS tests the two first-order factors account for a considerable proportion (26.5%) of the common variance. This suggests that while each test battery



appears to measure a single language ability, the ETS tests provide relatively more information about specific language abilities than do the FCE papers. While these similarities in factor structures would appear to reflect similarities in the abilities of the subjects in the study, they also suggest that these two sets of tests measure these abilities in much the same way.

TABLE 6
Exploratory Factor Analysis of ETS Tests

| VARIABLE | COMMUNALITY |
|---|---|
| TOEFL 1 | .53783 |
| TOEFL 2 | .57999 |
| TOEFL 3 | .57389 |
| TEW | .37555 |
| SPEAK GRAM | .80099 |
| SPEAK PRON | .60725 |
| SPEAK FLCY | .77096 |
| SPEAK COMP | .89596 |

| FACTOR | EIGENVALUE | % OF VAR | CUM % |
|---|---|---|---|
| 1 | 4.93914 | 61.7 | 61.7 |
| 2 | 1.17112 | 14.6 | 76.4 |
| 3 | .55103 | 6.9 | 83.3 |
| 4 | .39869 | 5.0 | 88.2 |
| 5 | .38118 | 4.8 | 93.0 |
| 6 | .28032 | 3.5 | 96.5 |
| 7 | .20493 | 2.6 | 99.1 |
| 8 | .07357 | 0.9 | 100.0 |

Orthogonalized Factor Matrix with Second-Order General Factor

| | GENERAL FACTOR | FACTOR 1 | FACTOR 2 | $h^2$ |
|---|---|---|---|---|
| TOEFL 1 | .654 | .275 | .258 | .569 |
| TOEFL 2 | .648 | -.004 | .532 | .703 |
| TOEFL 3 | .642 | -.036 | .559 | .726 |
| TEW | .534 | .094 | .341 | .410 |
| SPEAK GRAM | .668 | .576 | -.032 | .779 |
| SPEAK PRON | .651 | .404 | .126 | .603 |
| SPEAK FLCY | .684 | .567 | -.009 | .791 |
| SPEAK COMP | .750 | .650 | -.038 | .986 |
| Eigenvalue | 3.444 | 1.325 | .798 | 5.567 |
| % of $h^2$ | 43.100 | 16.600 | 9.900 | 69.600 |

### *Across Battery Factor Structures*

In order to examine the relationships between the two test batteries, the correlations among the scaled scores for the five FCE Papers and for the eight ETS test scores were analyzed using the procedures described above. Although the scree test suggested that only two or three factors should be extracted, the parallel analyses criterion indicated five. Therefore, orthogonal and oblique solutions with two, three, four and five principal axes were examined. The solution that appeared to optimize the simplicity and interpretability criteria was a four-factor oblique solution with highly correlated factors. The higher-order solution which the Schmid-Leiman transformation produced is presented in Table 7. As would be expected with very high correlations among the first-order factors, all of the measures have salient loadings on the second-order general factor, which accounts for 49.2% of the common variance. The first-order factors can be characterized as follows: Factor 1 (10.6% of common variance) - SPEAK ratings and FCE Paper 5; Factor 2 (4.4%) - TOEFL Sections 2, 3 and TEW; Factor 3 (1.9%) - FCE Papers 1, 2, 3; and Factor 4 (1.5%) - FCE Paper 4 and TOEFL Section 1.

These loadings suggest that all these tests measure, to a considerable degree, a common portion of the language abilities that characterize the test takers in the sample. After this general or common ability, the next largest component appears to be associated with speaking ability. This is followed by two components that appear to be combinations of ability (reading, structure and writing) and test method ("ETS test method" and "FCE test method"). Finally, there is a relatively small component associated with listening ability. Given that all of the measures examined load most heavily on a higher-order general factor and that two of the first-order factors appear to be associated with aspects of language ability (speaking and listening) across both tests, these tests do, in general, appear to measure the same abilities. That the other two factors appear to be associated in part with specific tests suggests that some of the observed differences in performance across the two test batteries are attributable to differences in the methods used in testing.



TABLE 7
Exploratory Factor Analysis of FCE Papers and ETS Tests

| VARIABLE | COMMUNALITY |
|---|---|
| FCE Paper 1 | .58958 |
| FCE Paper 2 | .52228 |
| FCE Paper 3 | .66459 |
| FCE Paper 4 | .48009 |
| FCE Paper 5 | .42786 |
| TOEFL 1 | .59892 |
| TOEFL 2 | .60101 |
| TOEFL 3 | .61938 |
| TEW | .39597 |
| SPEAK GRAM | .80465 |
| SPEAK PRON | .62949 |
| SPEAK FLCY | .77734 |
| SPEAK COMP | .89563 |

| FACTOR | EIGENVALUE | % OF VAR | CUM % |
|---|---|---|---|
| 1 | 7.48415 | 57.6 | 57.6 |
| 2 | 1.32523 | 10.2 | 67.8 |
| 3 | .65258 | 5.0 | 72.8 |
| 4 | .57734 | 4.4 | 77.2 |
| 5 | .55311 | 4.3 | 81.5 |
| 6 | .50183 | 3.9 | 85.3 |
| 7 | .38833 | 3.0 | 88.3 |
| 8 | .37212 | 2.9 | 91.2 |
| 9 | .34312 | 2.6 | 93.8 |
| 10 | .27587 | 2.1 | 96.0 |
| 11 | .25262 | 1.9 | 97.9 |
| 12 | .20044 | 1.5 | 99.4 |
| 13 | .07325 | 0.6 | 100.0 |

TABLE 7 (Continued)
Orthogonalized Factor Matrix with Second-Order General Factor

| | GENERAL FACTOR | FACTOR 1 | FACTOR 2 | FACTOR 3 | FACTOR 4 | $h^2$ |
|---|---|---|---|---|---|---|
| FCE Paper 1 | .754 | -.031 | .078 | .175 | .104 | .617 |
| FCE Paper 2 | .711 | .074 | -.023 | .270 | .002 | .584 |
| FCE Paper 3 | .820 | -.026 | .028 | .341 | -.004 | .789 |
| FCE Paper 4 | .704 | -.004 | -.048 | .053 | .236 | .556 |
| FCE Paper 5 | .621 | .173 | -.028 | .015 | .165 | .443 |
| TOEFL 1 | .776 | .058 | .059 | -.043 | .284 | .692 |
| TOEFL 2 | .680 | .049 | .573 | -.004 | .010 | .793 |
| TOEFL 3 | .711 | -.085 | .419 | .032 | .102 | .699 |
| TEW | .581 | .074 | .223 | .097 | .012 | .402 |
| SPEAK GRAM | .642 | .621 | .015 | -.000 | -.000 | .798 |
| SPEAK PRON | .707 | .333 | -.026 | .131 | .030 | .630 |
| SPEAK FLCY | .676 | .555 | -.021 | .007 | .045 | .767 |
| SPEAK COMP | .705 | .719 | .040 | .011 | -.034 | 1.018 |
| Eigenvalue | 6.401 | 1.377 | .570 | .252 | .189 | 8.789 |
| % of $h^2$ | 49.200 | 10.600 | 4.400 | 1.900 | 1.500 | 67.600 |

## DISCUSSION

### Adequacy of Sample

In terms of test performance, the sample subjects were representative of both the December, 1988 FCE candidature and "typical" ETS test takers. In terms of test-taker characteristics, the sample subjects were similar in age to ETS undergraduate degree planners, but they included a higher proportion of female test takers than in the general ETS population. Since little is known about the characteristics of typical FCE candidates, no generalizations can be made in this regard.

### Reliability

FCE Papers 1, 3 and 4 were somewhat less reliable than the ETS tests, while the reliabilities of Papers 2 and 5 could not be estimated. Although this does not necessarily mean that Papers 2 and 5 are unreliable, the inability to estimate their reliability is a recognized deficiency that will be remedied through the on-going program of research and development described below.



### Comparability of Abilities Measured

The factor structure for any given set of test scores will be a function of both the profile of language abilities of the specific group(s) of individuals tested and the characteristics of the specific tests used. The large proportions of variance accounted for by the general factors in our analyses suggest that the FCE papers and ETS tests administered in this study appear to measure, to a large degree, the same common aspect of the language proficiency of the subjects in our sample. We feel that at present there is no basis for interpreting this general factor as anything other than a common aspect of language proficiency shared by these subjects as measured by these tests. That is, this general factor does not necessarily represent the same aspect of language proficiency as do the general factors that have been found in other sets of language tests with other groups of subjects (e.g., Oller, 1979; Carroll, 1983; Bachman & Palmer, 1982; Sang et al., 1986).

In addition to a common, general aspect of language ability, the test batteries in our study appear to reflect shared specific abilities and different testing formats. The primary factor that accounts for the largest proportion of variance is associated with measures of speaking, especially the SPEAK. The primary factor that accounts for the least amount of variance is associated with measures of listening. The FCE oral interview loads almost equally on both the speaking and listening factors. A third primary factor (associated with ETS measures of structure, reading and writing) can be identified as an "ETS written test factor," while a fourth primary factor (associated with the FCE measures of structure, writing and reading) can be identified as an "FCE written test factor."

### Score Comparisons Across Test Batteries

Since the forms of the FCE and ETS tests of EFL examined in this study appear to measure nearly the same aspects of the subjects' English language proficiency, score comparisons across tests are justified and could be made in a meaningful way. However, because of differences in the levels of reliability across the two test batteries, as well as a lack of demonstrated equivalence of different FCE forms, such score comparisons could best be made on the basis of subsequent studies, once reliability and equivalence are better assured.

## Future Research

This study has provided an opportunity for Cambridge to study its EFL examinations in a way that has not generally been done in the past, which includes an on-going program of research and development consistent with the long-range objectives of the study. This research, which will address both practical test development issues and research questions that are of theoretical interest to the field of language testing, currently includes the following specific projects:

1. *The investigation into the reliability of Paper 2 and Paper 5 ratings.* This will involve a 3-facet G-study for Paper 2, with rater, occasion and topic as facets, using papers from several different administrations, as well as a 3-facet G-study for Paper 5, with interview mode (individual vs. group), rater and occasion as facets, using taped interviews from the June 1990 administration.
2. *The investigation into the comparability of FCE forms.* This will involve two stages: 1) investigation of content and comparability, followed by the establishment of procedures for insuring content comparability across forms, and 2) investigation of statistical equivalence of forms.
3. *The investigation into the relationship between test content and test performance.* This will involve both a continuation and extension of the type of analyses of this study's data that have been reported elsewhere and the content analysis of new forms of the FCE.
4. *The investigation into the relationship between test taker characteristics and test performance.* This will involve the analysis of data on test-taker characteristics in this study as well as the development of a questionnaire to be used to operationally gather such information on a regular basis.
5. *The investigation into the relationship between self-reported test-taking strategies and test performance.*
6. *The setting of standards for the content, design, development and use of language proficiency tests.* These would initially be standards to be used by Cambridge for its own EFL exams, but would hopefully provide a basis for developing more general standards for language tests.



## Contributions of This Study to Language Testing

One of the most pressing issues in the field of foreign language testing at present is that of defining the construct "communicative competence" precisely enough to permit its assessment. A related issue involves defining what we mean by a "communicative" or "authentic" test and determining whether test takers perform differentially on "communicative" and "non-communicative" language tests. These issues are of crucial importance for the development and use of language tests, since considerable effort is currently being expended in developing "communicative" language tests to measure "communicative competence" or "communicative language ability." The content analysis instruments developed as part of the CTCS, based on theoretical models of communicative language ability and test method facets (Bachman, 1990), provide a starting point for accurately describing the content of language tests and for investigating the relationship between test content and test performance.

Furthermore, while the focus of the CTCS was not on construct validation, much of the information that was gathered about the measures examined is relevant to the validity of construct interpretations. In this regard, the finding that measures as diverse as those examined in the CTCS tap virtually the same sets of language abilities is remarkable, although not particularly surprising, given the long history of such findings in language testing. At the same time, it is encouraging to find both that the theoretical constructs which are claimed to inform the measures are reflected, to a large degree, in patterns of performance and that the methodological approaches employed are useful in making these patterns interpretable.

Since the CTCS employed a variety of empirical research approaches, both qualitative and quantitative, the experience gained may thus be useful not only for continued multi-modal research, as has been proposed by Bachman & Clark (1987), but also for future test comparison studies. While some operational procedures that were planned had to be either changed or abandoned in the course of the study, we believe that in general the CTCS design, procedures and analyses provide a useful model for the comparison of different batteries of language tests.

Finally, the CTCS explored the complexities of cross-national comparative research, which involve issues such as the types of negotiations and compromises that are necessary in such

studies, whether these compromises vitiate or enrich the results of cross-national research and whether the results are worth the effort. This complex topic is taken up in Davidson & Bachman (forthcoming). Suffice it here to say that although we feel the CTCS illustrates the benefits to be derived from cross-national comparison studies, we have no delusions of having resolved these issues in this study. However, by bringing them to the fore we believe we have made some contribution to a better understanding of the similarities and differences between two approaches to EFL proficiency testing, an understanding that we hope will spur collaborative projects in which the subjective, qualitative judgments of "experts" are complemented by objective, quantitative research and development methods.

## Notes

[1]Revised version of a paper presented at the 12th Annual Language Testing Research Colloquium, San Francisco, 3-5 March 1990.

## REFERENCES


Alderson, J. C. (1990). Judgements in language testing. Paper presented at the 12th Annual Language Testing Research Colloquium, San Francisco.

Alderson, J. C. & Lukmani, Y. (1986). Reading in a second language. Paper presented at the 4th Colloquium on Research on Reading in a Second Language, 20th Annual TESOL Convention, Anaheim.

Alderson, J. C., Henning, G. & Lukmani, Y. (1987). Levels of understanding in reading comprehension tests. Paper presented at the 9th Annual Language Testing Research Colloquium, Miami.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. & Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science 490*, 20-33.

Bachman, L. F., Davidson, F. & Lynch, B. (1988a). Test method: The context for performance on language tests. Paper presented at the Annual Meeting of the American Association for Applied Linguistics.

Bachman, L. F., Davidson, F., Ryan, K. & Choi, I-C. (1989a). The Cambridge-TOEFL comparability study: Final report. Cambridge: University of Cambridge Local Examinations Syndicate.

Bachman, L. F., Davidson, F., Lynch, B. & Ryan, K. (1989b). Content analysis and statistical modeling of EFL proficiency tests. Paper presented at the 11th Annual Language Testing Research Colloquium, San Antonio,Texas.





Bachman, L. F., Kunnan, A., Vanniarajan, S. & Lynch, B. (1988b). Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries. *Language Testing 5*, 2, 128-59.

Bachman, L. F. & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 4, 449-65.

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 80-107). Rowley, MA: Newbury House.

Carroll, J. B. (1989). Exploratory factor analysis programs for the IBM PC (and compatibles). Chapel Hill: Author.

Christie, T. & Forrest, G. M. (1981). *Defining public examinations standards*. London: Macmillan Education.

Crick, J. E. & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.

Davidson, F. & Bachman, L. F. (forthcoming). The Cambridge-TOEFL comparability study: An example of the cross-national comparison of language tests. *AILA Review, 7*.

Educational Testing Service. (1982). *Test of spoken English: Manual for score users*. Princeton: Author.

Educational Testing Service. (1985). *Guide to SPEAK*. Princeton: Author.

Educational Testing Service. (1987). *TOEFL test and score manual (1987-88 ed.)*. Princeton: Author.

Educational Testing Service. (1989). *Test of Written English guide*. Princeton: Author.

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

Perkins, K. & Linnville, S. E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing 4*, 2, 125-141.

Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J. & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing 3*, 1, 54-79.

SAS Institute, Inc. (1988). *SAS guide to personal computers: Language* (Version 6). Cary, NC: Author.

Schmid, J. & Leiman, J. M. (1957). The development of heirarchical factor solutions. *Psychometrika 22*, 53-61.

Spolsky, B. (1978). Linguists and language testers. In B. Spolsky (Ed.), *Approaches to language testing* (pp. v-x). Arlington, VA: Center for Applied Linguistics.

SPSS Incorporated. (1988a). *SPSS-X user's guide* (3rd ed.). Chicago: Author.

SPSS Incorporated. (1988b). *SPSS/PC+ V2.0 base manual*. Chicago: Author.

SPSS Incorporated. (1988c). *SPSS/PC+ advanced statistics V2.0*. Chicago: Author.

Tucker, L. R. (n.d.). *Functional representation of Montanelli-Humphreys weights for judging the number of actors by the parallel analysis technique*. Champaign: Author.

Tucker, L. R. & Finkbeiner, C. T. (1981). *Transformation of factors by artificial personal probability functions* (Research Report No. 81-58). Princeton: Educational Testing Service.

University of Cambridge Local Examinations Syndicate. (1988). *Cambridge examinations in English: Survey for 1988.* Cambridge: Author.

Wilson, K. M. (1982). *A comparative analysis of TOEFL examinee characteristics, 1977-1979* (TOEFL Research Report No. 11). Princeton: Educational Testing Service.

Wilson, K. M. (1987). *Patterns of test taking and score change for examinees who repeat the Test of English as a Foreign Language* (TOEFL Research Report No. 22). Princeton: Educational Testing Service.

**Lyle F. Bachman** is a professor of applied linguistics at the University of California, Los Angeles. He is the author of *Fundamental Considerations in Language Testing* (Oxford University Press, 1990). He has also published in several journals regarding language testing and program evaluation.

**Fred Davidson** is an assistant professor of applied linguistics at the University of Illinois, Urbana-Champaign. His publications are in the areas of language testing and related applied linguistics concerns. His current research interests are language test construction and theory, decision theory and expert systems in language assessment, and issues in the assessment of language minority students in school settings.

**John Foulkes** is a test development officer in the Council for Examination Development, University of Cambridge Local Examinations Syndicate.

(Received March 29, 1990)



APPENDICES

APPENDIX A
Intercorrelations among FCE Papers

| | Paper 1 | Paper 2 | Paper 3 | Paper 4 | Paper 5 |
|---|---|---|---|---|---|
| Paper 1 | 1.00000 | | | | |
| Paper 2 | .58054 | 1.00000 | | | |
| Paper 3 | .70730 | .66754 | 1.00000 | | |
| Paper 4 | .53710 | .49427 | .56614 | 1.00000 | |
| Paper 5 | .46457 | .44448 | .47369 | .49554 | 1.00000 |

APPENDIX B
Intercorrelations Among ETS Tests

| | TOEFL 1 | TOEFL 2 | TOEFL 3 | TEW | SPK GRAM | SPK PRON | SPK FLCY | SPK COMP |
|---|---|---|---|---|---|---|---|---|
| TOEFL 1 | 1.00000 | | | | | | | |
| TOEFL 2 | .55197 | 1.00000 | | | | | | |
| TOEFL 3 | .56483 | .71597 | 1.00000 | | | | | |
| TEW | .44761 | .54038 | .51739 | 1.00000 | | | | |
| SPK GRAM | .58395 | .43680 | .38214 | .38866 | 1.00000 | | | |
| SPK PRON | .58779 | .46374 | .47350 | .45019 | .64165 | 1.00000 | | |
| SPK FLCY | .61484 | .41868 | .41163 | .43372 | .78437 | .67494 | 1.00000 | |
| SPK COMP | .64026 | .47326 | .44279 | .43503 | .89268 | .75249 | .87286 | 1.00000 |

APPENDIX C
Intercorrelations Among FCE Papers and ETS Tests

| | TOEFL 1 | TOEFL 2 | TOEFL 3 | TEW | SPK GRAM | SPK PRON | SPK FLCY | SPK COMP | FCE 1 | FCE 2 | FCE 3 | FCE 4 | FCE 5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TOEFL 1 | 1.00000 | | | | | | | | | | | | |
| TOEFL 2 | .55354 | 1.00000 | | | | | | | | | | | |
| TOEFL 3 | .56500 | .71801 | 1.00000 | | | | | | | | | | |
| TEW | .43919 | .54284 | .51782 | 1.00000 | | | | | | | | | |
| SPK GRAM | .58721 | .43657 | .38292 | .38729 | 1.00000 | | | | | | | | |
| SPK PRON | .58578 | .47240 | .47637 | .44754 | .64102 | 1.00000 | | | | | | | |
| SPK FLCY | .61371 | .42348 | .40753 | .42906 | .78855 | .66837 | 1.00000 | | | | | | |
| SPK COMP | .64193 | .47819 | .44714 | .43305 | .89324 | .74693 | .87416 | 1.00000 | | | | | |
| FCE 1 | .60579 | .57021 | .62738 | .44062 | .47310 | .52405 | .50064 | .53396 | 1.00000 | | | | |
| FCE 2 | .53629 | .51820 | .50162 | .47045 | .50129 | .56657 | .48799 | .53821 | .57632 | 1.00000 | | | |
| FCE 3 | .58680 | .62876 | .63841 | .52807 | .49123 | .59153 | .52199 | .54648 | .69970 | .66234 | 1.00000 | | |
| FCE 4 | .61123 | .44312 | .49670 | .42396 | .49685 | .52697 | .52437 | .53489 | .52980 | .49364 | .56273 | 1.00000 | |
| FCE 5 | .55493 | .40958 | .40419 | .36558 | .54278 | .53077 | .56055 | .57177 | .45530 | .45189 | .47376 | .49156 | 1.00000 |

# Orality, Oral-Based Culture, and the Academic Writing of ESL Learners

**Donald L. Rubin**
**Rosemarie Goodrum**
*University of Georgia*

**Barbara Hall**
*DeKalb College*

*Although ESL learners are often quite sensitive to interference of their native language in second language writing, they tend to be less aware of the interference of native culture rhetorical patterns in writing Western academic exposition. Conceptually and pedagogically, however, the construct of interference is inadequate because it implies that native linguistic and rhetorical abilities must be suppressed in order to achieve competence in the target culture's discourse. An alternative approach is to recognize that many native culture rhetorical patterns can be integrated into the discourse norms of academic writing, even when these discourse patterns are oral-based. Expert writers learn to reintegrate oral-based discourse strategies into their writing after becoming aware of the differences between written and oral codes. Although the concept of oral culture is problematic, oral-based cultures can be identified. ESL learners from oral-based cultures thus need not completely divorce themselves from their native rhetorical patterns when they learn to write academic English exposition. Instead, they can learn to capitalize on certain oral-based discourse strategies, such as metaphor and narrative as proof, direct second-person address and elements of redundancy.*

From time to time, we find sitting in our ESL writing classes an extraordinary writer whose voice (even in English) rings especially strongly with the cadences and tones of his or her native culture. Such a student is a mixed blessing: although we may relish

Issues in Applied Linguistics
ISSN 1050-4273
Vol. 1 No. 1 1990 56-76