DOCUMENT RESUME

ED 389 711 TM 024 191

AUTHOR	Henning, Grant; And Others
TITLE	Analysis of Proposed Revisions of the Test of Spoken English. TOEFL Research Reports 48.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-95-1
PUB DATE	Mar 95
NOTE	50p.
PUB TYPE	Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Adults; College Students; Comparative Analysis; *English (Second Language); Health Personnel; Higher Education; *Interrater Reliability; *Language Proficiency; Language Tests; Patients; Psychometrics; *Scoring; *Teaching Assistants; Test Construction; Test Validity
IDENTIFIERS	*Nonnative Speakers; Test of English as a Foreign Language; *Test of Spoken English

ABSTRACT

A prototype revised form of the Test of Spoken English (TSE) was compared with the current version of the same test, comparing interrater reliability, frequency of rater discrepancy at all score levels, component task adequacy, scoring efficacy, and other concurrent and construct validity evidence, including the oral proficiency interview correlations for a subsample of the examinees. The study employed a representative sample of 342 nonnative speakers of English, purposely sampled from two professional domains of prospective university teaching assistants (N=184) and prospective licensed medical professionals (N=158). In an attempt to involve the persons most at risk in the judgment process, 16 naive adult raters (8 first-year university students and 8 nondegreed medical outpatients) were used in addition to the usual group of trained raters. The naive raters provided concurrent judgments of the comprehensibility and communicative effectiveness of a subset of 40 examinations. In general, the evidence appeared to underscore the psychometric quality of the prototype revised TSE and to support conclusions of its adequacy as an instrument to make judgments of oral English language proficiency of nonnative speakers. Some suggestions on scoring are provided. Six appendixes give additional scoring information and include sample tests. (Contains 10 tables and 10 references.) (SLD)

RR-95-1

TOEFL®

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 48
MARCH 1995

Analysis of Proposed Revisions of the Test of Spoken English

Grant Henning
Mary Schedl
Barbara K. Suomi

ETS
Educational Testing Service

Analysis of Proposed Revisions of the Test of Spoken English

Grant Henning
Mary Schedl
Barbara K. Suomi

Educational Testing Service
Princeton, New Jersey

RR-95-1

Abstract

This research was conducted to compare a prototype revised form of the Test of Spoken English (TSE®) with the current version of the same test. The study compared interrater reliability, frequency of rater discrepancy at all score levels, component task adequacy, scoring efficacy, and other concurrent and construct validity evidence, including oral proficiency interview correlations for a subset of the examinee sample. The study employed a representative examinee sample of 342 nonnative speakers of English, purposely sampled from among the two professional domains of prospective university teaching assistants ($\underline{N}$ = 184) and prospective licensed medical professionals ($\underline{N}$ = 158).

One somewhat unusual component of the study was the attempt to involve persons most at risk in the judgment process. Thus, in addition to employing the usual group of trained raters for the scoring of the current and prototype versions of the test, 16 naive adult raters were purposely selected (eight first-year university students from four broad academic disciplines and eight nondegreed prospective medical outpatients within four broad age levels) for having limited exposure to foreign languages and cultures. These 16 naive raters (eight females and eight males) provided concurrent judgments of the comprehensibility and communicative effectiveness of a subset of 40 recorded prototype examinations.

In general, the comparative evidence gathered appeared to underscore the psychometric quality of the prototype revised TSE and to support conclusions of its adequacy as an instrument used to make judgments of the oral English language proficiency of nonnative speakers in the targeted populations. Some additional suggestions are provided on ways to implement the scoring of the prototype version of the test.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

✣ ✣ ✣

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. The Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. Many projects require the cooperation of other institutions, however, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that data confidentiality will be protected.

Current (1994-95) members of the TOEFL Research Committee are:

Paul Angelis	Southern Illinois University at Carbondale
James Dean Brown	University of Hawaii
Carol Chapelle	Iowa State University
Joan Jamieson	Northern Arizona University
Linda Schinke-Llano	Millikin University
John Upshur (Chair)	Concordia University

Table of Contents

List of Tables

List of Appendices

Background and Rationale

The Test of Spoken English for assessing the speaking ability of nonnative speakers of English has been the subject of research investigation since its introduction more than a decade ago (Bejar, 1985; Boldt, 1992; Boldt and Oltman, 1991; Clark and Swinton, 1979, 1980; Powers, 1983). A number of developments in language assessment and several concurrent research inquiries have had implications for potential improvements in testing format (Boldt, 1992; Cascallar, 1987, 1992; Henning, 1983).

For several years now, the desirability of improving the format and scoring of TSE has been a subject of discussion. This has become more critical with the recent growth in the number of nonnative English-speaking examinees from two major user populations; i.e., prospective university teaching assistants and prospective licensed medical professionals. One objective was to enhance the communicative quality and duration of the elicited speech of TSE by eliminating some parts of the current version of the test and substituting or strengthening other parts, without greatly affecting the current total time of test administration. Another priority was to provide a stronger theoretical rationale for the proposed speaking activities of the test and to streamline the scoring process by eliminating unnecessary procedures. In the revision process, a constant underlying concern has been that no changes be adopted that would compromise the current reliability and validity of the test.

The revision process included discussion among test development, program direction, statistical analysis, and research personnel involved with the current version of the test, as well as meetings with external scholars knowledgeable in the field of oral language assessment. These meetings culminated in a prototype revised TSE that represents an attempt to introduce improvement within the constraints imposed by large-scale operational limitations, such as the need for machine administration of the test, the requirement for rapid scoring and reporting, the desire to maintain low testing fees, and the obligation to preserve fairness and accountability.

Future manuals accompanying TSE will include a more thorough account of all of the changes introduced in the revision process. Here it must suffice to catalogue only the primary features of the revision. The following list includes the major changes incorporated in the prototype revised test (for a copy of this test, see Appendix F):[1]

[1] The materials presented here represent a stage in the development of the revised TSE and other revisions may be made for the final operational version.

1. Three of the original seven sections of the test have been dropped (i.e., the read-aloud section, the sentence-completion section, and the single-picture section). This change has eliminated the sections that were most problematic in terms of content and face validity and has provided a corresponding reduction in time and space needed for procedural instructions.

2. The performance rating scale has moved from a four-step to a five-step scale (i.e., 20, 30, 40, 50, and 60, instead of 0, 1, 2, and 3). This will allow greater reliability and discriminability of individual ratings, implying that fewer items would be needed to achieve the same overall consistency of scoring.

3. The elicitation of examinee language as a proportion of all language used in the test has increased due to the use of more tasks requiring longer responses, the reduction in the variety of instructions needed to accompany sections, and the use of fewer items within sections resulting in fewer pauses and scoring breaks.

4. Although the test is still machine-administered, the communicative orientation of the test content has been strengthened. This was achieved by eliminating certain less communicatively oriented sections of the test and by providing a more systematic communicative rationale for the development of the test and the rating of the speech sample.

5. The score-reporting scale has been changed from the current 0-to-300 scale to a proposed 20-to-60 scale. This change is more consistent with an attempt to conceive of speaking performance within descriptor-accompanied bands, as is more typical of criterion-referenced assessment than the current assessment procedure.

6. Instead of providing subscale scores for fluency, pronunciation, and grammar accuracy that are not averaged into the final comprehensibility score, as was the practice with the current version, it was decided to report scores on a scale of communicative language ability. Available evidence suggested that the subscale scores were highly intercorrelated and, thus, somewhat redundant, and user institutions reported that they made little use of subscale information. This change in scaling procedure entailed a much-needed streamlining of the scoring process. (For copies of the prototype revised scoring rubric and score sheets, see Appendices A and C.)

These and other proposed changes in the format and scoring of the test raised certain questions about the adequacy of the new version for meeting psychometric standards. Such far-reaching changes could not be implemented operationally without supporting research evidence that the modifications would not compromise the reliability, validity, fairness, and effectiveness of the test.

Purpose of the Study

Although the prototype revised Test of Spoken English introduces improvements in the content and format of the test and provides a better match between assessment theory and practice, there remained a need for some empirical evidence that the proposed new version of the test satisfied the same reliability and validity standards as the current version. The purpose of this study was to compare a current version of TSE with the proposed new version with reference to interrater reliability (Tables 4 and 5), distributions of ratings and frequencies of rater discrepancies at all score levels for relevant groups of examinees and raters (Tables 1-3), component task adequacy as reflected in item difficulties (Table 6) and item-total score correlations (Table 7), correspondence of scores over possible scoring steps (Table 8), and other concurrent and construct validity evidence including correlations with oral language proficiency interviews (Table 9) and item-level predictions of holistic overall ratings of speaking ability (Table 10).

In addition to investigating the effects of the proposed changes in test format and scoring, this study was also designed to consider the measured speaking performance construct from the perspective of those most at risk. Of particular concern have been those undergraduate university students and medical patients who have limited exposure to foreign languages and cultures and who may have greater difficulty in understanding nonnative English speech than trained experts would have. It was recognized that just as "beauty is in the eye of the beholder," comprehensibility is in the ear of the comprehender. That is, judgments about the comprehensibility or communicative effectiveness of a person's speech may vary depending on the amount of exposure the listener has had to speakers of other languages. People with wider exposure, including trained raters, might be more sympathetic listeners. For this reason one component of the study was an investigation of the potential judgment differences between naive and trained raters.

Method

Subjects

The sample consisted of nonnative English-speaking examinees from the two major examinee-user populations; viz., prospective university graduate teaching assistants and prospective licensed medical professionals. These two groups represented the primary TSE examinee population. Prospective university teaching assistants included university graduate students who were under consideration for teaching assistantships, and were therefore subject to local review of qualifications such as evidence of requisite English language speaking ability. Prospective licensed medical professionals included foreign medical graduates who were seeking license to practice as physicians, nurses, veterinarians, or pharmacists in the United States. Accordingly, 342 subjects participated by informed consent from these two populations -- 158 prospective licensed medical professionals, and 184 prospective university graduate teaching assistants. The original proposal called for 400 subjects -- 200 from each subsample; however, some attrition was experienced for a variety of reasons (e.g., defective tape recordings accounted for approximately 20 percent of the 58-subject shortfall). And, in general, it was more difficult to enlist the participation of medical professionals since they were not all available at a convenient number of university campuses as was the case with prospective graduate teaching assistants. Subjects participated by informed consent after responding to local public announcements of the opportunity to participate. All subjects were paid to participate.

Subjects represented more than 20 language backgrounds. The most common languages reported by the academic group in order of frequency were Chinese, various Indian, various African, Portuguese, Russian, Korean, and Spanish (80 percent). The predominant languages reported by the medical group were various Indian, Chinese, Russian, Spanish, French. various African, Burmese, and Arabic (73 percent). The most frequent academic disciplines reported by the academic group were engineering, economics, science, and computer science (60 percent).

Academic subjects (i.e., prospective teaching assistants) were drawn from three participating United States universities and one Canadian university. Medical subjects participated from one United States university and from the greater New York City and Philadelphia cosmopolitan areas. Time of residence in the United States or Canada varied greatly from a few weeks to more than 27 years.

Raters

Two types of raters participated in the project: trained and naive. The trained raters consisted of approximately 40 persons who had been certified for rating the current TSE version according to the established training procedures. The formal training for the current TSE version required one full day of instruction and practice rating with feedback on performance, followed by an evening assignment to rate a set of sample tapes at home. Those trainees who failed to complete the group training satisfactorily, or who were found too discrepant in their ratings of the sample tapes, were not certified as raters.

The formal training for the prototype TSE version required one and a half days but did not include the additional rating of a set of sample tapes at home. As with the current TSE version, training involved repeated rating of prerated performance tapes, using the respective rating forms, and receiving feedback on the accuracy of rating performance. Only those persons who had already successfully passed training in the current TSE version were enlisted for prototype TSE training. In addition to meeting training standards, all of the raters were experienced teachers of English as a first or second language, and all had at least a bachelor's degree in a related field. These raters served as needed over the two sessions required to rate the audiotapes of both versions of the test.

The naive raters were purposely chosen to be less sophisticated in their experience with nonnative English speakers. These raters were selected from the student and patient populations because they represented groups most likely to be disadvantaged by nonproficient speakers of English in the test population. These two groups of raters were enlisted through advertisements in a university paper and by telephone contact outside of the university community. The first group consisted of eight university freshmen (four females and four males). One male and one female were chosen from each of four broad academic disciplines (sciences, humanities, business, and social sciences). Criteria for selection also included the requirements that none had traveled abroad, none were language majors or otherwise fluent in a foreign language, none had professional acquaintances or friends who were nonnative speakers of English, none reported any hearing deficiencies. The appendices include a questionnaire that each applicant was required to complete to provide necessary screening information.

The second group of naive raters, potential medical patients, was chosen from within prescribed age ranges from among the community surrounding one of the participating university campuses. These were persons from the general population who had had occasion to seek professional medical assistance at some time in their lives and who were considered likely to seek it again. This group also consisted of eight persons (four females and four males). One male and one female were selected from within each of four age groups (16-25, 26-35, 36-45, and 46 and above). Criteria included the same limits of exposure to nonnative English speech as were required of the freshmen. In addition, in order to further ensure limited exposure to persons from foreign language backgrounds, this group was required not to have had a university education. The same background questionnaire was completed by applicants to this group as was completed by university freshmen applying to serve as raters. All raters were paid equally to rate a total of 10 tapes for communicative effectiveness as indicated in the naive raters' rating form in the appendices. In order not to compromise the naive status of these raters, who have had limited exposure to nonnative English speech, no formal training in rating procedures was provided. However, they were given the rating form shown in the appendices and asked to rate assigned tapes using that form in the same language-laboratory environment. They were also assisted in using the recording equipment. (For a copy of the naive rater rating form, see Appendix E.)

Instrumentation

The instruments employed in the study included a current version of TSE, a prototype revised version of TSE, an oral language proficiency interview (LPI), rating forms for both versions of TSE, and demographic questionnaires for naive raters.

Procedures

For the trained-rater part of the study, in order to minimize sequence or practice effects, a random subsample (consisting of approximately 50 percent of the examinees) participated by responding to the current version of the test before responding to the revised version of the test. The remaining subsample responded to the test versions in the opposite sequence. The current version of the Test of Spoken English was administered under standard conditions using approved materials and equipment. The prototype TSE version was administered the same day as the current version, using the same recording equipment and facilities; however, the test format differed as indicated earlier for the two versions. Examinees were paid to participate. All tests were administered and recorded by high-quality tape recorders in simulated or real laboratory conditions.

Rating of the current TSE tapes proceeded in the established manner, as described in the *TSE Manual*. Using the 0-3 rating scale, a minimum of two trained raters provided independent ratings for all items of the test. In those few cases where total scores exhibited discrepancies exceeding the established criteria, a third and possibly a fourth independent rater were enlisted to resolve the discrepancies. Similarly, rating of the prototype TSE required independent ratings on the part of at least two trained raters for each item of the test. For the prototype TSE ratings, however, the 20-60 rating scale was used, and overall holistic ratings were required in addition to the individual item ratings. Any discrepancy in overall holistic ratings required adjudication by a third and possibly a fourth independent rater, until any two raters agreed on the exact same overall holistic score--which then became the agreed overall holistic score for the test. These agreed overall holistic scores were later compared with item-average scores to ascertain the preferred method of scoring the test.

Analyses

The comparison of the current and prototype versions of TSE called for a variety of statistical analyses. Means and standard deviations were computed for total and partial scores for both the current and prototype versions of the test. For the prototype version, this information was computed with respect to the ratings by trained and by naive raters. Frequency distributions for the assignment of ratings were computed for all tests.

Correlations were computed between the two test versions and for those versions with oral proficiency interview ratings. Interrater reliability estimates were made that consisted of Pearson product-moment correlations between scores assigned by independent raters, which were subsequently adjusted by use of the Spearman-Brown Prophecy Formula (Gulliksen, 1987). Also, item-total correlations were computed for the prototype version, and score-correspondence regression analyses were computed for the two versions.

Results

Current Form TSE Score Data (Table 1)

Means, standard deviations, score frequency distributions, and rater discrepancies are presented in Table 1 for 342 examinees responding to the current version of the Test of Spoken English. Also note from Table 1, the score data are reported separately for the 158 medical professional and the 184 academic professional examinees. Interestingly, the academic subjects scored about one score-step higher on average than the medical subjects. A stronger potential ceiling effect was more evident for academic subjects than for medical subjects, suggesting that the average speaking superiority of the academic subjects may have been even greater than the test was capable of measuring.

Discrepancies were determined as rater overall-average comprehensibility score disagreements in excess of one standard deviation of the total scoring distribution, for comparative research purposes. Because the intention was to compare discrepancy rates between current and revised versions of TSE, and those versions use different scoring scales, it seemed best to base comparative judgments on the proportions of rater discrepancies that exceeded one standard deviation of each respective scoring distribution in order to control for scoring scale differences. Note that, by this criterion, there were 40 rater discrepancies that would require adjudication with the current version of TSE. This represented 11.7 percent of the total examinee sample.

Prototype Revised Form TSE Score Data with Trained Raters (Table 2)

Table 2 reports the same kind of score data as Table 1, except Table 2 presents this information for the prototype revised Test of Spoken English. Note that the sample size here included 12 fewer subjects than were reported in Table 1. This attrition was due primarily to defective tape recordings of the subjects' speaking performances. Although there are differences between the current and prototype versions in scale and measurement construct, some rough comparisons are possible. The overall mean score of 50.27 suggests, by comparison with the mean of 216.17 in Table 1, that a prototype revised TSE score of 50 was roughly equivalent to a current version TSE score of 220 (a more complete table of equivalencies appears in Table 8). Such comparisons mean simply that those examinees obtaining a score of 220 on the current TSE would be expected to obtain a score of approximately 50 on the revised TSE version, but these comparisons are approximate and are not intended to imply that the tests are, or even should be, measuring exactly the same speaking constructs.

The score distribution in Table 2 clearly shows negative skew with some apparent ceiling effect. Since it was known that some skewing can bias parametric correlation coefficients, a variety of score transformations were attempted unsuccessfully to compensate for skew. The creation of a 70 (distinguished score) category, for those few examinees who received ratings of 60 by both raters in all components of the test, was a technique that corrected this skewing and enhanced the interrater reliability and empirical validity of the test. None of the subsequent analyses in the report employed this transformation, but this finding is reported here for purposes of future reference.

Note that in Table 2, unlike Table 1, two methods were employed for identifying rater discrepancies. Use of two methods was dictated by the experimental nature of the prototype TSE test. The first method was proposed by the TSE committee (external scholars who represent the language testing and TSE score user communities), and the second method was added to determine if it would be more accurate and to enable more meaningful comparison with the method used in Table 1. The first method treated all overall score disagreements as discrepancies calling for adjudication. By this method, if one rater assigned an overall holistic rating of 50 and a second rater assigned an overall rating of 60, then a third rater was required and possibly a fourth rater until any two raters were found to agree on one score which then became the reported score used for determining the means and standard deviations in this table. Although this method was straightforward and easy to implement, it had the profound operational-cost disadvantage that 122 (or 37.0 percent) of the scores required adjudication. However, rater training procedures were not as well developed for the prototype as they were in the current TSE version. It is possible that future improvements in training may reduce the proportion of discrepancies requiring adjudication.

The second method for determining rater discrepancies entailed computing rater scores that were averages over the 15 component items of the test, figuring the differences between these average scores for the first two raters, and comparing these differences to the standard deviation of the distribution to establish whether they should be classified as discrepancies requiring adjudication. Disagreements in excess of one standard deviation of the score distribution were classified as discrepancies for purposes of comparison. Because these item-average scores reflected more sensitive scoring increments than the holistic overall ratings, the differences between raters diminished by this method. By this second method, the number of discrepancies diminished to 43 (or 13.0 percent of the distribution). When the second method was employed, there was no statistically significant difference between the current TSE version and the prototype revised TSE version in the frequency of discrepant ratings requiring adjudication. This underscores the benefit of using rater-average scores rather than rater-overall-holistic scores to determine whether scores require adjudication.

Prototype Revised Form TSE Score Data with Naive Raters (Table 3)

One important part of the present study involved the use of untrained, linguistically naive raters to judge the communicative effectiveness and comprehensibility of the examinees responding to the prototype revised TSE version. Table 3 reports the overall holistic score means, standard deviations, and scoring distributions for a group of 16 naive raters who were paid to rate 40 revised-version audiotapes (each rater was responsible for a total of 10 tapes). A more complete description of the raters appears in the Method section of this report. A copy of the simplified rating form appears in Appendix E.

Because the same number of steps and similar descriptors appeared on the rating scales for both the trained-rater and the naive-rater versions of the prototype revised TSE, these ratings were easily translatable from one scale version to the other. Thus, the naive-rater mean rating of 3.71 and standard deviation of 0.85 could be judged equivalent to a mean rating of 47.1 and standard deviation of 8.5 on the trained-rater scale. To compare the ratings of trained and naive raters, however, it was necessary to employ groups of these raters judging the same 40 prototype revised TSE audiotapes, as reported in Table 5. It may be sufficient to indicate at this point that the naive raters appeared to make better use of the lowest scale step than did the trained raters. Recall that in Table 2 trained raters did not use the lowest scale step for any of their overall holistic ratings. Since naive raters did not have access to scale-step behavioral descriptors and did not undergo formal training, this difference between trained and naive raters in application of the lowest step of the rating scale may suggest that information presented in the 20-scale-step behavioral descriptor and/or in the training procedures may have inhibited use of that scale step in actual ratings by trained raters.

Note that there was also a slight tendency for university freshmen to be more critical of the communicative effectiveness of examinees than were prospective patients (i.e., a mean rating of 3.61 versus a mean rating of 3.80). The apparent ceiling effect for patients' ratings suggested that this difference might have been even greater had the scale reflected performance above the scale-step value of 5. Patients, on average, were older and less formally educated than either the naive-freshmen or trained-rater groups, but it is not known if these factors accounted for their seemingly greater tolerance of nonnative-English speech.

Interrater Correlations and Reliabilities for Current and Prototype Revised TSE Versions (Table 4)

Table 4 reports the trained-rater means, standard deviations, intercorrelations, and interrater-reliability estimates for both the current and prototype revised TSE versions. Note that interrater correlations (0.817 and 0.817) and interrater-reliability estimates (0.899 and 0.899) were the same for both versions when final scores for the revised version were determined on the basis of 15-item average ratings. When holistic overall ratings for the revised version formed the basis of interrater correlation and reliability estimation, correlation and reliability estimation dropped to 0.754 and 0.860 respectively. The use of Spearman-Brown formula adjustments to determine reliability estimates in this case implies that final reported scores would need to be the average of rater 1 and rater 2 ratings to achieve this level of reliability. Since overall holistic scores were determined on the basis of agreement by any two raters, in this case reliability was estimated as the Spearman-Brown adjusted mean of rater 1, rater 2 correlations since the actual relationship could not be computed.

Apparently, no interrater-reliability loss resulted from the change to the prototype revised version from the current TSE version. However, this was true only when final scores on the prototype revised version were determined on the basis of 15-item averages, and not when overall holistic scores were used as final reported scores.

Naive-Rater Versus Trained-Rater Means, Standard Deviation, Correlations and Reliability Estimations (Table 5)

Table 5 reports comparable means, standard deviations, interrater correlations, and reliability estimates for naive and trained raters on the prototype revised TSE version and for trained raters on the current TSE version using a subsample of 40 examinees' audio-tape recordings. Most notably, when naive-rater-scale to trained-rater-scale conversions were made of mean scores reported in Table 5, as mentioned in the interpretation of Table 3, the naive-rater overall-holistic-score mean (3.71) became 47.10 by comparison with the trained-rater overall-holistic-score mean of 45.50. This scale conversion was achieved by simple linear transformation of the naive-rater scale (1-5) into the prototype revised scale (20-60), so that scale steps changed as follows: 1 = 20, 2 = 30, 3 = 40, 4 = 50, and 5 = 60. Thus, naive raters tended to be less critical of speaking performance than trained raters (although they were using different scales). This same tendency was exacerbated in the case of 15-item-average overall scores where the naive-rater mean became 50.60 by comparison with the trained rater mean of 45.37. Clearly, naive raters did not tend to judge nonnative-speaking performance as less comprehensible or less communicatively effective than did trained raters.

Naive-rater holistic ratings were less reliable (0.792 versus 0.979) than holistic trained-rater ratings; this is not surprising in view of the lack of formal training on the part of naive raters. (If the comparison is made based on two raters only in each situation then the comparable coefficients would be .488 and .958, indicating that judgments of naive raters were predictably much less reliable than those of trained raters.) And, although reliability estimates for naive-rater 15-item average ratings improved to 0.981, this high figure was due in part to the fact that these estimates were based on the composite ratings of four naive raters instead of two raters in the case of trained raters. The trained-rater reliability estimates for the revised TSE version were similar.

TSE Prototype Revised-Form Component-Score Means and Standard Deviations (Table 6)

Table 6 reports the means and standard deviations of ratings for each item and section of the TSE prototype revised version as rated by trained raters. Note that, with the possible exception of the final section, section scores reflected a planned tendency for the test sections to become progressively more difficult. The most difficult item of the test was item 1 of part 4 (the mean score equalled 48.470), and the easiest item of the test was item 2 of part 1 (the mean score equalled 52.924). The total mean at the bottom of the table was computed as the mean of section scores rather than as the mean of item scores. In this way this score gave equal weight to each section score. The part 5 overall score mean was the same as that for part 5, item 1, because part 5 had only one item.

TSE Prototype Revised Component-Total Correlations (Table 7)

Table 7 reports the Pearson product-moment correlations of section and item scores with two different total scores for the TSE revised version as rated by trained raters. The first total score (labeled FOAS in the table) reflected the final overall agreed holistic score, and the second total score (labeled FRMS in the table) reflected the 2-rater-by-5-part mean score which was shown earlier to be more reliable than the holistic score.

This table provides important indications about component-score discrimination and comparative internal consistency and construct validity of test sections and items. Although all correlations were positive, highly significant, and showed little variation in magnitude, the weakest sections of the test from the perspective of maximizing internal consistency, discriminability, and construct validity were part 5 (coefficients = 0.849, 0.938) and part 1 (coefficients = 0.859, 0.941). In the case of part 5, the lower coefficients were clearly due to the comparative paucity of items in that section rather than to problems with the nature of the task, since the single item of part 5 showed the highest item-total correlation. The weakest items of the test were part 2, item 2 and part 1, item 1 (coefficients = 0.758, 0.836, and 0.782, 0.866, respectively). By the same criteria, the strongest parts of the test were parts 3 and 4 and item 1 of part 3 and item 1 of part 5.

Correspondence of Scores for Current and Prototype Revised Versions (Table 8)

Table 8 reports the correspondence of scores on the current and revised versions of the test as rated by trained raters. Note that the values in parentheses at the bottom of the table represent the linear-regression estimates of current TSE scores corresponding to agreed holistic scores on the revised version. Consider, however, that there is a high degree of standardized error associated with these estimates (22.4), so it would be appropriate to assume that actual corresponding scores would fall within a broad range as occurred in Table 8.

Oral Proficiency Interview Data for Current and Prototype Revised Versions (Table 9)

To provide further concurrent validity evidence regarding the TSE scores, a subsample of 39 examinees were administered a formal oral proficiency interview (LPI), recognized by the American Council on the Teaching of Foreign Languages, the Foreign Service Institute, and the Interagency Language Roundtable. Scores on this interview ranged from 1 to 4, with a mean of 2.26 and a standard deviation of 0.72. Plus ratings between steps on the scale were coded as 0.8 for purposes of computation. Three subjects were interviewed for whom TSE scores were not available, so that the correlations in Table 9 are based on the remaining 36 subjects. Note that the TSE-score correlation with the oral proficiency interview was higher for the revised TSE version than for the current TSE version (0.819 versus 0.748). This comparison suggests that the revised TSE version did not show any reduction in available empirical validity in spite of the fact that the revised version represented a reduction in the total number of sections and items from the current TSE version.

Stepwise Multiple Regression Analysis of Prototype TSE Items Used to Predict Holistic Overall Ratings of Speaking Ability (Table 10)

The stepwise multiple regression results reported in Table 10 suggest that, from the standpoint of optimal prediction of holistic ratings of overall speaking performance, many items of the prototype TSE could be considered redundant. No significant new variance was accounted for in the prediction by using items beyond the six items that first entered the regression equation. This evidence, however, does not suggest that the remaining items should be dropped from the test. In fact, such redundancy contributes to reliable measurement.

Discussion and Conclusions

This study attempted to compare a current version of the Test of Spoken English with a proposed new version with regard to estimates of interrater reliability (Tables 4 and 5), distributions of ratings and frequencies of rater discrepancies at all score levels for relevant groups of examinees and raters (Tables 1-3), component task adequacy as reflected in item difficulties (Table 6) and item-total score correlations (Table 7), correspondence of scores over possible scoring steps (Table 8), and other concurrent and construct validity evidence, including correlations with oral language proficiency interviews (Table 9) and item-level predictions of holistic overall ratings of speaking ability (Table 10).

Results show that the prototype revised TSE did not fall behind the current version of the test in any available measure of reliability or validity. There is also some limited correlational evidence to suggest that the revised version may be a better reflection of the kind of oral proficiency that is measured by the oral proficiency interview of the Interagency Language Roundtable and the American Council on the Teaching of Foreign Languages. These outcomes are encouraging because the revised version of the test was designed to be more communicative and included fewer overall sections, items, and subscales than the current version (although the actual amount of examinee language elicited in the revised version, within the same amount of examination time, is greater than in the current version).

Scoring

It was also noted that there was no significant difference between the revised version of TSE and the current TSE version in the frequency of rater discrepancies (11-13 percent of the total scores exhibited differences exceeding one standard deviation of the distribution), provided that item-average scores formed the basis of comparison for both versions. When final agreed holistic scores were used as the basis of total score determination for the revised version of TSE, however, the rater discrepancies for which rater adjudication were required rose to 37 percent of the total.

It was clearly more advantageous to use item or section average scores than overall holistic scores for the purpose of determining which ratings required additional-rater adjudication and for the purpose of maintaining the same level of interrater reliability that is available with the current version of the test. It is likely that average scores would also be more accurate for purposes of formal equating of subsequent versions of the test.

Construct Validity

In spite of the content and format changes represented in the revised version of the Test of Spoken English, the speaking ability construct measured by the revised version is highly related to the current version measured construct. The correlation between scores for the two versions was 0.831 (Table 8), and when this correlation was disattenuated by removing the effects of unreliability, the correlation approached unity (.983 - 1.059). The correlations shown for the current and revised versions of TSE with an oral language proficiency interview (Table 9) differed only slightly (0.819 for the revised version versus 0.748 for the current version). Although this difference seemed to indicate higher concurrent validity for the revised version, because the sample of interviewees was small, the difference was not significant. Correlations between items and total scores were universally positive and significant for the revised version of the test. This also demonstrates the uniformity of the measurement construct.

The combination of all these reliability, validity, and scoring-distributional outcomes, when considered alongside the practical and theoretical improvements in the content and format of the test, strongly suggests that the prototype revised version of TSE should be adopted in replacement of the current version of the test.

Theoretical Improvements

Although the psychometric information reported in this study leads to the conclusion that this prototype revised TSE is very similar to the current version of TSE in terms of empirical estimates of reliability and construct validity, this in no way detracts from the theoretical improvements introduced by the revision effort.

The prototype revised TSE has become more communicative in focus through the deletion of read-aloud and sentence-completion type tasks and through the more appropriate selection of topical content and performance criteria. The revisions discussed here introduce greater content validity and theory-based rationale in the assessment. Similarly, the revisions increase the proportion of relevant examinee speech elicited within the same examination time, so that it can be maintained that the revisions enhance face validity of the test as well.

References

Bejar, I. (1985). A preliminary study of raters for the Test of Spoken English. TOEFL Research Report No. 18. Princeton, New Jersey: Educational Testing Service.

Boldt, R. F. (1992). Reliability of the Test of Spoken English revisited. TOEFL Research Report No. 40. Princeton, New Jersey: Educational Testing Service.

Boldt, R. F., and Oltman, P. (1993). Multimethod construct validation of the Test of Spoken English. TOEFL Research Report No. 46. Princeton, New Jersey: Educational Testing Service.

Sarwark, S. M., Smith, J., MacCallum, R., and Cascallar, E. (in press) A study of characteristics of the SPEAK test. TOEFL Research Report. Princeton, New Jersey: Educational Testing Service.

Cascallar, E. (1992). Automated evaluation of English pronunciation. Unpublished TOEFL research proposal. Princeton, New Jersey: Educational Testing Service.

Clark, J. L. D., and Swinton, S. S. (1979). An exploration of speaking proficiency measures in the TOEFL context. TOEFL Research Report No. 4. Princeton, New Jersey: Educational Testing Service.

Clark, J. L. D., and Swinton, S. S. (1980). The Test of Spoken English as a measure of communicative ability in English-medium instructional settings. TOEFL Research Report No. 7. Princeton, New Jersey: Educational Testing Service.

Gulliksen, H. (1987). Theory of mental tests. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. Language Learning, 33 (3), 315-332.

Powers, D. E., and Stansfield, C. (1983). The Test of Spoken English as a measure of communicative ability in the health professions: validation and standard setting. TOEFL Research Report No. 13. Princeton, New Jersey: Educational Testing Service

TABLES

Table 1

Current TSE Means, Standard Deviations, Score Frequency Distributions, and Rater Discrepancies* for Medical and Academic Examinee Groups

				Score Distribution										
Group	N	Mean	SD	40	50	60	70	80	90	100	110	120	130	140
Medical	158	210.19	50.77	0	1	0	1	0	1	1	2	2	1	5
Academic	184	221.30	56.71	1	0	0	1	1	1	2	3	4	4	7
Total	342	216.17	54.25	1	1	0	2	1	1	3	5	6	5	12
Discrepancies	40	(11.7%)		0	0	0	1	1	0	0	1	0	0	2

	Score Distribution (cont.)															
Group	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290	300
Medical	8	5	6	11	16	15	14	14	7	8	10	3	6	6	6	9
Academic	2	2	3	7	15	16	11	13	14	10	9	9	5	12	20	12
Total	10	7	9	18	31	31	25	27	21	18	19	12	11	18	26	21
Discrepancies	1	1	1	3	5	1	2	2	6	4	7	1	1	0	0	0

* Discrepancies were determined as rater overall average comprehensibility score disagreements in excess of one standard deviation of the total scoring distribution (i.e., 54.25), for comparative research purposes.

Table 2

Prototype Revised TSE Means, Standard Deviations, Score Frequency Distributions, and Rater Discrepancies* for Medical and Academic Examinee Groups

				Score Distribution				
Group	N	Mean	SD	20	30	40	50	60
Medical	147	49.93	8.15	0	4	37	62	44
Academic	183	50.55	9.06	0	11	37	66	69
Total	330	50.27	8.66	0	15	74	128	113
Discrepancies								
Method A	122	(37.0%)		0	6	35	61	20
Method B	43	(13.0%)		0	2	16	21	4

* Discrepancies were determined by two different methods for comparative research purposes. By method A, any overall score difference between rater 1 and rater 2 was considered a discrepancy and called for additional independent rating. By method B, discrepancies were determined as rater-average-component-score disagreements in excess of one standard deviation of the total scoring distribution (i.e., 8.66). Thus, method B is comparable to the method used in Table 1 for the current Test of Spoken English. Assignment of discrepancies to discrete score categories by this method was done on the basis of rounding of average scores to the nearest scale step.

Table 3

Overall Score Means, Standard Deviations, and Communicative Effectiveness Scores Assigned by Naive Freshmen (4 Females and 4 Males) and Naive Patients (4 Females and 4 Males) for 40 Prototype Revised TSE Tapes; 4 Ratings per Tape

					Score Distribution				
Rater Group	Rater N	Rating N	Mean	SD	1	2	3	4	5
Freshmen	8	80	3.61	0.92	1	10	27	23	19
Patients	8	80	3.80	0.99	3	8	19	22	28
Total	16	160	3.71	0.85	4	18	46	45	47

Table 4

Rater Means, Standard Deviations, Pearson Product-Moment Correlations, and Interrater Reliability Estimates for Current TSE and Prototype Revised TSE Versions Rated by Trained Raters

TSE Version	Rating Method	N	Possible Range	Rater 1 Mean	SD	Rater 2 Mean	SD	$r_{r1,r2}$	r_{tt}
Current		342	0 - 300	214.92	56.26	217.34	56.60	0.817	0.899
Revised		330	20 - 60						
	Overall Score			50.21	9.44	50.42	8.88	0.754	0.860*
	15-Item Average Score			49.98	8.41	50.00	8.01	0.817	0.899

* The Spearman-Brown adjustment of r_1, r_2 for two raters is reported here. This number provides an estimate of the relationship that would be obtained if the average of the two ratings were used as the final score. (This should be a conservative estimate of the reliability of the score based on the adjudication procedure; the latter relationship cannot be computed.)

Table 5

Naive Rater Study Comparative Means, Standard Deviations, Pearson Product-Moment Correlations, and Interrater Reliability Estimates for Current TSE and Prototype Revised TSE Versions Using the Same 40 Randomly Chosen Examinees

TSE Version	Rating Method	$\underline{N}$	Possible Range	Male Freshman Raters Mean (SD)	Female Freshman Raters Mean (SD)	Male Patient Raters Mean (SD)	Female Patient Raters Mean (SD)	$\underline{r}_{r1,r2}$	$\underline{r}_{tt}$
Revised									
	Naive Rater	40	1 - 5#						
	Overall Score			3.62 (1.17)	3.60 (0.87)	3.83 (1.20)	3.78 (1.10)	0.488*	0.792
	15-Item Average Score			3.73 (0.85)	3.77 (0.77)	4.43 (0.99)	4.31 (1.03)	0.927*	0.981
				First Rater Mean (SD)		Second Rater Mean (SD)			
	Trained Rater	40	20 - 60	45.50 (10.17)		45.50 (11.08)		0.958	0.979**
	Overall Score								
	15-Item Average Score			45.80 (10.11)		44.93 (10.52)		0.946	0.972
Current									
	Trained Rater	40	0 - 300						
	6-Part Average Score			198.93	75.28	194.48	75.21	0.912	0.954

* Pooled single-rater reliability, based on four raters.

** Spearman-Brown adjustment of $\underline{r}_{r1',r2}$ for two raters is reported here.

To convert these (1-5 scale) mean and standard deviation values to the (20-60 scale) values employed by the trained raters, one may simply multiply means by 10 and add 10 (so that 3.62 becomes 46.2) and multiply standard deviations by 10 (so that 1.17 becomes 11.70).

Table 6

Prototype Revised TSE Item and Section Score Means and Standard Deviations Averaged Over Rater 1 and Rater 2 Scores (N = 330)

Section		Mean	SD
Part 1	Overall	51.051	7.544
	Item 1	50.379	8.375
	Item 2	52.924	7.325
	Item 3	49.848	8.468
Part 2	Overall	50.222	7.970
	Item 1	50.273	8.256
	Item 2	50.955	8.713
	Item 3	49.439	9.196
Part 3	Overall	49.852	8.281
	Item 1	49.606	8.567
	Item 2	50.576	8.637
	Item 3	48.818	9.182
	Item 4	50.409	8.633
Part 4	Overall	49.038	8.620
	Item 1	48.470	9.715
	Item 2	49.530	9.143
	Item 3	49.152	9.309
	Item 4	49.000	9.253
Part 5	Overall	50.500	8.619
	Item 1	50.500	8.619
Total	Mean of Part Scores	50.133	7.803

Table 7

Prototype Revised TSE Item and Section Score Correlations with Final Overall Agreed Score (FOAS) and Final 2-Rater-by-5-Part Mean Score (FRMS) (N = 330)

Section		FOAS Correlation	FRMS Correlation
Part 1	Overall	0.859	0.941
	Item 1	0.782	0.866
	Item 2	0.804	0.887
	Item 3	0.825	0.892
Part 2	Overall	0.864	0.947
	Item 1	0.798	0.873
	Item 2	0.758	0.836
	Item 3	0.811	0.887
Part 3	Overall	0.897	0.970
	Item 1	0.872	0.928
	Item 2	0.839	0.909
	Item 3	0.849	0.924
	Item 4	0.836	0.908
Part 4	Overall	0.893	0.957
	Item 1	0.807	0.871
	Item 2	0.840	0.896
	Item 3	0.804	0.859
	Item 4	0.840	0.900
Part 5	Overall	0.849	0.938
	Item 1	0.849	0.938
Total (FRMS)		0.918	1.000

Table 8

Correspondence of Current TSE Scores and Prototype Revised TSE Scores (N = 321)

Current Version TSE Score				
300				20
290				25
280			2	15
270			2	11
260			7	5
250			7	12
240			11	6
230			14	6
220			16	5
210		4	18	1
200		8	16	6
190	1	15	13	
180	1	7	10	
170		3	5	
160		5		
150		7	2	
140	2	8	1	
130	1	4		
120		5		
110	3	2		
100	3			
90		2		
80	1			
70	1			
60				
50	1			
40	1			
30				
20	30	40	50	60

Prototype Revised TSE Score

Pearson Correlation between current and revised version scores = 0.831

Regression equation for the estimation of current version TSE scores from the knowledge of overall agreed holistic scores of the revised TSE version:

Current Version TSE Estimated Score = -42.818 + 5.178 (Revised Version TSE Score)

Revised TSE		Current TSE	95% Upper & Lower Bound	Fitted SE	Predicted SE
20	=	60.75	48.74 - 72.75	6.10	30.77
30	=	112.53	104.12 - 120.93	4.27	30.45
40	=	164.31	159.19 - 169.43	2.60	30.27
50	=	216.09	212.78 - 219.41	1.68	30.20
60	=	267.87	262.89 - 272.85	2.53	30.26

Table 9

Oral Proficiency Interview (LPI) Means, Standard Deviations, and Pearson Product-Moment Correlations with Current and Prototype Revised TSE Scores

Test	N	Mean	SD	LPI Interview	TSE Current	TSE Revised
LPI Interview	39	2.26	0.72	1.000		
TSE Current	36	213.89	45.06	0.748	1.000	
TSE Revised	36	48.89	8.20	0.819	0.847	1.000

Table 10

Stepwise Multiple Regression Analysis of Prototype TSE Items Used to Predict Holistic Overall Ratings of Speaking Ability

(N=330) Multiple R = .921; Squared Multiple R = .848; Adjusted R Squared =.844

Step	Item	Cumulative R	Regression Coefficient	Standardized Coefficient	Error	t
1	C1	.872	.229	.226	.054	4.22 **
2	D4	.898	.106	.113	.047	2.25 *
3	A3	.908	.147	.144	.045	3.27 **
4	D2	.914	.126	.133	.045	2.79 **
5	D3	.917	.109	.117	.038	2.87 **
6	E1	.919	.123	.123	.051	2.41 *
7	B1	.920	.077	.073	.044	1.76
8	C2	.921	.082	.082	.050	1.64

* $p < 0.05$ (2 tail)

** $p < 0.01$ (2 tail)

All other coefficients were not significant.

APPENDICES

PROTOTYPE REVISED SCORING RUBRIC

Holistic Band Scores

60 COMMUNICATION ALMOST ALWAYS EFFECTIVE

- Almost always comprehensible to all listeners.
- Pronunciation, grammar, fluency, and vocabulary almost always accurate.
- Coherent with effective use of cohesive devices.
- Completely appropriate response to audience/situation.
- Successfully addresses communicative tasks.

50 COMMUNICATION GENERALLY EFFECTIVE

- Generally comprehensible to all listeners.
- Generally accurate pronunciation, grammar, fluency, and vocabulary.
- Coherent with some effective use of cohesive devices.
- Generally appropriate response to audience/situation.
- Addresses communicative tasks and/or successfully uses compensation strategies.

40 COMMUNICATION SOMEWHAT EFFECTIVE

- Somewhat comprehensible with some effort to all listeners.
- Somewhat accurate pronunciation, grammar, fluency, and vocabulary.
- Coherent with some use of cohesive devices.
- Somewhat appropriate response to audience/situation.
- Partially addresses communicative tasks; some use of compensation strategies.

30 COMMUNICATION GENERALLY NOT EFFECTIVE

- Generally incomprehensible to most listeners even with effort.
- Generally inaccurate pronunciation, grammar, fluency, and vocabulary.
- Incoherent with little to no use of cohesive devices.
- Generally inappropriate response to audience/situation.
- Unable to address most communicative tasks; little to no use of compensation strategies.

20 NO EFFECTIVE COMMUNICATION

- Either says nothing or is completely incomprehensible to all listeners.
- No accuracy in pronunciation, grammar, fluency, and vocabulary.
- Incoherent with no use of cohesive devices.
- Completely inappropriate response to audience/situation.
- Unable to address any communicative tasks; no use of compensation strategies.

DEFINITION OF TERMS USED IN PROTOTYPE RLVISED SCORING RUBRICS

Effective communication: The degree to which an intended message is successfully conveyed to the listener.

Comprehensibility: The degree to which a listener is able to correctly identify the intended meaning of the speaker. An utterance fails to be comprehensible if the listener cannot identify its meaning within the given context.

Accuracy: The degree to which pronunciation, grammar, fluency, and use of vocabulary approaches that of a native speaker.

Pronunciation: The skill with which a speaker uses phonemic and phonetic contrasts along with patterns of stress, intonation, and rhythm to produce intelligible speech.

Grammar: The skill with which a speaker controls simple and complex morphological and syntactic structures in the production of intelligible speech.

Fluency: The skill with which a speaker controls pausing and smoothness of flow in the production of intelligible speech.

Vocabulary: The skill with which a speaker selects and employs words and expressions that are appropriate for the intended message.

Intelligibility: The degree to which a listener is able to identify the linguistic structure or form of the speaker's language. An utterance fails to be intelligible if the listener cannot identify and repeat the words of the utterance.

Coherence: The degree to which the various components of the speaker's utterances are explicitly connected to each other and to the listener's knowledge of the world.

Use of cohesive devices: The degree to which a speaker uses reference, substitution, conjunction, and vocabulary replacement to tie utterances together.

Response to audience/ or the situation: The sensitivity of the speaker to the sociocultural context of the listener situation in which the utterance occurs. Such sensitivity is demonstrated by the speaker's choice of exponents to directly or indirectly accomplish a sociolinguistic function, choice of vocabulary, the use of idiomatic expressions register, degree of politeness, relative complexity of the utterance, speed, volume, and tone of voice.

Addressing the communicative task: The degree to which a speaker successfully conveys the message requested by the directions given for the task.

Uses compensation strategies: The degree to which the speaker uses such strategies as paraphrase, examples, collocation, synonyms, redundancy, topic identification, comparison/contrast, demonstration, writing, spelling, and avoidance to convey information that would not necessarily be understood without use of the strategy.

TSE SCORE SHEET

Each diagnostic area rated on a 0 – 3 scale

Use a No. 2 pencil only. Completely fill in the circle that corresponds to your intended response Erase any errors completely

NOT RATABLE circle is to be completed when the examinee is not to be rated for that item

EXAMINEE I.D.							TYPE CODE
							Ⓐ
⓪	⓪	⓪	⓪	⓪	⓪	⓪	
①	①	①	①	①	①	①	Ⓟ
②	②	②	②	②	②	②	
③	③	③	③	③	③	③	
④	④	④	④	④	④	④	
⑤	⑤	⑤	⑤	⑤	⑤	⑤	
⑥	⑥	⑥	⑥	⑥	⑥	⑥	
⑦	⑦	⑦	⑦	⑦	⑦	⑦	
⑧	⑧	⑧	⑧	⑧	⑧	⑧	
⑨	⑨	⑨	⑨	⑨	⑨	⑨	

RATER I.D.		
⓪	⓪	⓪
①	①	①
②	②	②
③	③	③
④	④	④
⑤	⑤	⑤
⑥	⑥	⑥
⑦	⑦	⑦
⑧	⑧	⑧
⑨	⑨	⑨

CENTER No.					
Ⓐ	Ⓛ	Ⓦ	⓪	⓪	⓪
Ⓑ	Ⓜ	Ⓧ	①	①	①
Ⓒ	Ⓝ	Ⓨ	②	②	②
Ⓓ	Ⓞ	Ⓩ	③	③	③
Ⓔ	Ⓟ		④	④	④
Ⓕ	Ⓠ		⑤	⑤	⑤
Ⓖ	Ⓡ		⑥	⑥	⑥
Ⓗ	Ⓢ		⑦	⑦	⑦
Ⓘ	Ⓣ		⑧	⑧	⑧
Ⓙ	Ⓤ		⑨	⑨	⑨
Ⓚ	Ⓥ				

SCORE SHEET No.
①
②
③
④

TEST DATE		
MONTH	YEAR	
○ JUL		⓪
○ AUG		①
○ SEP		②
○ OCT		③
○ NOV		④
○ DEC		⑤
○ JAN		⑥
○ FEB		⑦
○ MAR		⑧
○ APR	⑨	⑨
○ MAY		
○ JUN		

SECTION II READING ALOUD

Pronunciation	Fluency	Comprehensibility
⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable

SECTION III SENTENCE COMPLETION

Grammar	Comprehensibility	Grammar	Comprehensibility
1. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	6. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
2. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	7. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
3. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	8. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
4. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	9. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
5. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	10. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable

SECTION IV PICTURE SEQUENCE

Pronunciation	Fluency	Comprehensibility
⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable

SECTION V SINGLE PICTURE

Pronunciation	Grammar	Fluency	Comprehensibility
1. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
2. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
3. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
4. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable

SECTION VI [illegible]

Pronunciation	Fluency	Comprehensibility
1. ⓪ ① ② ③ ○ Not [illegible]	[illegible]	[illegible]
2. ⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable
[illegible]	[illegible]	[illegible]

[illegible]

Pronunciation	Fluency	Comprehensibility
⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable	⓪ ① ② ③ ○ Not Ratable

QCODE 1044 1 N.208807 70808 • 04808 • TF122P20X LB82107

3

PROTOTYPE REVISED SCORE SHEET

OVERALL SCORE

EXAMINEE I.D.

RATER I.D.

(CIRCLE THE CORRECT No.)

SCORE SHEET No. 1 2 3 4

NR" should be entered for responses that can not be scored for technical reasons.
(i.e. poor sound fidelity, faulty recording, background noise interference)

SECTION 1	
1	
2	
3	
SECTION 2	
1	
2	
3	
SECTION 3	
1	
2	
3	
4	
SECTION 4	
1	
2	
3	
4	
SECTION 5	
1	

APPENDIX D

RATER QUESTIONNAIRE

Name: ____ ______________

Address: ____

Telephone: _________________________ Birth date: ______________

Kind of work or academic major field: _____________________________________

Indicate your highest level of education:

elementary school: ________

junior high or high school: ________

some college study: ________ (number of semesters: ________)

undergraduate degree (BA, BS, etc.): ________

graduate degree (MA, PhD, etc.): ________

Indicate the total number of years you have studied any foreign language(s):

none: ________

up to one year: ________

one to two years: ________

more than two years: ________

Indicate the time you have spent in travel to countries where English is not the official language:

never traveled abroad: ________

less than one month abroad: ________

less than three months abroad: ________

three to six months abroad: ________

more than six months abroad: ________

Indicate your level of contact with non-native speakers of English:

never any personal or work acquaintances: ________

a few casual acquaintances: ________

regular frequent continuing relationships: ________

Do you have any known hearing disability: Yes ________ No ________

APPENDIX E

RATING FORM FOR SPEAKING COMPETENCE

TAPE NUMBER ____________________________

RATER NAME ____________________________

OVERALL SCORE
1 2 3 4 5

DIRECTIONS: For each part of the recorded speaking session indicated below, show your judgment of how understandable the speaker's language seemed to you. Imagine the speaker was to serve as your teacher or as your physician. Circle the appropriate number in the row opposite each part.

SECTION 1 **PART**	**No Effective Communication**	**Communication Generally Not Effective**	**Communication Somewhat Effective**	**Communication Generally Effective**	**Communication Almost Always Effective**
(1)	1	2	3	4	5
(2)	1	2	3	4	5
(3)	1	2	3	4	5
SECTION 2 **PART**					
(1)	1	2	3	4	5
(2)	1	2	3	4	5
(3)	1	2	3	4	5
SECTION 3 **PART**					
(1)	1	2	3	4	5
(2)	1	2	3	4	5
(3)	1	2	3	4	5
SECTION 4 **PART**					
(1)	1	2	3	4	5
(2)	1	2	3	4	5
(3)	1	2	3	4	5
SECTION 5 **PART**					
(1)	1	2	3	4	5

APPENDIX F

ID #: ______________________________

Form 7

ENGLISH ORAL PROFICIENCY PRETEST

DO NOT OPEN TEST BOOK UNTIL YOU ARE TOLD TO DO SO.

THIS TEST BOOK MUST NOT BE TAKEN FROM THE ROOM.

ETS CONFIDENTIAL

GENERAL DIRECTIONS

In the test, you will be able to demonstrate how well you speak English. The test has five different sections and will last approximately twenty minutes. In each section you will be asked questions by an interviewer. The questions are printed in the test book and the time you will have to answer each one is written in parentheses after the question. You are encouraged to answer the questions as completely as possible. While most of the questions on the test may not appear to be directly related to your academic or professional field, each question is designed to tell the raters about your oral language ability. The raters will evaluate how well you communicate in English.

As you speak, your voice will be recorded. Your score for the test will be based on your speech sample. Be sure to speak loudly enough for the machine to record clearly what you say. Now, please start your tape recorder so that it will record what you say. Your tape recorder should now be running and recording. Do not stop your tape recorder at any time during the test. If you have a problem with the tape recorder, notify the test supervisor immediately.

Now, please go on to Section One.

SECTION ONE: DIRECTIONS

In this section, I am going to ask you some questions about yourself. After each question, you will have a short time to answer. The first three questions are for practice and will not be scored, but it is important that you answer them.

What is the ID number on the cover of your test booklet? (10 seconds)

How long have you been studying English? (10 seconds)

Why are you taking the English Oral Proficiency Pretest? (10 seconds)

Now the test will begin with question #1. Be sure to speak clearly and say as much as you can in responding to each question.

1. Imagine that I am coming to visit next Tuesday. You are going to meet me at the airport. Could you please tell me what you look like and what you will be wearing so that I can recognize you? (45 seconds)

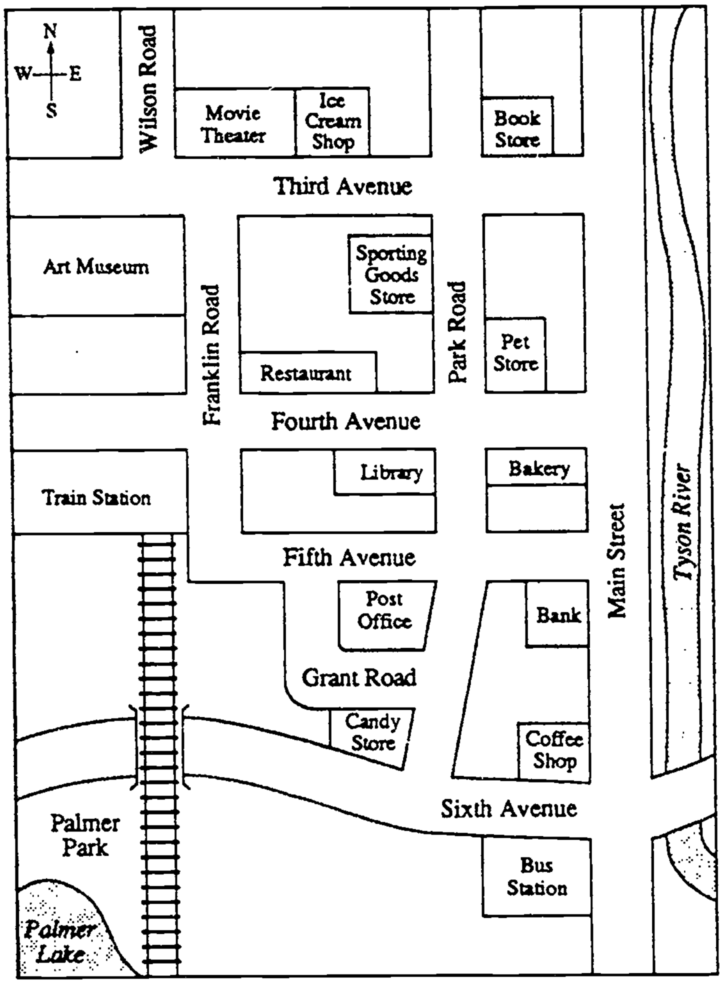
2. I would like to know more about your daily routine. What do you usually do on Tuesdays? (45 seconds)

3. When I come to visit, I would like to buy some gifts to take home with me. Please tell me what you think I should buy and why you made that suggestion. (45 seconds)

This is the end of Section One. Now, please go on to Section Two.

SECTION TWO: DIRECTIONS

Imagine that this is a map of a neighboring town which you have suggested that I visit. You will have 30 seconds to study the map. Then I will ask you some questions about the map.



Now please answer these questions about the map. Be sure to say as much as you can in responding to each question.

1. Choose one place on the map that you think I shou:ld visit and give me some reasons why you recommend this place. (30 seconds)

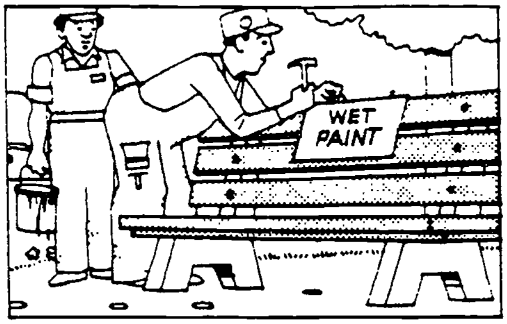
2. I would like to see a movie. Could you please give me directions from the bus station to the movie theater? (30 seconds)

3. Imagine that the movie theater is showing one of your favorite movies. Please tell me about the movie and why you like it. (60 seconds)

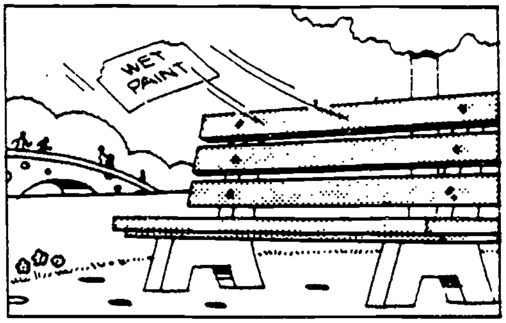
This is the end of Section Two. Now, please go on to Section Three.

SECTION THREE: DIRECTIONS

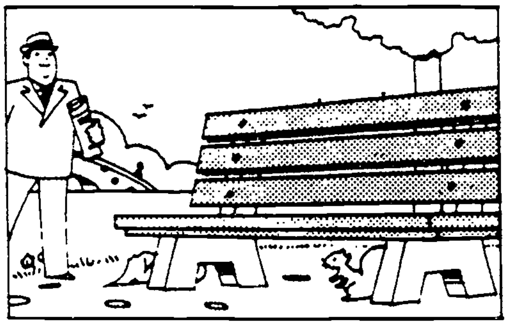
In this section, you will see six pictures. I would like you to tell me the story that the pictures show, starting with picture number one and going through picture number six. Please take one minute to look at the pictures and think about the story. Do not begin the story until I tell you to do so.



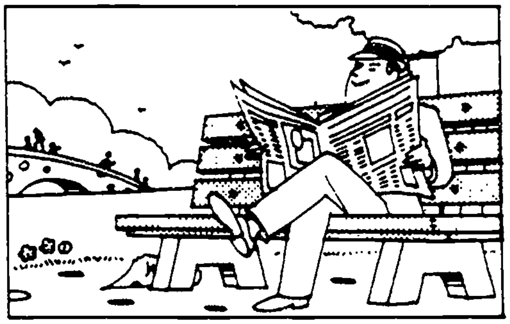
1



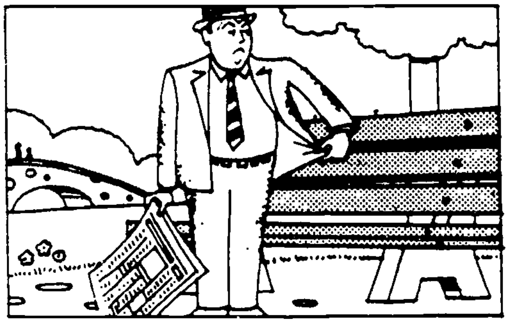
2



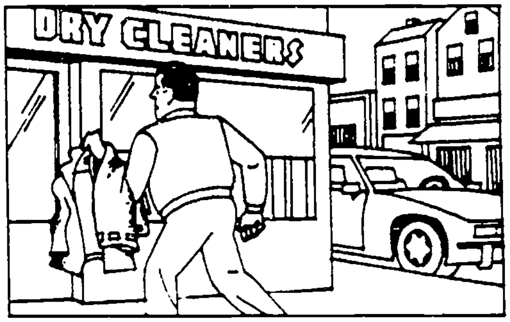
3



4



5



6

1. Tell the story that the pictures show. (60 seconds)
2. What would you do if this happened to you? (30 seconds)
3. What could the painters have done to prevent this? (30 seconds)
4. Imagine that this happens to you. After you have taken the suit to the dry cleaners, you find out that you need to wear the suit the next morning. The dry cleaning service usually takes two days. Call the dry cleaners and try to persuade them to have the suit ready late: today. (45 seconds)

This is the end of Section Three. Now, please go on to Section Four.

SECTION FOUR: DIRECTIONS

In this section, I would like to hear your ideas about a variety of topics. Be sure to say as much as you can in responding to each question. After I ask each question, you may take a few seconds to prepare your answer, and then begin speaking when you are ready.

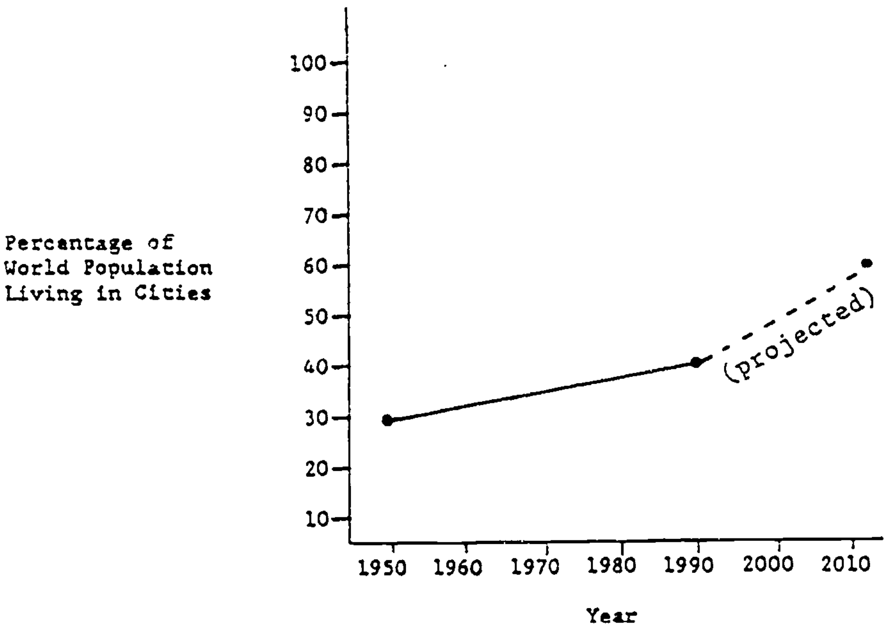
Topic 1. Tell me about a recent news event you have read or heard about. Who was affected by it? Why is it important? (60 seconds)

Topic 2. Many people enjoy visiting zoos and seeing the animals. Other people believe that animals should not be taken from their normal surroundings and put into zoos. I would like to know what you think about this issue and why. (60 seconds)

Topic 3. I am not familiar with your field of study. Select a term used frequently in your field and define it for me. (60 seconds)

Topic 4. The graph below presents the actual and projected percentage of the world population living in cities from 1950 to 2010. First, briefly describe to me the information given in the graph. Then, discuss what this might mean for the future. (75 seconds)

PERCENTAGE OF WORLD POPULATION
LIVING IN CITIES 1950-2010



This is the end of Section Four. Now, please go on to Section Five.

SECTION FIVE: DIRECTIONS

In this section of the test there is some information about a trip to Washington, D.C., that has been organized for the members of the Forest City Historical Society. Imagine that you are the president of this organization. At the last meeting you gave out a schedule for the trip, but there have been some changes. You must remind the members about the details of the trip and tell them about the changes indicated on the schedule. In your presentation do not just read the information printed, but present it as if you were talking to a group of people. Please take one minute to plan your presentation. Do not start speaking until I tell you to do so.

FOREST CITY HISTORICAL SOCIETY
TRIP TO WASHINGTON, D.C.

Date:	Saturday, April ~~5~~ 12
Transportation:	Chartered Bus
Depart:	~~8:30~~ 8:00 a.m. -- Community Center parking lot
Itinerary:	10:30 a.m. -- Guided Tour of White House
	12:30 p.m. -- Lunch* -- Rock Creek Park
	3:00 p.m. -- National Museum of History and Technology (lecture -- 4:00 p.m.)
	6:30 p.m. -- Dinner, ~~Embassy~~ Restaurant Clinton Inn ~~Georgetown~~
Return:	~~10:00~~ 9:30 p.m. (approximately)
Cost:	~~$20.00~~ $25.00 (excluding admissions and dinner)

* Bring your own

(90 seconds)

This is the end of the English Oral Proficiency Pretest. Than.. you for your cooperation.

Cover Printed on Recycled Paper

57906-09061 • Y25M 5 • 275516 • Printed in U S A