DOCUMENT RESUME

ED 384 674 TM 023 995

AUTHOR	Boldt, R. F.
TITLE	Reliability of the Test of Spoken English Revisited. Research Reports, Report 40.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-52
PUB DATE	Nov 92
NOTE	31p.
PUB TYPE	Reports - Research/Technical (143) -- Statistical Data (110)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adults; Analysis of Variance; *Interrater Reliability; *Language Proficiency; *Scaling; Tables (Data); *Test Format; Test Items; Test Reliability
IDENTIFIERS	*Nonnative Speakers; *Test of Spoken English

ABSTRACT

The Test of Spoken English (TSE) is an internationally administered instrument for assessing nonnative speakers' proficiency in speaking English. The research foundation of the TSE examination described in its manual refers to two sources of variation other than the achievement being measured: interrater reliability and internal consistency. Because the reported data were drawn from a 1980 study of the reliabilities based on two raters, newer data and more extensive analyses were needed. This study uses data from recent administrations of the TSE examination. Analysis of variance examined the effects of scale, section, examinee, and rater, as well as the interactions of these factors. Reliabilities were reported for item, section, and scale scores. Common and unique variance estimates were developed for each scale score for which a section was rated. Estimates of the effects of altering section lengths suggested that some sections should be lengthened and others omitted if reliability were to be maximized. Others suggestions are offered for improving reliability. Thirteen tables present study findings. An appendix describes TSE scale points. (Contains 16 references.) (SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 40
November 1992

Reliability of the Test of Spoken English Revisited

R.F. Boldt

Educational Testing Service

RELIABILITY OF THE TEST OF SPOKEN ENGLISH REVISITED

R. F. Boldt

Educational Testing Service
Princeton, New Jersey

RR-92-52

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

Abstract

The Test of Spoken English (TSE) is an internationally administered instrument for assessing nonnative speakers' proficiency in speaking English. The research foundation of the TSE® examination is currently described in the Manual for Score Users. This publication describes technical characteristics of the test, including such psychometric characteristics as level of difficulty, reliability, and validity. Consistent with the ETS Standards for Quality and Fairness, the Manual refers tc two sources of variation other than the achievement being measured: interrater reliability and internal consistency. Because the reported data on both issues were drawn from a 1980 study of reliabilities based on two raters, newer data and more extensive analyses were needed.

The present study uses data from recent administrations of the TSE examination. Analysis of variance examined the effects of scale, section, examinee, and rater, as well as the interactions of these factors. Reliabilities were reported for item, section, and scale scores. Common and unique variance estimates were developed for each scale score for which a section was rated. Estimates of the effects of altering section lengths suggested that some sections should be lengthened and others omitted if reliability were to be maximized. Other suggestions were offered for improving reliability.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were 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, and 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 and 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. Currently 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. However, many projects require the cooperation of other institutions, 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 the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

Table of Contents

Pages

List of Tables

Introduction

The Test of Spoken English (TSE) is an internationally administered instrument given 12 times per year to an average of approximately 1,100 examinees per administration. The research background is reported in the TSE Manual for Score Users (Test of Spoken English Program, 1990). The principal reliability study reported in the Manual was conducted by Clark and Swinton (1980), who obtained estimates of rater agreement by correlating two raters' evaluations of the same 134 examinees. Reliabilities and intercorrelations were given for pronunciation, grammar, and fluency, as well as for overall comprehensibility.

The present study supplements the Clark and Swinton results using more recent and extensive data and more extensive analyses.

Description of TSE

TSE comprises the following seven sections:

- Section One, which is not scored, is for warmup and allows the examinees to tell why they are taking the test.
- Section Two requires examinees to read aloud.
- Section Three is sentence completion.
- Section Four calls for examinees to tell a story based on their examination of a set of related pictures.
- Section Five requires examinees to answer questions about a single picture.
- Section Six asks for descriptions of common objects or discussions of issues of general interest.
- Section Seven requires explaining a schedule of group activities.

The stems are displayed visually or aurally (tape) and the examinees' oral responses recorded. The replayed responses are then assessed by qualified and trained raters. Examinees receive a score report with ratings on four scales: overall comprehensibility, as well as "diagnostic ratings" on pronunciation, grammar, and fluency.

Description of the Rating Procedure

TSE tape raters are experienced teachers and specialists in English or English as a Second Language who are trained to use the TSE scoring key. During their training, the potential raters listen to actual responses from TSE examinees who have already received scores ranging from low to high performance. Appropriate ratings are discussed in group sessions until discrepancies are resolved. Potential raters then score several TSE examinations that are presented in random order. The scores are discussed with and evaluated by TSE staff, who determine which raters have mastered the rating procedures well enough for operational purposes. Experienced raters undergo regular recalibration and are retrained if scoring discrepancies indicate the need. Previously scored answer tapes are played prior to each rating session to help the raters maintain consistent use of the scoring guidelines.

Raters gather for the tape rating sessions within three weeks of each test administration, as a rule. The ratings are completed within two weeks after a test and scores are generally mailed within three to five weeks.

For this study, each tape was rated by two raters. No further review occurred if their ratings, as averaged over sections, agreed within .95 of a point for each scale. If they differed more than .95 of a point for some scales, an adjudication process was initiated. In this process a third rater evaluated the examinee's performance. If, within the required tolerance, this rater agreed with one of the first two raters, the ratings in closest agreement were used. If all three evaluations differed excessively, a fourth rater was used, but four was the maximum. Two raters were sufficient for 91 percent and 95 percent of the 1,528 and 1,366 examinees from the October and November 1990 administrations, respectively. (As of July 1992, three raters were used only for the overall comprehensibility scale.)

Integer ratings of zero to three were assigned for each item in each section. Brief definitions of the rating levels for each scale are given in the Appendix.

Score Computations

Though all of the sections are rated on at least one of the four possible scales--pronunciation, grammar, fluency, and overall comprehensibility--not all sections are rated on all scales. If ratings by two raters agree sufficiently, the TSE scores are computed as follows:

For a given scale, ratings on the items contributing to that scale are averaged for each section and the section averages are averaged. The distributions of items to sections are given in Table 1. As an example, Table 1 indicates that 10 items appear in Section III and that four items appear in Section V. The items in Section III and Section V are both scored on the grammar scale, so that two grammar averages are produced, one for Section III based on 10 ratings, and one for Section V based on four ratings. The average of these two averages, rounded to two significant figures, comprises the grammar score because that scale is not used in any other section. All the averages are unweighted.

Table 1

Numbers of Items and Rating Scales for Each Section of TSE

Section Number	Name	Number of Items	Scales Rated[a]
I	Warmup	--	None
II	Reading Aloud	1	P,F,C
III	Sentence Completion	10	G,C
IV	Picture Sequence	1	P,F,C
V	Single Picture	4	P,G,F,C
VI	Free Response	3	P,F,C
VII	Short Presentation	1	P,F,C

[a] The Initials P, G, F, and C refer to pronunciation, grammar, fluency, and overall comprehensibility, respectively.

Source and Completeness of Data

Data Source. Data for the present study were from the compl^te operational computer records of the October and November 1990 administrations. For the October administration a total of 1,528 records were examined; for November that total was 1,366. Parallel analyses of data from the two administrations were conducted to provide comparisons.

Missing Responses. It has been mentioned that items were rated twice unless the ratings disagreed sufficiently, in which case one or two additional ratings were secured. The data files, however, contain, for each item, only the scores assigned by the first two raters, regardless of whether other raters were involved. These item records, therefore, provide information on the frequency with which raters failed to assign a score or agree.

There being data for 1,528 and 1,366 examinees from the October and November administrations--2,894 examinees in all--and two raters per examinee, 5,788 ratings were recorded for each item for each applicable scale. For example, 2,894 examinees responded to the first item in Section II and two raters evaluated each response for pronunciation. Thus, 5,788 ratings for pronunciation were recorded in the data file for that item. These unadjudicated scores constitute a census of initial evaluations of that item for that scale.

The computer files used here, like most computer files, included positions for which scores had not been entered. The missing scores were counted for each item for each scale for which that item was scored, with the following results. The maximum number of missing scores for any item on a single scale was 29, or .36 percent, which yields a minimum complete data figure of 99.64 percent. Those cases with missing scores were dropped from the analysis.

Missing Raters. Because there were up to four raters, there are up to six possible pairs of first to fourth raters who were in agreement within the required tolerance. Table 2 presents the numbers and combinations of the agreeing rater pairs in the October and November administrations. In this table the raters are numbered 1 through 4 for the order in which they scored each particular tape.

Table 2

Distribution of the Number of Rater Patterns for the October and November TSE Administrations

Rater	October		November	
Agreement Patterns	Number	Percent	Number	Percent
1-2	1397[a]	91	1279[b]	94
1-3	66	4	40	3
1-4	2	0	0	-
2-3	60	4	47	3
2-4	1	0	0	-
3-4	2	0	0	-

[a] Of the 1528, 1494 or 98% had complete item data.
[b] Of the 1366, 1344 or 98% had complete item data.

Examination of Table 2 indicates that there was seldom a need for more than two raters. Further, for both administrations 98 percent of the examinees had complete item-rating data. For this study the examinees with complete item data from the first two raters were used, whether or not another rater was required. This decision was made because it was felt that the unadjudicated rating should be more generalizable. The results should be similar to those that would be obtained if the adjudicated cases were included because adjudication was so seldom needed.

Sources of Score Variance

Modern reliability analysis focuses on assessing the amount of score variance from various sources and the effects on these amounts of various test changes. Table 3 presents an analysis of variance by scale, section, examinee, and raters within examinees. The term "raters within examinee" is explained as follows. It has already been pointed out that raters differ from examinee to examinee, and that tape assignments to raters are more or less random. While it is true that a "Rater 1" and a "Rater 2" are identified on data tapes, Rater 1 for one examinee is not necessarily the same person as Rater 1 for another examinee. Rather, each pair of raters is regarded as being drawn at random for each examinee, a procedure that is consistent with regarding the ratings as being nested within examinee. Therefore, the data are configured in a "scale by section by examinee by rater within examinee" design, with 3, 5, N, and 2 as the number of levels for each facet of the design, where N is the number of examinees--1,494 for October and 1,344 for November.

Note there are only three scales in the design and five sections. This is because Section III scores and the grammar scores were omitted from the analysis in order to achieve a balanced design. Section III and grammar scores were, however, subsequently analyzed and their unique effects assessed.

Table 3
Analysis of Variance and
Component Analysis

October Data

		Sums of	Mean	
Source	d.f.	Squares	Squares	Sample %
Scales(A)	2	16.47	8.24	.09
Sections(B)	4	362.16	90.54	1.67
AxB	8	162.81	20.35	1.12
Examinees(E)	1,493	10,112.65	6.77	37.40
AxE	2,986	807.09	.27	4.47
BxE	5,972	2,052.67	.34	9.49
AxBxE	11,944	10,014.89	.08	7.03
Judges w. E(JwE)	1,494	1,671.31	1.12	12.35
AxJwE	2,988	505.82	.17	5.60
BxJwE	5,976	1,090.03	.18	10.07
AxBxJwE	11,952	767.92	.06	10.64
	November Data			
Scales(A)	2	3.37	1.68	.02
Sections(B)	4	306.13	76.53	1.65
AxB	8	152.56	19.07	1.23
Examinees(E)	1,343	8,722.12	6.49	37.80
AxE	2,686	575.28	.21	3.74
BxE	5,372	1,728.84	.32	9.36
AxBxE	10,744	848.66	.08	6.89
Judges w. E(JwE)	1,344	1,557.54	1.16	13.49
AxJwE	2,688	374.17	.14	4.86
BxJwE	5,376	915.19	.17	9.90
AxBxJwE	10,752	677.99	.06	11.01

With the exception of the last column on the right, the entries in Table 3 are the usual ones of source, degrees of freedom, and mean square. If examinees and judges within examinees are regarded as random effects with all others fixed, the Tukey-Cornfield algorithm (Winer, 1962, p. 195-199) indicated the following mean-square ratios for significance tests: A, B, and AB against the corresponding interactions with examinee; and E, AE, BE, and ABE against the corresponding interactions with JwE. The AxBxJwE interaction might be used as an error term for JwE and its other interactions, but it is not ideal. This is so because it contains one more variance component than the ideal, which cannot be estimated. Even so, it is the smallest mean square in the table. This indicates that all F-values would exceed one. Thus, given the very large degrees of freedom, there is no reason to conclude that any of the variance components are zero.

Because they are affected by the numbers of scales, raters, examinees, and sections, the mean squares do not give a clear indication of the magnitudes of the variance components. Therefore the right-hand column of Table 3, which indicates the relative magnitudes of those components, was included. This column gives the percent of total variance contributed by each source if single observations were drawn with replacement a large number of times, with each effect, including examinee and judge within examinee, being regarded as fixed. Thus, column three regards each data set as fixed regardless of how the raters and examinees were associated, or how the examinees were obtained. With these assumptions, the Tukey-Cornfield algorithm yields divisors that were used to remove the confounding of data structure with size of variance component. The resulting estimates of the components were converted to percents and given in column five of Table 3.

The percentages in the right-hand column of Table 3 indicate that the largest proportion of variance by far was contributed by the examinees, with relatively small contributions by scale, section, and their interactions. But to the extent that section and scale contributed differential variance to the scores, the interactions of these variables with each examinee should be substantial. It was especially desirable that the scale by examinee interaction be large, because the scales should be measuring different skills. However, they are smaller than the variance attributable to rater differences (JwE). The exploration of common and unique variances in the scores is pursued in a section that follows, where it will be seen that Section III makes a greater unique contribution than did those used in the analysis that led to Table 3.

Reliabilities

Item Score Reliabilities. Tables 4 through 7 present reliabilities of October and November scores on the pronunciation, grammar, fluency, and overall comprehensibility scales, respectively. Two types of coefficients were used: intraclass correlations (Model I in Shrout and Fleiss, 1979), and intraclass correlations corrected for two raters using the Spearman-Brown prophesy formula (Gulliksen, 1987). The intraclass correlation estimates the ratio of item true variance to item true plus error variance.

Table 4

Item Intraclass Rs for One and Two Raters: Pronunciation Scores

	October		November	
Section	ICR[a]	R_2[b]	ICR	R_2
II[c]	.55	.71	.52	.68
IV	.53	.69	.48	.65
V	.39	.56	.46	.63
V	.46	.63	.51	.68
V	.50	.67	.48	.65
V	.50	.66	.50	.67
VI	.55	.71	.53	.69
VI	.51	.68	.54	.70
VI	.56	.72	.51	.68
VII	.53	.70	.53	.69

[a] ICRs are intraclass correlation coefficients.
[b] R_2s and ICRs corrected for double length.
[c] There is one row per item.

Table 5

Intraclass Rs for One and Two Raters: Grammar Scores

	October		November	
Section	ICR[a]	R_2[b]	ICR	R_2
III[c]	.56	.72	.62	.77
III	.52	.69	.59	.74
III	.63	.78	.69	.82
III	.61	.76	.66	.80
III	.64	.78	.62	.77
III	.62	.76	.63	.77
III	.64	.78	.65	.79
III	.58	.73	.63	.77
III	.60	.75	.63	.77
III	.58	.74	.62	.77
V	.47	.64	.46	.63
V	.55	.71	.54	.70
V	.52	.69	.55	.71
V	.50	.67	.55	.71

[a] ICRs are intraclass correlation coefficients.
[b] R_2s and ICRs corrected for double length.
[c] There is one row per item.

Table 6

Intraclass Rs for One and Two Raters:
Fluency Scores

	October		November	
Section	ICR[a]	R_2[b]	ICR	R_2
II[c]	.39	.56	.42	.59
IV	.48	.65	.52	.69
V	.36	.53	.43	.60
V	.53	.69	.53	.69
V	.53	.69	.54	.70
V	.50	.67	.53	.69
VI	.58	.73	.52	.68
VI	.54	.70	.57	.73
VI	.57	.73	.56	.72
VII	.53	.69	.54	.70

[a] ICRs are intraclass correlation coefficients.
[b] R_2s and ICRs corrected for double length.
[c] There is one row per item.

Table 7

Intraclass Rs for One and Two Raters,
Overall Comprehensibility

Section	ICR[a]	R_2[b]	ICR	R_2
II[c]	.54	.70	.46	.63
III	.54	.70	.57	.73
III	.52	.69	.57	.72
III	.59	.74	.59	.74
III	.59	.74	.63	.77
III	.54	.70	.51	.68
III	.53	.69	.55	.71
III	.62	.77	.59	.74
III	.51	.67	.50	.67
III	.58	.73	.61	.76
III	.56	.71	.56	.72
IV	.53	.69	.52	.69
V	.39	.56	.45	.62
V	.53	.69	.54	.70
V	.58	.74	.55	.71
V	.53	.69	.53	.69
VI	.59	.74	.52	.69
VI	.54	.70	.60	.75
VI	.60	.75	.59	.74
VII	.55	.71	.53	.69

[a] ICRs are intraclass correlation coefficients.
[b] R_2s and ICRs corrected for double length.
[c] There is one row per item.

Note that the October and November results were highly similar. This result was obtained all through the analyses and will be not commented on further. As item reliabilities, the intraclass correlations were substantial, even more so when the use of two raters (R_2) was taken into account.

Section Score Reliabilities. As explained above, scores on ratings for items in a section are averaged to produce a scale score. For example, an examinee's ratings on grammar for each of the 10 items in Section III are averaged to produce a grammar score for that section. The reliabilities for these scores are presented in Table 8.

Table 8

Intraclass Rs for One and Two Raters:
Scale Scores by Section

October		November		Sect.	No.
ICR	R_2	ICR	R_2	No.	Items
		Pronunciation			
.55	.71	.52	.68	II	1
.53	.69	.48	.65	IV	1
.55	.71	.57	.73	V	4
.61	.76	.58	.73	VI	3
.53	.69	.53	.69	VII	1
		Grammar			
.73	.84	.73	.84	III	10
.62	.77	.63	.78	V	4
		Fluency			
.39	.56	.42	.59	II	1
.48	.65	.52	.69	IV	1
.58	.73	.60	.75	V	4
.65	.79	.62	.76	VI	3
.53	.69	.54	.70	VII	1
	Overall	Comprehensibility			
.54	.70	.46	.63	II	1
.66	.79	.67	.80	III	10
.53	.69	.52	.68	IV	1
.62	.77	.62	.76	V	4
.67	.80	.65	.79	VI	3
.55	.71	.53	.69	VII	1

Scale Score Reliabilities. After an examinee's item responses are scored on the appropriate scales and the item scores are averaged within sections, as explained above, the section scores are again averaged to produce scores on the four scales: pronunciation, grammar, fluency, and overall comprehensibility. For example, an examinee's average item scores for Sections II, IV, V, VI and VII are averaged to produce the final rating for fluency. These scores are then averaged. Intraclass correlations yield single-rater reliabilities, which are presented in Table 9 along with reliabilities for two raters obtained using the Spearman-Brown correction for double length. The corrected reliabilities reflect the effect of averaging the two raters.

Table 9

Intraclass Rs for One and Two Raters:
Scale Scores

	October		November	
Scale	ICR	R_2	ICR	R_2
Pronoun	.66	.79	.64	.78
Grammar	.72	.83	.71	.83
Fluency	.65	.79	.66	.79
Comp	.71	.83	.69	.81

Effect of Dropping Sections on Scale Reliabilities. One way of assessing the contribution of a section to scale score reliability is to note the reliability of the scale score that results when the section is dropped. Table 10 presents such data. In Table 10, the number of the section that was dropped is given in the fifth column. Thus, the first line of Table 10 indicates that the reliability of the pronunciation score in October was .64 if Section II was not used in computing that score. The portion of Table 10 that presents data for the grammar score further clarifies the entries. Inasmuch as only two sections contribute to the grammar score, the reliability of that score is based only on the second when the first is dropped, and vice versa. For this reason, the first four entries in the first grammar line of Table 10 are the same as the first four entries in the second grammar line of Table 8.

Table 10

Intraclass Rs for One and Two Raters:
Effect of Dropping Sections

October		November		Sect.	No.
ICR[a]	R_2	ICR	R_2	No.	Items
Pronunciation					
.64	.78	.63	.77	II	1
.66	.80	.64	.78	IV	1
.65	.79	.62	.77	V	4
.65	.79	.63	.77	VI	3
.65	.79	.63	.77	VII	1
Grammar					
.62	.77	.63	.77	III	10
.73	.84	.74	.85	V	4
Fluency					
.66	.80	.67	.80	II	1
.64	.78	.65	.79	IV	1
.63	.77	.64	.78	V	4
.61	.76	.64	.78	VI	3
.64	.78	.65	.79	VII	1
Overall Comprehensibility					
.71	.83	.70	.82	II	1
.70	.82	.67	.80	III	10
.71	.83	.68	.81	IV	1
.70	.82	.68	.81	V	4
.69	.82	.68	.81	VI	3
.71	.83	.69	.81	VII	1

[a] ICRs are intraclass correlation coefficients computed after dropping the section indicated in column 5.
[b] R_2s and ICRs corrected for double length.

Dropping Sections II, III, and V has been suggested. When these sections were dropped from the October calculations the resulting reliabilities were .64, .64, and .67 for pronunciation, fluency, and overall comprehensibility, respectively. The corresponding figures for November were .61, .64, and .65. Correcting these figures for two raters using Spearman-Brown yields .78, .78, and .80 for October, and .76, .78 and .79 for November. No figure is available for the grammar scale because it is based only on Sections III and V, which were not used in the calculations.

Common and Unique Factor Variance

The reliability figures in the previous section index agreement on two components: (a) an "error" due to disagreement between raters' evaluations of the same tape, and (b) an item's "true score." The true score for an item reflects rater agreement on evaluation of that item performance. It is possible, however, that raters might agree on their ratings of performance on each of two items, but might judge a different aspect of performance for one item than for the other (even though instructed not to do so). Thus, ratings could also be considered as having three components: (c) an "error" that indexes disagreement between raters' evaluations of the same tape, (d) whatever it is that is reliably rated that is common to the items rated, and (e) whatever it is that is rated reliably that is unique to the items rated. Components (d) and (e), described in the previous sentence, result from a partitioning of component (b), the true score. Component (c), which is the same as component (a), and (e) are often referred to as "specificity." Thus, reliable variance is regarded as being made up of true-score variance and uniqueness, and specificity is regarded as being made up of uniqueness and error.

One indication of the existence of unique variance in the TSE ratings comes from the data on Section III. The 10 items in this section are scored on grammar. The items have similar interrater reliabilities that, for October, were .598 on the average. The Spearman-Brown prophesy formula (Gulliksen, 1987) can be used to calculate what the reliability of a test would be with a given number of parallel items. This formula indicates that a test with 10 parallel items, each item having a reliability of .598, should have a reliability of .94. Note however, that the reliability for the grammar score, based on items in Section III, is only .73. Thus the Spearman-Brown formula results were not accurate.

The discrepancy observed, .94-.73, arises because TSE scoring allows one to estimate the reliability for an item apart from the other items, which is not possible in most applications of the Spearman-Brown formula. To use Spearman-Brown appropriately we must assume that the true score that produces item reliability, i.e., rater reliability, is the same true score that produces item intercorrelations. This is so because the "item reliability" on which the Spearman-Brown derivation is based is an average item intercorrelation. It works for multiple-choice items because such item reliabilities are based on item scores whose reliabilities are, in contrast to the reliabilities of the TSE items, not separately assessed. That is, it did not work here because item reliability could be estimated apart from item intercorrelations and because there are substantial specific components in the grammar ratings, as will be seen.

Table 11 contains partitionings of rating variance into fractions of common, unique, and error variance by section for each scale. For each particular month's data for a scale, the partitioning was done as follows. First, the variance-covariance matrix of scores was computed. Then, a single factor was extracted using the Minres criterion (Harman, 1966). This method of factor extraction fits factor loadings to the off-diagonal entries of the matrix, i.e., the covariances, in this case. Squaring these loadings yields an estimate of the common factor variance that, when subtracted from the estimated true-score variance, yields an estimate of the specific variance (because true comprises common and unique).

Table 11

Partitioning of Rating Variance

October			November				No.
Common	Unique	Error	Common	Unique	Error	Section	Items
Pronunciation							
.25	.01	.22	.21	.01	.20	II	1
.25	-.02	.20	.22	-.03	.21	IV	1
.18	-.02	.13	.18	-.01	.13	V	4
.28	-.04	.15	.25	-.04	.15	VI	3
.28	-.04	.21	.27	-.03	.21	VII	1
Grammar							
.09	.04	.05	.08	.05	.05	III	10
.19	-.02	.10	.18	-.01	.10	V	4
Fluency							
.11	.04	.23	.12	.03	.21	II	1
.25	-.02	.24	.24	-.01	.21	IV	1
.16	.00	.12	.17	.01	.12	V	4
.29	-.01	.15	.27	-.03	.15	VI	3
.27	-.02	.22	.26	-.02	.20	VII	1
Overall Comprehensibility							
.18	.04	.19	.15	.01	.19	II	1
.13	.02	.08	.13	.02	.07	III	10
.23	-.01	.20	.22	-.01	.20	IV	1
.19	.00	.11	.19	.00	.12	V	4
.27	.00	.14	.25	-.01	.13	VI	3
.26	-.02	.2	.25	-.03	.20	VII	1

Examination of Table 11 reveals a number of negative numbers in the columns pertaining to unique variance. But one should reasonably expect unique variance to be zero or positive, because uniqueness is a part of true score. Why, then, were the estimates of unique variance sometimes negative? The reason is that factor model and the intraclass model differ somewhat in structure and were not constrained to produce numerical consistency. One way in which they differ in this project is that the factor model used data from all the sections to estimate communalities, but the intraclass model applied to the data by section separately to estimate the true-score variance. The two models are somewhat different, though not inconsistent, views of the data. In this case, if uniqueness is zero, then true-score variance estimated in the reliability calculation should equal the common variance from the single-factor extraction, though they were not constrained to do so. The occurrence of several instances where the estimates of communality substantially exceeded the estimates of true-score variance would be very troublesome, indicating a problem with the models or computations. In this case, the negative estimates of uniqueness were small.

There are two cases where the true-score variance substantially exceeded the communality, both occurring in the grammar ratings for Section III. In these cases the unique variance is on the order of magnitude of the error variance, a condition that holds for the uniqueness of no other section or scale. Though the rater agreement is very high for this scale, relatively more of its variance is unique. Noting this fact, a single factor was extracted at the item level. This result revealed substantial unique variance at the item level for Section III grammar ratings.

Optimum Section Length. Several section-related statistics determine the reliability of a scale score: section reliability, variance common to sections, uniqueness, rater-error variance, and section length in terms of time. These quantities were all available as a result of the present study and were used to explore the effects of varying section length and scoring weight on scale reliability.

Table 12 displays the information used to examine optimum section length. The common and unique entries in Table 12 are derived from those of Table 11 as follows: If the uniqueness was positive, the common and unique entries were just copied from Table 11 to Table 12. Wherever the uniqueness in a row of Table 11 was negative, the following steps were followed: (a) that uniqueness was added to the entry in the column headed "common" for the same row and month, (b) the resulting sum was copied to the column headed "common" for the same row and month in Table 12, and (c) .00 was entered as the corresponding uniqueness in Table 12. This procedure imposed the true-score variance as the upper bound for common factor variance in Table 12. Also in Table 12, the testing time in seconds for a section is included in column eight. Testing times were taken from the Bulletin of Information (TOEFL/TSE Services, 1989).

Table 12

Re-Partitioning of Rating Variance for Optimum Section-Length Analysis

October			November			Testing	
Common	Unique	Error	Common	Unique	Error	Section	Time
			Pronunciation				
.25	.01	.22	.21	.01	.20	II	120[a]
.23	.00	.20	.19	.00	.21	IV	120
.16	.00	.13	.17	.00	.13	V	108
.24	.00	.15	.21	.00	.15	VI	135
.24	.00	.21	.24	.00	.21	VII	120
			Grammar				
.09	.04	.05	.08	.05	.05	III	100
.17	.00	.10	.17	.00	.10	V	108
			Fluency				
.11	.04	.23	.12	.03	.21	II	120
.23	.00	.24	.23	.00	.21	IV	120
.16	.00	.12	.17	.01	.12	V	108
.28	.00	.15	.24	.00	.15	VI	135
.25	.00	.22	.24	.00	.20	VII	120
		Overall	Comprehensibility				
.18	.04	.19	.15	.01	.19	II	120
.13	.02	.08	.13	.02	.07	III	100
.22	.00	.20	.21	.00	.20	IV	120
.19	.00	.11	.19	.00	.12	V	108
.27	.00	.14	.24	.00	.13	VI	135
.24	.00	.2	.22	.00	.20	VII	120

[a] Given in seconds.

The scale reliability calculation consisted of maximizing a type of reliable variance holding error variance constant. The calculation used weighted sums of item averages for sections of variable length. The two types of reliable variance that were maximized were: (a) common factor variance, and (b) common plus unique factor variance.

It was not possible to express the formula for optimum section length in closed form, but by iteration, the section lengths and weights could be calculated. (Optimum lengths were found for equal weights, then optimal weights for the lengths just found, then optimal lengths, and so on.) The result of the iterations was that all but one section would be reduced to zero length, with all testing time devoted to the remaining section. This outcome suggested comparing section reliabilities that would be obtained if the time limits for each section were equal to the total examination time currently in effect. The results are presented in Table 13.

Table 13

Estimated Scale Reliabilities for Lengthened[a] Sections

	October		November	
Section	Non-U[b]	U[c]	Non-U	U
		Pronunciation		
II	.84	.87	.83	.87
IV	.87	.87	.84	.84
V	.89	.89	.89	.89
VI	.89	.89	.88	.88
VII	.87	.87	.87	.87
		Grammar		
III	.66	.95	.58	.95
V	.92	.92	.92	.92
		Fluency		
II	.58	.79	.65	.81
IV	.85	.85	.87	.87
V	.90	.90	.86	.91
VI	.91	.91	.89	.89
VII	.87	.87	.87	.87
		Overall Comprehensibility		
II	.71	.87	.78	.83
III	.81	.93	.81	.94
IV	.87	.87	.86	.86
V	.92	.92	.91	.91
VI	.91	.91	.91	.91
VII	.88	.88	.87	.87

[a] Times include actual prompt examination and answering (703 sec.), but not directions.
[b] Reliabilities omitting unique variance in numerator.
[c] Reliabilities including unique variance in numerator.

As has been mentioned, the entries in Table 13 assume that all the testing time would be devoted to the sections indicated. This procedure puts the figures on a common base. Note that most of the figures are substantial. Thus the finding that optimum reliability iterations drive out all sections but one indicates that the optimization might be capitalizing on small differences. These results will be extended in the discussion.

Discussion

Examination of the data from two administrations of the Test of Spoken English indicated that very few data were missing either by item or by rater. Complete data were available for the first two raters for over 98 percent of the examinees, so these examinees were used for the subsequent analysis. The data were sufficiently complete that the impact of adjudication on score reliability was felt to be minimal. This is not to suggest that adjudication of scores is not necessary. Because the data developed were so complete, the results of this study are more generalizable than they would have been if extensive use of adjudication procedures had been necessary. Reliabilities that are realized by obtaining repeated ratings when disagreements occur must be regarded as spurious, at least in terms of their implications about the agreement of trained judges in general. The occurrence of the need for frequent adjudication of rater disagreements indicates a problem with the rating system, and the eventual success of adjudication does not contraindicate the existence of the problem. However, this does not seem to be a problem for the Test of Spoken English.

The variance of scores was analyzed using section, scale, examinee, judge within examinee, and their interactions as variance sources. All sources were found to make significant contributions. The percent contribution from each source to the variance of single observations drawn randomly with replacement was computed. These percents were used to index the magnitude of variance contribution from the sources. The largest contributor to this variance was the examinee, followed in size by the interaction of scale and section with examinee, and then by judge within examinee. These percents indicated that there was some unique contribution of scales and sections to examinee scores, but the examinee-by-scale interaction, which is the most desirable in terms of the purpose of the test, was not large. That is, the large size of the examinee main effect indicated a substantial limitation of possible differential validity of the diagnostic scores. However, this interpretation should be tempered by the fact that Section III, Sentence Completion, and the grammar scale were not included in the analysis of variance. Correlation and factor analysis techniques were then used to explore variance components further.

This study's use of intraclass correlations is a practice that contrasts with that used by Clark and Swinton (1980), who computed product-moment correlations between scores assigned by two raters. The operational data used here did not permit this approach for two reasons. First, the largest number of cases rated by the same rater pair at either administration was 36, and the rest of the rater pairs were used even less frequently. Hence, only very small numbers of cases would have been available for product-moment correlations. Second, the product-moment correlation approach requires a set of ordered data pairs. It was pointed out in an earlier section that there is no logical way to order these data. Hence, generating correlations of "Rater 1" scores with "Rater 2" scores, while computationally feasible, would not be meaningful. The rater session seemed to be much better described as a one-way layout with examinees as the main effect and with rating pairs being nested within examinees. This latter description lent itself to the intraclass correlation.

Correlations were developed by item, by scale for each section, and for each scale. In sum, these results indicated no strikingly defective items, but differences were noted by section and scale. Section III was found to yield the most reliable item and scale scores, and elimination of that section without other adjustment might reduce the reliability of the ratings of grammar and overall comprehensibility. Partitioning or rater variance indicated that the grammar score based on Section III ratings had a high specificity relative to rater disagreement, as indexed by error variance.

An analysis was undertaken that sought to determine test lengths that shou'd yield optimum reliabilities for the scales. The research revealed that maximum reliability would be achieved for each scale by selecting a single section and lengthening it to the time available. Based on this result, the reliabilities were calculated for standard section times using the total testing time as the standard. These calculations indicated maximum reliabilities for Sections V and VI for pronunciation and fluency. For grammar, the reliability for Section III was highest if specificity was included as a reliable variance, but the reliability for Section V was highest if specificity was not regarded as a contributor to reliability. The overall comprehensibility scale reliability would be highest if based on Section III, unless uniqueness was not counted, in which case Section V (October) or Sections V and VI (November) were highest.

These results suggest that scores from some combination of Sections III, V, and VI could improve the reliability of TSE-scaled scores. Most reliabilities in Table 13 are on the order of .9, which is a generally acceptable standard of reliability for scores used in decision making.

Using estimates based on equal test time yielded reliabilities for Sections II, IV, and VII that were never largest for any scale. The differences among these reliabilities were often not large. Even so, the following comment on Sections II, IV, and VII seems warranted. These sections each involve a single item that requires two minutes to generate a very free-flowing answer, which yields only one rating. In contrast, Sections III, V, and VI all produce multiple, highly constrained responses and ratings. Perhaps if some method were found to obtain several ratings on specific aspects of the responses of Sections II, IV, and VII, their reliabilities could be enhanced. The time-standardized reliabilities of Section VII were often nearly as high as the reliability of the most reliable section for the scale being rated, and the use of rather natural-language responses is an asset. It might, therefore, be particularly fruitful to seek an improved scoring method for that section.

APPENDIX

BRIEF DESCRIPTIONS OF TSE SCALE POINTS

Individuals raters rate examinees' item performances on a four-point scale (zero to three). The item ratings within each section are averaged and the section averages averaged in turn to produce a scale average. Brief descriptions are provided to the raters for anchoring their judgments of the item performances. These descriptions are given below.

Pronunciation

0. Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be unintelligible.

1. Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be occasionally unintelligible.

2. Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is intelligible.

3. Occasional nonnative pronunciation errors, but speaker is always intelligible.

Grammar

0. Virtually no grammatical or syntactical control except in simple stock phrases.

1. Some control of basic grammatical constructions but with major and/or repeated errors that interfere with good intelligibility.

2. Generally good control in all constructions, with grammatical errors that do not interfere with overall intelligibility.

3. Sporadic minor grammatical errors that could be made inadvertently by native speakers.

Fluency

0. Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible.

1. Numerous nonnative pauses and/or a nonnative flow that interferes with intelligibility.

2. Some nonnative pauses but with a more nearly native flow so that the pauses do not interfere with intelligibility.

3. Speech is smooth and effortless, closely approximating that of a native speaker.

APPENDIX (Con't)

Overall Comprehensibility

0. Overall comprehensibility too low in even the simplest type of speech.

1. Generally not comprehensible because of frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary items, or lack of grammatical control.

2. Comprehensible with errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing.

3. Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Campbell, D. T. & Fiske, D. W. (1967). Convergent and discriminant validation by multitrait-multimethod matrix. In D. N. Jackson & S. Messick (Eds.), Problems in human assessment. (pp. 124-131). New York: McGraw-Hill, Inc.

Clark, J. L. & Swinton, S. S. (1979). An exploration of speaking proficiency measures in the TOEFL context (TOEFL Research Report 4). Princeton, NJ: Educational Testing Service.

Clark, J. L. & Swinton, S. S. (1980). The Test of Spoken English as a measure of communicative ability in English-medium instructional settings (TOEFL Research Report 7). Princeton, NJ: Educational Testing Service.

DeMauro, G. E. (1988). Construct validity and redundancy of TSE scoring scales. Internal report for TOEFL Programs. Princeton, NJ: Educational Testing Service.

Educational Testing Service (1987). ETS standards for quality and fairness. Princeton, NJ: Author.

Gulliksen, H., 1987. Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Harman, H. H. & Jones, W. H. (1966). Factor analysis by minimizing residuals (Minres). Psychometrika, 31, 351-368.

Powers, D. E. & Stansfield, C. W. (1983). The Test of Spoken English as a measure of communicative proficiency in the health-related professions (TOEFL Research Report 13). Princeton, NJ: Educational Testing Service.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psych. Bull., 86, 420-438.

Society for Industrial and Organizational Psychology (1987). Principles for the validation and use of personnel selection procedures: third edition. College Park, Md.: Author.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (Ed.), *Direct testing of speaking proficiency: theory and application* (pp. 1-12). Princeton, NJ: Educational Testing Service.

Test of Spoken English. (1990). *Test of Spoken English manual for test users*. Princeton, NJ: Educational Testing Service.

TOEFL/TSE Services. (1989). *Bulletin of information for TOEFL and SE*. Princeton, NJ: Educational Testing Service.

Wilds, C. P. (1975). The Oral Interview Test. In R. Jones and B. Spolsky (Eds.), *Testing language proficiency* (pp. 29-44). Arlington, VA: Center for Applied Linguistics.

Winer, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

57906-01202 • Y102M 5 • 275628 • Printed in U S A