DOCUMENT RESUME

ED 383 724 TM 023 145

AUTHOR Boldt, Robert F.; Oltman, Philip K.
TITLE Multimethod Construct Validation of the Test of Spoken English. Report 46.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-58
PUB DATE Dec 93
NOTE 30p.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Audiotape Recordings; *Construct Validity; Correlation; *English (Second Language); Factor Structure; *Interrater Reliability; *Language Proficiency; Multidimensional Scaling; Performance; Scores; *Scoring; Test Construction
IDENTIFIERS Exploratory Factor Analysis; *Test of Spoken English

ABSTRACT

Administration of the Test of Spoken English (TSE) yields tapes of oral performance on items within six sections of the test. Trained scorers subsequently rate responses using four proficiency scales: pronunciation, grammar, fluency, and overall comprehensibility. This project examined the consistency of statistical relations among TSE scores with the measurement constructs these scores purport to reflect by examining dimensions underlying the scores. Multidimensional scaling analyses revealed that three-dimensional solutions fit the scale intercorrelations with low stress values and that the coordinates of the scales fell into three clusters in three-dimensional space, those clusters being defined primarily by test section rather than by proficiency scales. An exploratory factor analysis revealed that a single dimension dominated the variation of the 18 section-scale scores. However, when tapes for the scale scores with substantial discrepancies were rerated, agreement between the order of original and rerated scale scores far exceeded chance. This indicated that raters were able to modify their judgments according to the scale being rated. Subsequent exploratory factor analysis indicated that both section and scale factors contribute to score variation. The factors were highly correlated, and the predominance of the single factor in the exploratory analysis was seen as arising from those correlations. Six tables present analysis results. (Contains 12 references.) (Author/SLD)

ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 96-2
DECEMBER 1996

Multimethod Construct Validation of the Test of Spoken English

Robert F. Boldt

Philip K. Oltman

Educational Testing Service

Multimethod Construct Validation of the Test of Spoken English

R. F. Boldt and P. K. Oltman

Educational Testing Service
Princeton, New Jersey

RR-93-58

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

Abstract

Administration of the Test of Spoken English (TSE) yields tapes of oral performance on items within six sections of the test. Trained scorers subsequently rate responses using four proficiency scales: pronunciation, grammar, fluency, and overall comprehensibility. This project examined the consistency of statistical relations among TSE scores with the measurement constructs these scores purport to reflect.

Analyses included factor analysis and multidimensional scaling, which examined dimensions underlying the scores. These dimensional methods were applied to the 18 scores yielded by the combinations of section and scale. Another analysis was applied to scale scores averaged over sections. This analysis compared the ranking of pairs of scale scores obtained during the original scoring of selected taped performances with the ranking resulting from a rescoring by different raters.

Multidimensional scaling analyses revealed that three-dimensional solutions fit the scale intercorrelations with low stress values, and that the coordinates of the scales fell into three clusters in three-dimensional space, those clusters being defined primarily by test section rather than by proficiency scales.

An exploratory factor analysis revealed that a single dimension dominated the variation of the 18 section-scale scores. However, when tapes for scale scores with substantial discrepancies were rerated, agreement between the order of original and rerated scale scores far exceeded chance. This indicated that raters were able to modify their judgements according to the scale being rated. Subsequent exploratory factor analysis indicated that both section and scale factors contribute to score variation. The factors were highly correlated, and the predominance of the single factor in the exploratory analysis was seen as arising from those correlations.

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, 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 data confidentiality will be protected.

Current (1992-93) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins (Chair)	Southern Illinois University at Carbondale
Linda Schinke-Llano	Millikin University
John Upshur	Concordia University

Table of Contents

List of Tables

Introduction

The principal aim of the proposed study was to determine whether ratings obtained when scoring the TSE, one of the tests offered as part of the ETS package for assessing the English language skills of nonnative speakers, are consistent with the constructs they are intended to reflect. Administration of the TSE yields tapes of oral performance on six sections of the test. Trained scorers subsequently evaluate the taped performances using four proficiency scales. The scales are pronunciation, grammar, fluency, and overall comprehensibility. Three approaches to developing evidence of relating construct validity were used: factor analysis, multidimensional scaling, and rerating of selected tapes. After a more detailed description of the test and of the context of this study--test standards and research background--these three approaches will be described in more detail.

Background

The TSE includes seven sections, six of which are scored (the first section elicits biographical information, and is used for "warmup.") Each of the sections presents a different situation in which spoken response is required, and examinees' responses are recorded on tape for later analysis. The scoring of the TSE responses is carried out by trained judges, who listen to each response and score it on a scale of 0 to 3 on two or more of the following four dimensions, depending upon the section:

- Pronunciation
- Grammar
- Fluency
- Overall Comprehensibility

The sections, and the dimensions on which each is scored, are as follows:

Dimensions / Sections	Pronunciation	Grammar	Fluency	Overall Comprehensibility	Number of Dimensions Scored
Reading Aloud	Scored		Scored	Scored	3
Sentence Completion		Scored		Scored	2
Picture Sequence	Scored		Scored	Scored	3
Single Picture	Scored	Scored	Scored	Scored	4
Free Response	Scored		Scored	Scored	3
Short Presentation	Scored		Scored	Scored	3
Sections Scored	5	2	5	6	Total scores= 18

The scores on individual items within sections are averaged, as are the scores across sections. The score report consists of an overall Comprehensibility score, and diagnostic scores for Pronunciation, Grammar, and Fluency.

Administration and scoring of the TSE requires individual or language laboratory testing conditions, and subsequent scoring by trained raters of the six operational sections on four scales. The process is effective in producing valid measurement of the English speaking proficiency of persons whose native language is not English. However, the procedure as presently constituted is costly. In what is proposed below, analyses would be conducted that would provide potentially useful information for designing modifications of the TSE format to reduce its cost while preserving its measurement effectiveness.

The Test of Spoken English Manual for Score Users (Test of Spoken English, 1990) outlines several types of validity evidence for TSE. It describes procedural steps that are taken to ensure the quality of the scores provided, including evidence of interrater reliability for the constructs measured. A recently completed study by Boldt (1992) provides additional information on TSE reliability. The manual also summarizes several studies. Clark and Swinton (1979, 1980) performed studies of the relation of TSE scores to performance on an oral proficiency interview (Wilds, 1975; Sollenberger, 1978), and with ratings of instructor performance. Powers and Stansfield (1983) reported evidence that subject matter experts in health professions could identify

levels of TSE performance that varied in their acceptability if occurring in health care situations. Finally, DeMauro (1988) examined construct validity and redundance in the TSE for the fluency, pronunciation, and overall comprehensibility scales. Because different sections measure the same constructs but in different ways, DeMauro was able to examine their convergent and discriminant validity (Campbell and Fiske, 1967). That is, correlations among scores on a construct measured in different ways should exceed correlations among scores that measure different constructs. He found a partial confirmation of these types of construct validity in his data. However, using regression methods DeMauro also found a great deal of redundancy in the measures. The present study updates and supplements the existing validation research.

The constructs of the Test of Spoken English (TSE) are those with respect to which the four reported scale scores provide evidence, i.e., pronunciation, grammar, fluency, and overall comprehensibility. The test manual (Test of Spoken English, 1991), contains descriptors of scale points for each of these constructs. The descriptors are meant to refer to discriminable aspects of speech behavior to which the constructs refer, hence item responses and their derived scores should relate to each other in ways that are consistent with our notions of the constructs. Because these scores evaluate samples of recorded speech, TSE assumes that there are four distinct properties of speech on which appropriately instructed raters can agree as to the relative quality of examinees' performances. Clearly, a complete evaluation of these assumptions is well beyond a single, or perhaps even several projects, so the present project can be only a partial validation of the TSE.

Research Approaches

Data

Data for the present study were from the October and November, 1990 administrations. The complete operational computer records of the administrations were examined. For the October administration a total of 1,528 were found; for November that total was 1,366. These data were used in a previous study by Boldt (1992) who reported that 98 percent of the cases had complete item data, and that the first two raters of 91 percent and 94 percent of the October and November cases, respectively, agreed within the standards required for operational use without obtaining further ratings. Because the present study is not concerned with problems of rater disagreement, which was rare, and of incomplete data, which was also rare, we used only those examinees for which the first two raters agreed within the tolerance limits of the program, and for which complete data were present. This yielded 1,369 and 1,261 cases for October and November, respectively.

Factor Analysis

Factor analysis models base the description of observed data on a few underlying "factors." These models feature "factor scores" and "factor loadings." Factor scores associate with examinees, and factor loadings associate with scale scores for sections. When combined, factor scores and factor loadings yield an approximation of the raters' assignments of scale-section scores to examinees. The procedures of this study (Rao, 1955) combine factor scores and loadings to obtain least squares approximations of observations.

Exploratory v. Confirmatory Analyses

Two types of analyses, exploratory and confirmatory, are used in the present study. These analyses impose different degrees of structure on the assumed statistical models. Where a factor analysis is exploratory, the investigator posits some number of factors to account for variation in a data set, and then determines how well that many factors actually approximate the data. The approximation improves as the number of factors posited increases, and at some point additional factors are not warranted.

A confirmatory analysis imposes additional restrictions on an exploratory model. In our study, an exploratory model might posit four factors that contribute to every observed score regardless of the rating scale used, while a confirmatory model might require that one factor contribute only to ratings of pronunciation, another only to ratings of grammar, and so forth.

The exploratory analyses in this study are multidimensional scaling and exploratory factor analysis. In addition, a confirmatory factor analysis was conducted. A rerating analysis was also confirmatory.

Multidimensional Scaling

Multidimensional scaling can be understood in terms of a spatial analogy in which an examinee's scores define a point in N dimensional space (18 dimensional space if we refer to scores from section-scale combinations), and the configuration of points for all examinees forms a cloud in this space. Multidimensional scaling explores the dimensionality of this cloud, which would be 1 if the cloud is essentially a line, 2 if it is essentially a plane, etc. In this analysis, distances between the points making up the cloud comprise the basic data input to the scaling procedure.

The multidimensional scaling results were used as input to a cluster analysis. The cluster analysis isolates from the total

cloud of data points those points that have similar coordinates. That is, multidimensional scaling fits the data into a set of dimensions, and then cluster analysis helps locate regions in the space where data points cluster. The nature of the clusters may reveal the nature of the construct or constructs being measured by the test.

Rerating

Since they are intended for interpretation as different aspects of speech performance, the three diagnostic scores reported for TSE examinees should reflect different aspects of those performances, and should generalize to somewhat different domains of behavior. Thus, according to the Bulletin of Information for TOEFL and TSE (Test of English as a Foreign Language, TOEFL,1992), pronunciation refers to the similarity of sound patterns produced by the examinee to those produced by persons who are proficient in English speech, and to some extent uses intelligibility as a standard; grammar refers to adherence of the word structure to the rules of the language; and fluency refers to the rate of and effort required for speaking. To the extent that these descriptions refer to distinguishable behaviors, speech samples can be strong or weak in different ways. For example, a speech passage that is quickly and easily delivered with well pronounced and easily comprehended words that are, however, arranged and structured ungrammatically should receive different scores than one in which the words are delivered haltingly but understandably and with correct grammatical structure. The first would score low on grammar and high on fluency and pronunciation; the second would score high on pronunciation and grammar and low on fluency.

In the paragraph above, the scores are thought to differ because the speech samples differ. However, what if one found that judgments differ from one measure to another for only a few speech samples? Do the few differences arise because the judges are appreciating real but infrequent occurrences, or is their occurrence a "fluke?" Either is possible.

We selected tapes that differed in that they displayed expected differences in ratings, in contrast with the majority of tapes whose ratings did not display expected differences. The tapes were supplied to raters who rescored the tapes under standard operational conditions. If the differences in the reratings consistently reflected the original differences, then the original differences were not a fluke. One could believe that the measures did indeed reflect different aspects of performance as intended, and that the observation that a single factor existed occurred only because the factoring procedure was not sensitive enough to pick up the seldom occurring differences.

Analyses and Results

Exploratory Factor Analyses

It has been mentioned that in exploratory factor analysis the investigator posits that some number of factors underlies a set data, and then determines how well a system of that many factors reproduces the data. Factors are added until at some point further numbers of factors cease to improve the data reproduction. The data reproduction provided by a single factor in this study abbreviated the exploratory analysis.

Table 1 presents the percents of standard score variation accounted for by each of the 18 factors present. For both the October and November data, examination of the table reveals a large first factor followed by small and gradually diminishing subsequent factors. This gradual progression from one large factor through a series of gradually diminishing factors is common when one factor is present with errors. Hence, Table 1 indicates that, for both months, there is essentially one factor with additional factors accounting for very little score variance. It will be seen that more factors were identified by other methods, ten in all, than were identified initially by this exploratory method.

Table 1

Percent of Section-Scale Scores Accounted for by Each Factor

Factor Number	Month Oct	Nov
1	72.12%	72.34%
2	5.45%	4.99%
3	4.15%	3.99%
4	3.02%	3.50%
5	2.91%	2.59%
6	2.56%	2.55%
7	1.85%	1.91%
8	1.37%	1.53%
9	1.15%	1.11%
10	0.99%	1.00%
11	0.96%	0.93%
12	0.68%	0.71%
13	0.63%	0.68%
14	0.58%	0.55%
15	0.49%	0.47%
16	0.45%	0.45%
17	0.35%	0.36%
18	0.30%	0.32%

Multidimensional Scaling & Cluster Analysis

Multidimensional scaling was the second exploratory method used in this study. The October and November data sets were analyzed separately and independently. Intercorrelations were computed among the 18 scores that were derived from the dimensions that were scored on the six sections of the test. Multidimensional scaling was performed on the correlation matrices using the nonmetric method of Kruskal (1962). Although there are several approaches to deriving multidimensional scaling analyses, they perform essentially the same function. Coordinates for a set of points in a space are computed so as to fit the empirical similarities among a set of objects (test items, scales, and so forth). Multidimensional scaling aims to maximize goodness of fit by minimizing a value known as *stress*; values near zero indicate good fit. Stress is defined as the square root of the sums of squares of discrepancies between interpoint distances from the scaling plot and smooth distances predicted from the similarities among the objects. For the calculation of stress, each set of distances is normed so that the sum of squares is one. Three-dimensional solutions yielded low stress values in both samples, and using higher numbers of dimensions did not improve the fit very much. The stress values for the three-dimensional solutions were .08 and .07 for the October and November data, respectively.

The matrix of coordinates of each of the 18 scores in three-dimensional space was subjected to hierarchical cluster analysis using Ward's method applied to the Euclidean distances among the points. The points corresponding to the 18 scores appeared to congregate into either two or three main clusters in the space, depending on the criterion used for determining the number of clusters. At each stage of the clustering process, a "fusion coefficient" or "cluster diameter" is calculated, which is a numerical value at which cases merge into clusters. One heuristic approach is to plot the number of clusters against the fusion coefficient, and look for a marked flattening of the curve, which suggests that no new information is added by forming additional clusters (this process is analogous to the "screen test" of factor analysis). Using this procedure indicated three clusters. Another heuristic defines the flattening of the curve somewhat more objectively (Mojena, 1977). In this method, the fusion coefficient at a given step in the clustering must exceed a critical value based on the mean and standard deviation of the N-2 fusion coefficients (where N = the number of objects being clustered).

In both data sets, the clusters seemed to be defined by section proficiencies rather than by scale proficiencies. This can be seen when we compare cluster membership with scale, ignoring section, and compare cluster membership with section, ignoring scale. When cluster memberships were cross-tabulated

with scale proficiencies, scales straddled several clusters, as shown in Table 2 below, suggesting that the clusters are not homogeneous with regard to the language variables that were rated. Entries in Table 2 are the numbers of scores (out of the 18 that were used) that fell into one or another of the clusters. Thus the three in cluster 2 for the October pronunciation scale means that three scores derived from the reading items fell in cluster 2. For purposes of these tables, three clusters are shown.

Table 2

Number of Rating Scale Scores Falling into Each Cluster

	Cluster Membership					
	October			November		
Rating Scales	1	2	3	1	2	3
Pronunciation	1	3	1	1	2	2
Grammar	0	0	2	0	0	2
Fluency	1	3	1	1	3	1
Overall Comprehensibility	1	3	2	1	3	2

However, when cluster membership was cross-tabulated with section proficiencies, no straddling of clusters occurred in the October data, and only one straddle occurred in the November data, as can be seen in Table 3 below. "Straddling" occurs when a particular type of score falls into more than one cluster.

Table 3

Number of Test Section Scores Falling into Each Cluster

	Cluster Membership					
	October			November		
Test Sections	1	2	3	1	2	3
Reading Aloud	3	0	0	3	0	0
Sentence Completion	0	0	2	0	0	2
Picture Sequence	0	3	0	0	3	0
Single Picture	0	0	4	0	0	4
Free Response	0	3	0	0	2	1
Short Presentation	0	3	0	0	3	0

Note that the pattern of cluster memberships was identical for the two data sets (with one exception), whether the cross-tabulation was by rating scale or by section. This result suggests that the two samples had a similar structure.

Using the test sections to define the names of the clusters, and assuming three clusters, both samples yielded the following:

1. A cluster containing the Reading Aloud items.

2. A cluster made up of the Picture Sequence, Free Response, and Short Presentation items.

3. A cluster containing the Sentence Completion and Single Picture items.

If we accept the more conservative Mojena heuristic, then both samples contain just two clusters -- a Reading Aloud cluster, and everything else.

Thus, to the extent that these highly correlated scores can be said to have clustered into similar groupings, they did so more on the basis of similarity in the section, or item type, from which they were drawn, and less so because of similarity in the language variables being rated.

Stability Upon Repeated Ratings

For both exploratory factor analysis and multidimensional scaling, which is also exploratory, ratings on different scales were not reflected in the data structure. To test whether scale effects were nevertheless present, the following sub-study featuring rerating selected examinee tapes was conducted.

If there were indeed only one factor in TSE ratings as the exploratory factor analysis suggests, then a discrepancy between an examinee's scores is due to chance, regardless of the discrepancy's size; the order between any pair of scale scores would, on rerating, be repeated only 50 percent of the time. A significantly greater percent of agreement would indicate that something more than chance is operating. If the order of scales agreed significantly more than 50 percent of the time when rerating occurs, then the unifactor hypothesis is rejected.

A preliminary exploration of scale score differences revealed that the rarest number of large discrepancies occurred between the pronunciation and comprehensibility score scales, and between the fluency and comprehensibility score scales. A standard of .5 minimum absolute value for the discrepancy between pronunciation and comprehensibility, and between fluency and comprehensibility, was used to select cases for the study. This standard yielded 68 and 47 cases for October and November, respectively. All of these cases were used, so that a number of instances of the least numerous discrepancies, in addition to the more common discrepancies, would appear in the study. The tapes obtained at the TSE administration for these cases were rescored by different raters.

Before testing for agreement between original ratings and reratings, the scales were corrected by computing means and subtracting them from the individual scale ratings. This was done separately for the original ratings and the reratings. The correction was applied in order to avoid inflating agreement between comparisons of one scale with another simply because the scale scores differ in magnitude, on the average.

After the corrections were applied, agreement relative to average ratings was compared for original ratings and reratings. For example, when the pronunciation rating was being compared with the grammar rating an agreement was counted if the pronunciation rating, less the pronunciation mean, was greater than the grammar rating, less the grammar mean, for both the original and reratings, OR if the pronunciation rating, less the pronunciation mean, was smaller in both cases. The null hypotheses here is that agreement and disagreement are equally likely, hence the expected number of agreements is half the total number of cases. The results are given in Table 4.

Table 4

Chi-squared Tests of Proportion Agreement
Between Scale Comparisons of Raters' Scores

Scales Compared[a]	Proportion Agreements	Chi-Squared (1 d.f.)
	October (n=64)	
P&G	.70	10.6**
P&F	.78	20.2**
P&C	.73	14.1**
G&F	.77	18.1**
G&C	.52	.1
F&C	.75	16.0**
	November (n=46)	
P&G	.63	3.1
P&F	.72	8.7**
P&C	.67	5.6*
G&F	.59	1.4
G&C	.48	.1
F&C	.57	.8

[a] P=pronunciation, G=grammar, F=fluency, and C=Overall Comprehensibility.
* Significant at the 5% level of confidence.
** Significant at the 1% level of confidence.

Note in Table 4 that, there being more cases in October than November, the significance tests are more powerful in October. Table 4 tends to reject the single-factor result obtained from exploratory factor analysis, especially for the October data. Those scoring the October tapes differentiated consistently between pronunciation and fluency. Though they may have treated

grammar and overall comprehensibility as a single scale, they did consistently differentiate between that scale and both pronunciation and fluency. The November results are substantially weaker. It can be seen, however, that the exploratory factor analysis was not sensitive to trends that exist in data as revealed in Table 4.

Confirmatory Factor Analyses

Results of the rerating sub-study indicated that scale factors were indeed present in the rating data, despite their non-detection when exploratory factor analysis or multidimensional scaling procedures were used. Confirmatory factor analyses provided a more extensive examination of the effects that were present in the data.

Presence of Section and Scale Factors. Of the six scored sections, five are scored for pronunciation, two for grammar, five for fluency, and six for overall comprehensibility, which yields 18 variables. The correct associations of the 18 variables with scales and patterns are displayed in Table 5.

Table 5

Patterns of Non-Zero Factor[a] Loadings

P	G	F	C	2	3	4	5	6	7
X,	-,	-,	-,	X,	-,	-,	-,	-,	-,
X,	-,	-,	-,	-,	-,	X,	-,	-,	-,
X,	-,	-,	-,	-,	-,	-,	X,	-,	-,
X,	-,	-,	-,	-,	-,	-,	-,	X,	-,
X,	-,	-,	-,	-,	-,	-,	-,	-,	X,
-,	X,	-,	-,	-,	X,	-,	-,	-,	-,
-,	X,	-,	-,	-,	-,	-,	X,	-,	-,
-,	-,	X,	-,	X,	-,	-,	-,	-,	-,
-,	-,	X,	-,	-,	-,	X,	-,	-,	-,
-,	-,	X,	-,	-,	-,	-,	X,	-,	-,
-,	-,	X,	-,	-,	-,	-,	-,	X,	-,
-,	-,	X,	-,	-,	-,	-,	-,	-,	X,
-,	-,	-,	X,	X,	-,	-,	-,	-,	-,
-,	-,	-,	X,	-,	X,	-,	-,	-,	-,
-,	-,	-,	X,	-,	-,	X,	-,	-,	-,
-,	-,	-,	X,	-,	-,	-,	X,	-,	-,
-,	-,	-,	X,	-,	-,	-,	-,	X,	-,
-,	-,	-,	X,	-,	-,	-,	-,	-,	X

[a]Columns are for factors, with P,G,F, and C indicating the rating scales and 2-7 indicating the sections. Arabic numerals are used for section factors to accommodate the space available.

Rows in Table 5 index variables that are ordered by section within scale, which is the order in which they appear in the program statistical files; column entries refer to factors. Thus row 1 in the table refers to variable 1 in the program data files, which by hypothesis contains contributions from the pronunciation factor and the factor for section II. Entries in Table 5 indicate a particular model of TSE variables in that all variables (rows) with an X in the same column have non-zero loadings on the factor indexed by the column; all loadings with a hyphen in that same column are assigned the value of zero and hence ensure that the factor associated with that column does not contribute to that variable. Thus, the first column is the column for the pronunciation factor, the second is for grammar, the third is for fluency, and the fourth is for overall comprehensibility.

To see if the data would confirm as correct a pattern equivalent to that of the first 4 columns of Table 5, we fit a factor model consistent with that pattern and computed the least squares fit to the data. Then we drew 50, 4-factor patterns at random, fit factor models consistent with those patterns, and recorded for each pattern the least squares fit to the data. For each of these patterns the number of non-zero loadings in a column was the same as the number of Xs in one of the corresponding first four columns of Table 5 (5,2,5 and 6 for the four columns respectively). The correct pattern fit the data better than all 50 of the randomly selected patterns, an event that in random sortings of 51 events would occur less than 2 percent of the time. Thus, the data support our hypothesis of the presence of scale factors as given in the first four columns of Table 5. This result confirms the rerating results rather than those of the exploratory factor analysis, indicating that the scorers do indeed respond to the tapes in patterns consistent with our understanding of what the scales measure.

Subsequently, several sequences of pattern testing were carried out; one tested for the existence of section factors as a group and another tested for individual section factors. The test for existence of section factors as a group compared the sums of squares of residuals from a six factor pattern consistent with the association of sections with variables, with the sums of squares of 50 randomly drawn patterns with columns containing 3,2,3,4,3 and 3 Xs, where 3,2,3,4,3 and 3 are the numbers of scales on which the 6 sections are scored. For these tests of section variables a "correct" pattern is given by the last six columns of Table 5. Here again, the sum of squares of residuals was the smallest for the correct pattern, which rejects at the 2 percent level the hypothesis that its fit is one of a random sort of fits from randomly selected patterns. This result supports the inference that section factors are present.

Table 6

Patterns of Non-Zero Factor[a] Loadings

Ideal Pattern											Random Pattern										
P	G	F	C	2	3	4	5	6	7		P	G	F	C	2	3	4	5	6	7	
X	-	-	-	X	-	-	-	-	-		X	-	-	-	X	-	-	-	-	-	
X	-	-	-	-	X	-	-	-	-		-	X	-	-	-	-	X	-	-	-	
X	-	-	-	-	-	X	-	-	-		-	X	-	-	-	-	-	X	-	-	
X	-	-	-	-	-	-	X	-	-		X	-	-	-	-	-	-	-	X	-	
X	-	-	-	-	-	-	-	X	-		X	-	-	-	-	-	-	-	-	X	
-	X	-	-	-	X	-	-	-	-		X	-	-	-	-	X	-	-	-	-	
-	X	-	-	-	-	-	X	-	-		X	-	-	-	-	-	-	-	-	X	
-	-	X	-	X	-	-	-	-	-		-	-	X	-	X	-	-	-	-	-	
-	-	X	-	-	-	X	-	-	-		-	-	X	-	-	-	X	-	-	-	
-	-	X	-	-	-	-	X	-	-		-	-	X	-	-	-	-	X	-	-	
-	-	X	-	-	-	-	-	X	-		-	-	X	-	-	-	-	-	X	-	
-	-	X	-	-	-	-	-	-	X		-	-	X	-	-	-	-	-	-	X	
-	-	-	X	X	-	-	-	-	-		-	-	-	X	X	-	-	-	-	-	
-	-	-	X	-	X	-	-	-	-		-	-	-	X	-	X	-	-	-	-	
-	-	-	X	-	-	X	-	-	-		-	-	-	X	-	-	X	-	-	-	
-	-	-	X	-	-	-	X	-	-		-	-	-	X	-	-	-	X	-	-	
-	-	-	X	-	-	-	-	X	-		-	-	-	X	-	-	-	-	X	-	
-	-	-	X	-	-	-	-	-	X		-	-	-	X	-	-	-	-	-	X	

[a]Columns are for factors, with P,G,F, and C indicating the rating scales and 2-7 indicating the sections. Note reversal of left half and right half entries between columns one and two of rows 3,4,6 and 7; other entries are the same.

The right-hand section of Table 6 contains a possible pattern randomly selected for the test of pronunciation against grammar. For reference, the left-hand portion of Table 6 is the same as Table 5. For this contrast the numbers of Xs in columns 1 and 2 must total 5 and 2 as in Table 5, and the rows for which the Xs may be reassigned are 1 though 7. The randomly selected pattern in the right-hand portion of Table 6 has variables (rows) 2 and 3 associated with grammar, in contrast with the left-hand (correct) pattern where grammar is associated with variables 6 and 7. Variables 6 and 7 were reassigned to pronunciation by the randomization.

The first set of statistical tests of variables, two at a time, was concerned with scale factors. All possible pairs of scales taken two at a time were tested, with each test selecting 50 random patterns as before for each test, computing the least squares fit for each pattern and then comparing the resulting 50 figures with that obtained using the "correct" pattern. Thus some 300 factor models were fit in evaluating scale contrasts.

For all the contrasts involving the overall comprehensibility scale the correct pattern gave the best fit of the 50. The correct pattern also gave the best fit when

pronunciation and fluency were contrasted. But when grammar was contrasted with pronunciation or with fluency, some randomly selected patterns fit better than the correct pattern given in Table 5. Results of contrasting pronunciation with grammar, and fluency with grammar are given below.

Sections III and V are rated on grammar, and sections II, IV, V, VI and VII are rated on either pronunciation or grammar, producing seven variables rated on one or the other of these scales. The correct assignment of variables to scales is only one of 21 combinations of seven variables taken two at a time, and we computed the least squares fit for all 21 of these patterns. Thus, we intentionally mixed variables reflecting grammar ratings with variables reflecting ratings of pronunciations to see which of these mixes produced a least squares fit that was superior to the correct one. We found that, for both the October and November data, the following changes each resulted in a superior least squares fit:

- Treat the grammar ratings of Sections III and V as ratings of pronunciations, and the pronunciation ratings of Sections II and IV as ratings of grammar.
- Treat the grammar rating of Section III as a rating of pronunciation, and the pronunciation rating of section II as a grammar rating.
- Treat the grammar ratings of Sections III and V as ratings of pronunciations, and the pronunciation ratings of Sections IV and V as ratings of grammar.

As with pronunciation, sections II,IV, V, VI and VII are rated on fluency, and the correct assignment of variables to scales is only one of 21 combinations of seven variables taken two at a time. Again we computed the least squares fit for all 21 of these patterns. Thus, this time we intentionally mixed variables reflecting grammar ratings with variables reflecting ratings of fluency to see which of these mixes produced a least squares fit that was superior to the correct one. Again we found that, for both the October and November data, the following changes each resulted in a superior least squares fit:

- Treat the grammar rating of Section III as a rating of fluency, and the fluency rating of section IV as a grammar rating.
- Treat the grammar rating of Section V as a rating of fluency, and the fluency rating of section II as a grammar rating.

- Treat the grammar rating of Section V as a rating of fluency, and the fluency rating of section IV as a grammar rating.

- Treat the grammar rating of Section V as a rating of fluency, and the fluency rating of section VII as a grammar rating.

- Treat the grammar ratings of Sections III and V as ratings of fluency, and the fluency ratings of Sections II and IV as ratings of grammar.

- Treat the grammar ratings of Sections III and V as ratings of fluency, and the fluency ratings of Sections IV and VII as ratings of grammar.

Correlations between factor scores ranged around .99, with the signs of correlations between pairs of section factors all being positive, as were signs of correlation between pairs of scale factors. However, the signs of correlations of section factors with scale factors were all negative. All factor loadings were positive in sign. Multiple correlations of factors from the 10 factor solution with the 18 derived scores used in this analysis were computed and found to be quite substantial, yielding multiple correlation coefficients ranging from .92 through .99.

Summary and Conclusions

This project was undertaken to provide more recent evidence on the construct validity of the Test of Spoken English than was available. It investigated the dimensional structure of the test, emphasizing whether the empirical results were consistent with the intended structure.

TSE provides four scale scores based on six types of items arranged in sections of the test. The four scores are provided in order to reflect proficiency in different aspects of spoken performance, and different types of items are used for the same reason. Eighteen scores are derived by obtaining ratings of selected scales on the different sections, and one might expect both section and scale factors to operate. If all scale and section factors do indeed affect the derived scores, those eighteen scores should produce ten factors. However, by the standards of exploratory analysis, extracting ten factors from eighteen scores constitutes overfactoring to a considerable extent.

Exploratory factor analysis detected only one factor and multidimensional scaling, which is exploratory in nature, detected two or three. Hence, results from both exploratory methods failed to detect differences in the rating scales that

are consistent with our expectation that pronunciation, grammar, fluency, and overall comprehensibility reflect different aspects of speech. Subsequent analyses revealed that neither method detected all the factors present.

The confirmatory portion of this project was undertaken to explore the possibility that proficiency in the scale-defined behaviors and tasks are different but highly correlated. To the extent that this is so, results of exploratory analyses can mislead the test researcher who does not have access to examinees who differ in the full extent of capacities that the test can potentially measure. Indeed, in our data the correlations between factor scores were extremely high, being in the high nineties in absolute magnitude. This is why the single exploratory factor fits the correlations so well.

The rerating and confirmatory analyses established the existence of factors that were undetected by the exploratory analysis. It should be pointed out that there is a great difference between extracting ten exploratory factors from an eighteen variable matrix, which would be an example of gross overfactoring, and the ten factor extraction used here. A ten factor exploratory analysis requires estimating 180 factor loadings because in that analysis every factor can have a non-zero loading on every derived score. But in our confirmatory analysis only 36 non-zero loadings were allowed (see Table 5), and this is the number of loadings that would be estimated in a two-factor exploratory analysis.[1]

The confirmatory analyses supported the superiority of the allocation of factors to derived scores as displayed in Table 5, which is consistent with current notions of what the scores measure. In the exploratory factor analysis, however, factor patterns that treated grammar ratings as ratings of pronunciation or fluency sometimes provided a better fit to the data than did the correct pattern, as given in Table 5. Only Section VI, the free response section, failed to figure in such patterns, i. e., the intended role for scale ratings assigned to performances on section VI were always the correct roles.

The correlations between factor scores included a result we have not previously seen. That result is the negative correlations between scale factor scores and section factor

---

[1]Our analysis fits many more factor scores, however, so the comparison of two factor exploratory and our confirmatory analyses is not entirely accurate. We know of no existing rigorous method of counting parameters for the two types of analyses. Our analysis capitalizes on chance more than a two-factor exploratory analysis, and hence the significance tests were essential.

scores. Factor scores are constructs and, hence, are to some degree arbitrary. In particular, the choice of sign is arbitrary so long as the relationship between the sign of the factor scores and the sign of the factor loading are kept consistent with the signs obtained in the estimation. One expects to be able to set the signs so that the loadings are all positive and the factor scores are positively intercorrelated, or that there is a sensible interpretation available when the signs are reversed. The results of the present study are not congruent with this expectation. All of the loadings are positive, hence an increase in the factor score implies an increase in the estimated derived score. Any reflection of the variables, while allowable, produces a negative loading, which in turn indicates an increase in ability and implies a decrease in the estimated derived score. Thus, there are three choices available, any of which is consistent with the estimates: (1) the factor score intercorrelations are all positive, but an increase in scale factor scores implies a decrease in estimated derived score; (2) the factor score intercorrelations are all positive, but an increase in section factor scores implies a decrease in estimated derived score; or (3) an increase in any factor score implies an increase in estimated derived score, but those who score high on scale factors are likely to be low scorers on section factors, or the reverse. The type of result described here is novel and not fully understood. At the present time we are unable to use it as a basis for test design.

We must emphasize that, as with many dimensional studies, this is a survey and not a controlled experiment. Causal inferences are sometimes drawn from studies such as this, but this should not occur. For example, one cannot conclude that if grammar skills improve then section-specific skills will diminish. However, the high intercorrelation between skills suggests that there might be some transfer of training between skill areas. Research in that area could be helpful in developing optimal training methodologies for speaking skills.

Conclusions that can be drawn from the results of this study are given below with the sources of the evidence given in parentheses:

- Variations in item format affect the proficiencies required (confirmatory factor analysis, rerating study, multidimensional scaling)
- Distinctions between pronunciation, fluency, and overall comprehension ratings were consistently reflected in the rating scale data (confirmatory factor analysis, rerating study)

- Distinctions between ratings on the grammar scale and other scales were, for selected cases, (rerating study) reliably drawn.

- Status on the various factors that affected performance was highly consistent across factors (factor score correlations were very high in absolute magnitude).

Some implications of these results for the examinations are as follow: first, within the range of tasks that the TSE currently comprises there is a great deal of redundancy in what is measured. Note, for example, that variations in the degree of structure in the speaking tasks contained in TSE are not reflected in the section factor intercorrelations. It is possible that this redundancy extends to other tasks that might be considered for inclusion in the TSE, or that are not included in the TSE but that might be regarded as aspects of speaking proficiency. To the extent that it does extend, knowledge of TSE performance level provides a reasonably accurate expectation of status on these other tasks. Thus to be credible, a claim that another speaking task provides substantial information, independent of the information by TSE, needs confirmation by correlational evidence. It is important to note the limitations of the speaking domain, however. For example, one would not necessarily expect to find consistent proficiency status across oral tests of widely disparate areas of knowledge because TSE is not intended for specialized areas. Also, variations in the social context within which speech takes place might also affect relative proficiency status.

Second, as with the speaking tasks, status on the rated aspects of speech are also highly correlated. This result might obtain because the four skills rated were highly correlated, or because having the same rater rate all the skills from the same performance produced a halo effect. However, the rerating study results suggest that raters would differentiate between the skills if the differences were there. It is important to remember that detecting differences in what variables measure requires a research design with the power to detect the differences when they exist. The rerating study is one such design but, in the case of TSE, a traditional correlation study is not.

Third, even though the ratings reflect differences in the taped performances, one cannot automatically infer that it is useful to record scores on all the scales. If, for the vast majority of the tested population, status on one scale is very like status on another, then recording more than one scale might be counterproductive. This is especially so if obtaining several scales inflates the cost of testing, or if consistent and effective procedures for using information about scale discrepancies do not exist. This implication is timely, because

change in TSE reporting scales is under consideration, with future reporting scales possibly being limited to overall comprehensibility.

The results of this study support the expected structure of proficiencies measured by the TSE, including that the diagnostic scores--pronunciation, grammar, and fluency--are indeed diagnostic. It does not, however, suggest that one should look to TSE for frequent diagnostic applications. Instead, based on these results, one should expect to find that TSE provides information on overall proficiency status, and that if remedial treatments are to be provided, such treatments should be designed with the expectation that improvement will be needed in all the diagnostic areas.

References

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 No. 4). Princeton, NJ: Educational Testing Service.

Clark, J. L. & 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, 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.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. Psychomet, 29, 115-129.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules--An evaluation. Computer Journal, 20, 359-363.

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 No. 13). Princeton, NJ: Educational Testing Service.

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. Psychomet, 20, 93-111.

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.

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

Educational Testing Service (1992). Bulletin of information for TOEFL/TWE and TSE. Princeton, NJ: Author.

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.

Cover Printed on Recycled Paper
57906 • Y123M.5 • 275598 • Printed in U.S.A.