DOCUMENT RESUME

ED 263 120 TM 850 519

AUTHOR Bejar, Isaac I.
TITLE A Preliminary Study of Raters for the Test of Spoken English.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-85-5; TOEFL-RR-18
PUB DATE Feb 85
NOTE 44p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Covariance; Cost Effectiveness; *English (Second Language); Evaluation Criteria; Evaluation Methods; Evaluators; Factor Structure; Feasibility Studies; Higher Education; *Interrater Reliability; Language Tests; *Oral English; *Scoring; *Speech Tests; Standards; Test Reliability; Test Validity
IDENTIFIERS *Test of Spoken English

ABSTRACT

The feasibility of reducing scoring costs for the Test of Spoken English (TSE) by using one rater was investigated. Currently, two raters are used. It was found that, because of the possibility of different standards used by potential raters, it does not appear feasible to use a single rater as the sole determiner of speaking proficiency under the current system. Other possible alternatives were also examined. One approach was the development of a quality control index which would predict the extent of the disagreement between two raters, but the index that was developed could not be validated. The best predictors of rater disagreement were the identities of the raters. Their disagreements, however, resulted from the differing standards they used. Raters agreed substantially about the ordering of examinees, but varied slightly in the severity of their ratings. (Author/GDC)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

A Preliminary Study of Raters for the Test of Spoken English

Isaac I. Bejar

EDUCATIONAL TESTING SERVICE

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 over 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 TOEFL 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. Local research may sometimes require access to TOEFL data. In such cases, the program may provide this data following approval by the Research Committee. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1984-85) members of the TOEFL Research Committee include the following:

Henry F. Holtzclaw, Jr. (chair)	University of Nebraska
Kathleen M. Bailey	Monterrey Institute of International Studies
Alison d' Anglejan-Chatillon	University of Montreal
H. Douglas Brown	San Francisco State University
Russell N. Campbell	University of California at Los Angeles
John Haskell	Temple University-Japan

A Preliminary Study of Raters for the Test of Spoken English

Isaac I. Bejar

Educational Testing Service
Princeton, New Jersey

RR-85-5

Abstract

The investigation was undertaken to provide information about the feasibility of reducing scoring costs by using one rater instead of the two that are now used for the TSE. It was concluded that because of the possibility of different standards among potential raters, it does not appear feasible to use a single rater as the sole determiner of speaking proficiency under the current system. Other possible alternatives to a single rating, relying on psychometric methodology and technology, are discussed. The approach was to first examine the possibility of developing a "quality control" index that would predict the extent of the disagreement between two raters. The index that was developed for this purpose could not be validated. It was found that the best predictors of rater disagreement were the identities of the raters. The disagreements, however, resulted from the differing standards used by different raters. That is, raters agree substantially about the ordering of examinees but vary slightly in the severity of their ratings.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Many individuals contributed generously to this project. I'd like to express my appreciation to Gary Driscoll for skillfully and promptly coordinating the data analyses, Spencer Swinton and Don Powers for reviewing an earlier draft, Rod Ballard for providing the data and many details about the program, Elsa Rosenthal for editing the manuscript, Elaine Guennel for the preparation of the final report, the TOEFL Research Committee for providing the necessary financial support, and Charles Stansfield for facilitating everything.

A Preliminary Study of Raters for the Test of Spoken English

Isaac I. Bejar

There is a growing consensus that speaking proficiency is best measured by evaluating directly the individual's speaking skills (Powers and Stansfield, 1983). The Interagency Language Round Table Oral Proficiency Interview is a well-known procedure that exemplifies this approach; the Test of Spoken English (TSE) developed at Educational Testing Service (ETS) is another. An important feature of each of these measures is that the score is derived solely from the ratings provided by language specialists. This means that a contingent of trained raters must be available to the program. One characteristic of testing programs that rely on raters is the high cost of "scoring" the tests. The proportion of the total budget that rating costs represent naturally decreases only slightly as volume increases but remains quite high since the cost of rating each examinee remains constant. By contrast, in a testing program that uses "objective" assessment procedures, the cost associated with scoring declines sharply as volume increases. This no doubt explains to some extent the preponderance of objective assessment procedures. Nevertheless, in domains such as speaking proficiency, ratings are the most appropriate and feasible measurement approach. Thus, it is important that cost-effective procedures be found to obtain valid measurement.

Overview of the Study

This investigation sought to provide information that could guide a decision on ways of reducing TSE scoring costs. The approach taken was to investigate the possibility of using one rater instead of the two currently used. Some previous research (Bolus, Hinofotis, and Bailey, 1982) has demonstrated that a single rater can in fact yield sufficiently adequate measures of proficiency. If this proved to be the case for the TSE, it should be possible to significantly reduce the costs of the program.

Because the records for the program have not been "computerized," our first task was to create a data base. (Care was taken to document the data base carefully since it will facilitate future analysis for either administrative or research purposes.) We then examined the measurement characteristics of the existing rating procedures, focusing on the possibility that disagreement among raters could be predicted statistically. Our rationale for this approach was that even if a single rater proved sufficiently accurate, we would still need a quality-control mechanism to identify instances in which there would have be a large discrepancy if another rater were involved. We then focused on the characteristics of individual raters to determine whether raters tend to apply similar standards.

Description of the Test

The TSE consists of seven sections designed to elicit different speech acts. The first section is a warm-up and is not scored. The composition of the test is presented in Table 1.

Table 1

Contents of the TSE

Section	Description
1	Warm-up consisting of questions concerning examinee background characteristics
2	Examinee is given a passage to read aloud.
3	Examinee completes 10 partial sentences in a meaningful manner.
4	Examinee tells a story about a series of related pictures.
5	Examinee responds to a series of questions posed by the examiner concerning a drawing.
6	Examinee is expected to provide lengthy responses about topics with which he or she is familiar.
7	Examinee sees a printed schedule and describes it aloud.

The test can be administered on an individual basis or to groups using a language laboratory. The test questions and response stimuli appear in the printed test book or are heard by the examinee on the test tape. Examinee responses are recorded on a separate tape that is sent to ETS.

Scoring Procedures

Performance on the TSE is evaluated by two raters, both randomly assigned from a pool of about twenty raters who have a background in language teaching and testing and have attended a one-day rater training workshop at ETS. Raters evaluate examinees' performance along four linguistic dimensions. Three of these--grammar, fluency, and pronunciation--are considered diagnostic scores; the fourth dimension, comprehensibility, is considered to be integrative.

To facilitate discussion of the different variables, we will adopt the following convention: the first and second rating of each examinee will be denoted as rating A and rating B, respectively. (It should be pointed out that these are arbitrary labels and that each rater has an equal chance of contributing a B or an A rating. That is, ratings A and B should be viewed for practical purposes as replicates of each other.) The four linguistic dimensions will be denoted by "Pron" for pronunciation, "Gram" for grammar, "Flue" for fluency, and "Comp" for comprehensibility. Finally, sections 2 through 7 will be denoted by the corresponding numerals.

Each examinee obtains two sets of ratings. To refer to scores obtained under either the first or the second rating, the variable name is preceeded with either A or B. The number of sections and items composing each dimension is shown in Table 2. The item score in each case ranges from 0 to 3; that is, each of the linguistic dimensions is rated on a four-category scale. For sections that contain more than one item, the mean rating across items is the score for that section. For example, Section 3 consists of ten items, each rated for grammar and comprehensibility. The scores on the ten items are averaged, and the mean score becomes the score for the section.

An overall score on each of the four dimensions is obtained by averaging across the section scores. For example, the overall score for grammar is the average of Gram3 and Gram5. The result is a set of four overall scores for each examinee from each rater. For score reporting purposes, the two sets are averaged; the average comprehensibility score is considered the score for reporting purposes.

Table 2

Sections that Contribute to TSE Scores and Number of Items per Section

Section	Pronunciation	Grammar	Fluency	Comprehensibility	Number of items
2	Pron2		Flu2	Comp2	1
3		Gram3		Comp3	10
4	Pron4		Flu4	Comp4	1
5	Pron5	Gram5	Flu5	Comp5	4
6	Pron6		Flu6	Comp6	3
7	Pron7		Flu7	Comp7	1
Overall	Pron	Gram	Flu	Comp	

If the two raters differ by more than .95 at the overall score level on any linguistic dimensions, a third rater is brought in. The third rating is averaged with the other two.

Description of the Data Base

The data base used in the study consisted of all available protocols from November 1981, the first official administration of the TSE, to June 1983. Altogether there were 560 examinees in the database, each rated by at least two raters. Table 3 shows some descriptive statistics for each of the scores under ratings A and B.

Table 3

Mean, Standard Deviation, Median, and Interquartile Range (IQR) of Ratings A and B (N = 560)

	Rating A				Rating B			
	Pron	Flu	Gram	Comp	Pron	Flu	Gram	Comp
Mean	1.98	2.12	2.32	2.11	2.00	2.14	2.30	2.14
S.D.	.60	.55	.50	.53	.62	.57	.51	.54
Med	2.00	2.05	2.40	2.07	2.00	2.07	2.38	2.10
IQR	.80	.70	.65	.62	.88	.74	.68	.70

Figures 1-4 show the distribution for rating A and rating B on each score. It is apparent that the distributions are quite similar, as would be expected given the fact that whether a rater provides a rating A or B is determined basically at random.

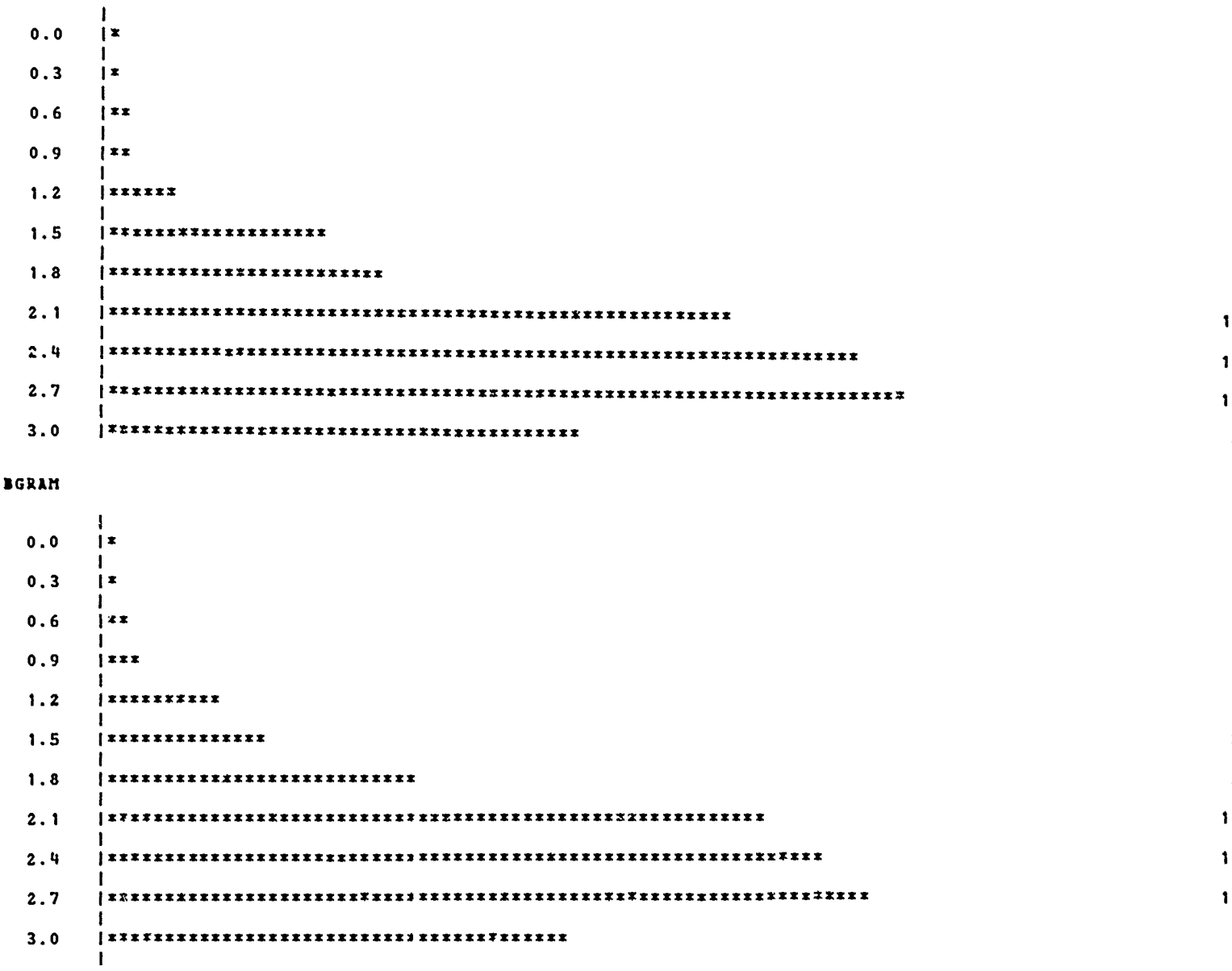
Differences Between Raters

For the purposes of this investigation it is important to characterize the differences among the ratings since, as indicated above, it is those differences that determine whether a third rater is used. Table 4 shows descriptive statistics on the differences for the four scores. The distribution of the differences tends toward normality, with the mean and median close to zero in all cases. However, the variability of the differences for pronunciation and fluency scores is markedly greater than the variability of the differences for grammar and comprehensibility.

These findings are reassuring. When differences among the raters are pure error, the distribution is precisely normal with a mean of zero. Since the means are very near zero, one can assume the differences are not systematic.

Figure 1

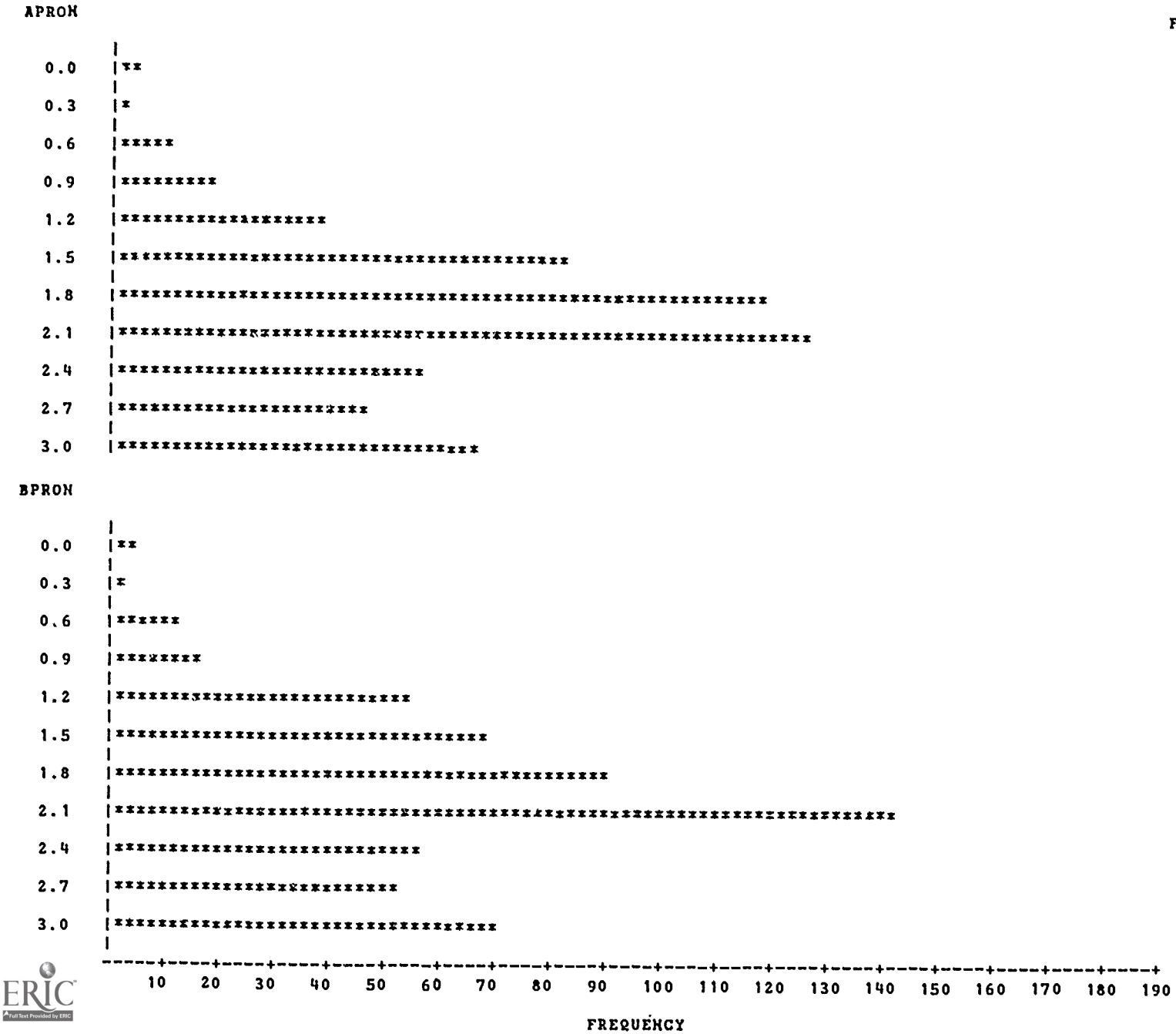
Grammar Ratings Assigned to 560 Examinees
Under Ratings A and B



	FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
AGRAM				
0.0	1	1	0.18	0.18
0.3	2	3	0.36	0.54
0.6	3	6	0.54	1.07
0.9	4	10	0.71	1.79
1.2	11	21	1.96	3.75
1.5	37	58	6.61	10.36
1.8	48	106	8.57	18.93
2.1	107	213	19.11	38.04
2.4	129	342	23.04	61.07
2.7	137	479	24.46	85.54
3.0	81	560	14.46	100.00
BGRAM				
0.0	1	1	0.18	0.18
0.3	2	3	0.36	0.54
0.6	3	6	0.54	1.07
0.9	6	12	1.07	2.14
1.2	19	31	3.39	5.54
1.5	27	58	4.82	10.36
1.8	54	112	9.64	20.00
2.1	113	225	20.18	40.18
2.4	123	348	21.96	62.14
2.7	132	480	23.57	85.71
3.0	80	560	14.29	100.00

Figure 2

Pronunciation Ratings Assigned to 560 Examinees
Under Ratings A and B



	FREQ	CUM. FREQ	PERCENT	CUM. PERCENT
APRON				
0.0	4	4	0.71	0.71
0.3	1	5	0.18	0.89
0.6	10	15	1.79	2.68
0.9	17	32	3.04	5.71
1.2	37	69	6.61	12.32
1.5	82	151	14.64	26.96
1.8	118	269	21.07	48.04
2.1	125	394	22.32	70.36
2.4	56	450	10.00	80.36
2.7	45	495	8.04	88.39
3.0	65	560	11.61	100.00
BPRON				
0.0	4	4	0.71	0.71
0.3	1	5	0.18	0.89
0.6	12	17	2.14	3.04
0.9	15	32	2.68	5.71
1.2	53	85	9.46	15.18
1.5	67	152	11.96	27.14
1.8	89	241	15.89	43.04
2.1	142	383	25.36	68.39
2.4	56	439	10.00	78.39
2.7	52	491	9.29	87.68
3.0	69	560	12.32	100.00

Figure 3

Fluency Ratings Assigned to 560 Examinees
Under Ratings A and B

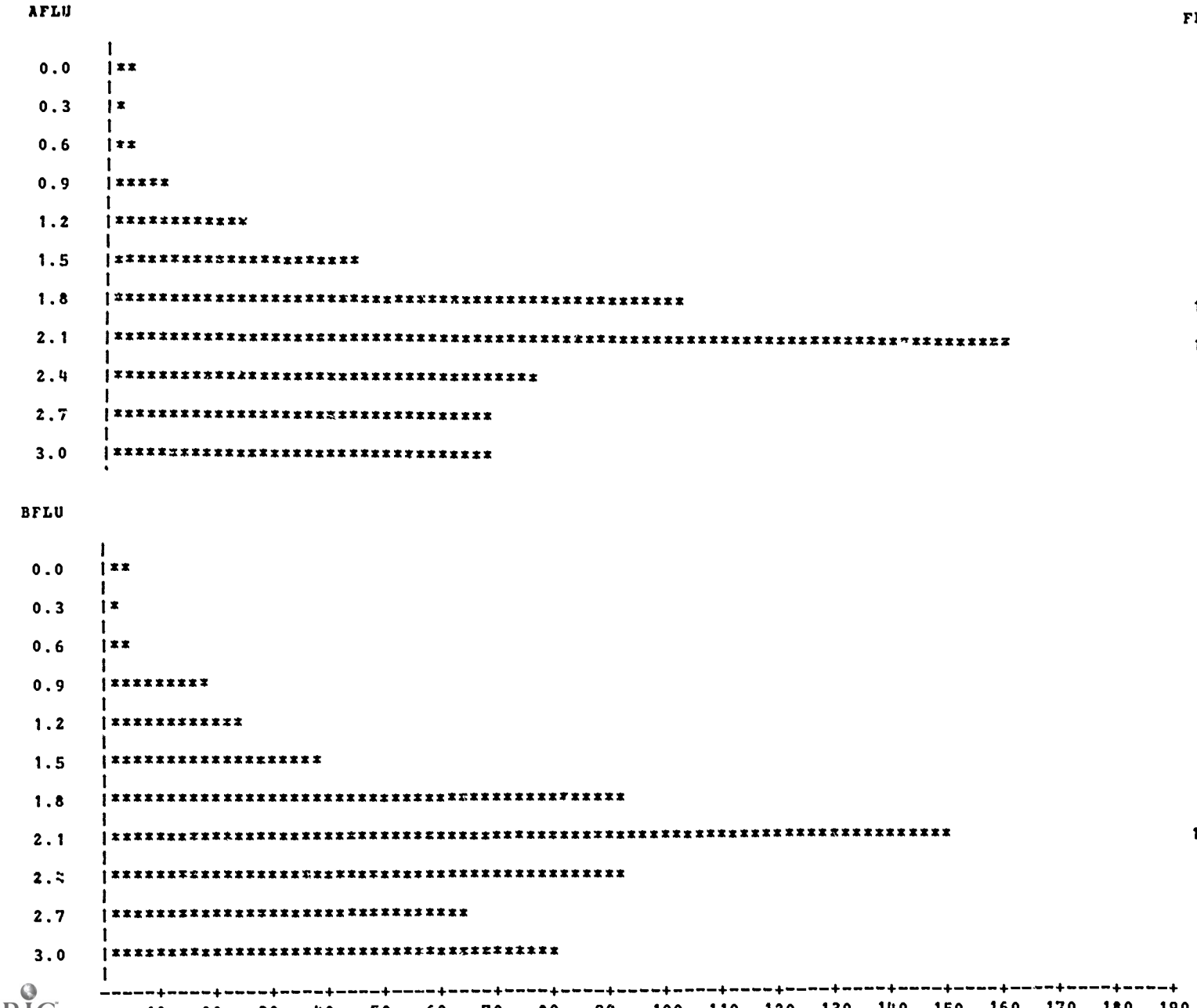


Figure 4

Comprehensibility Ratings Assigned to 560 Examinees
Under Ratings A and B

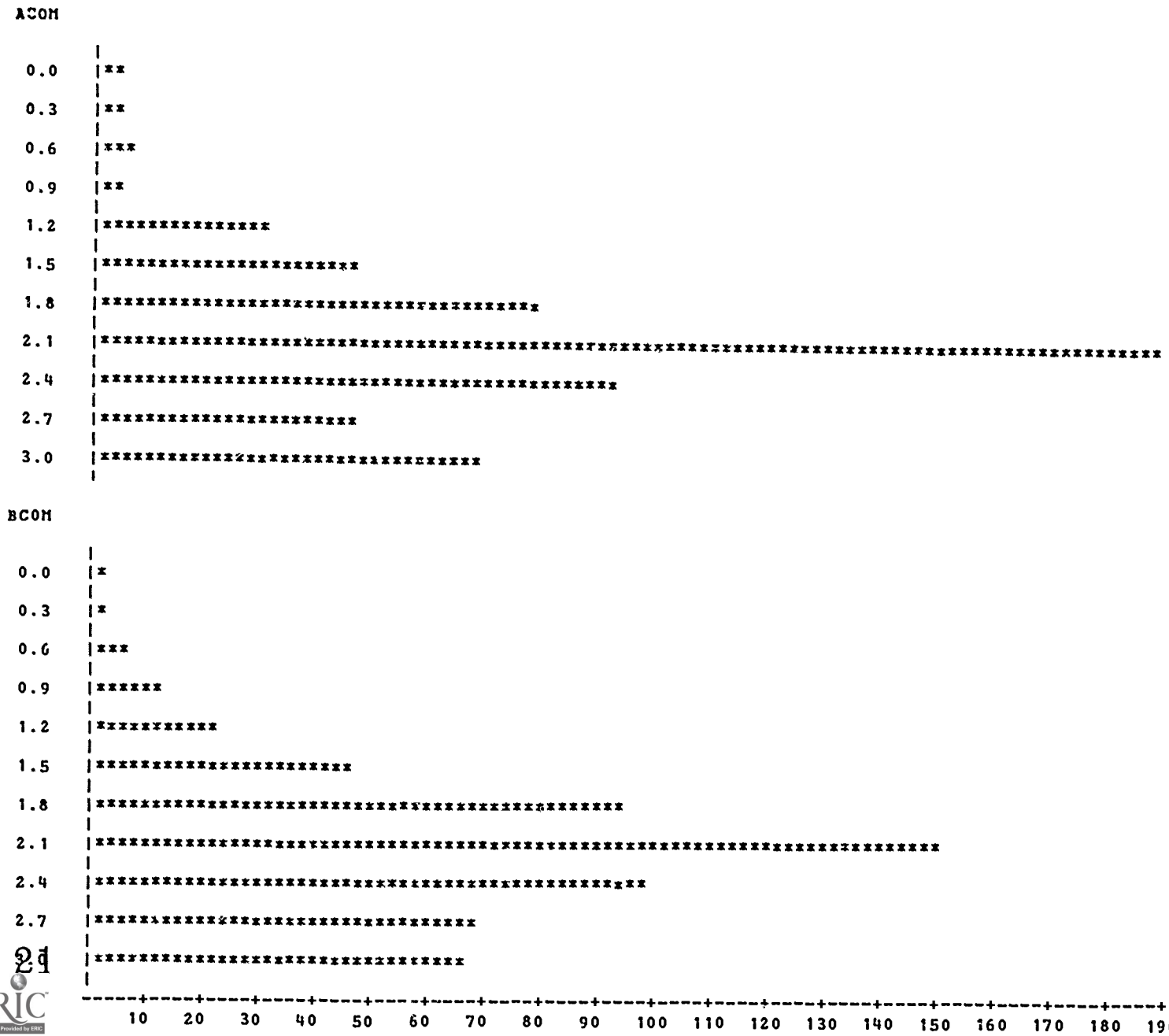


Table 4

Descriptive Statistics of Differences Between A and B Ratings on Four Linguistic Skills (N = 560)

Statistics	Pron	Flu	Gram	Comp
Mean	-.02	-.02	.02	-.03
Standard Dev.	.44	.45	.30	.38
Med	0	0	0	-.01
IQR	.60	.52	.35	.45
99th percentile	1.01	1.01	.71	.88
95th percentile	.78	.80	.50	.66
90th percentile	.60	.60	.42	.45
10th percentile	-.60	-.63	-.34	-.52
5th percentile	-.73	-.80	-.48	-.65
1st percentile	-.93	-1.09	-.80	-.91

Note: The interquartile range (IQR) is the range between the 25th and 75th percentiles.

Development of a Quality Control Index

As indicated earlier, a major objective of this investigation was to develop and validate an index that could be used to predict disagreement among raters. The fact that up to this point examinees have been rated by two raters allows us to validate such an index. However, to be useful, the index should work in such a way that it could be computed on the ratings provided by a single rater. One hypothesis investigated was that the language background of an examinee could be implicated In large disagreements between raters. Examination of the data, however, did not indicate that the frequency of rater disagreement was associated with language background. This left two possibilities for predicting disagreement: the identity of the rater and some rater-by-ratee interaction. We will first examine the latter possibility.

The rationale of the procedure to detect rater-by-ratee interactions is to investigate the underlying statistical model that accounts for the covariation among the ratings. Deviations from that underlying model are taken to be a possible indication of something unusual about the rater-ratee observation. The model that was postulated was a dimensionality model. Specifically, it was postulated that the ratings would be unidimensional. That is, that a single underlying variable would account for the covariation among the ratings. Such a model provides the simplest starting point even though it is inconsistent with the current view that linguistic skills are not based on a unitary factor. (See Oller, 1983; Vollmer & Sang 1983.) (It is beyond the scope of this report to discuss the dimensionality of linguistic skills. As we will see below, a

unidimensional model seems adequate to statistically account for the covariation among linguistic skills as perceived by raters, but this does not preclude the possibility that psychologically more than one factor is needed to account for linguistic performance.)

Dimensionality was investigated by means of factor analysis. Table 5 shows the correlations among the four scores for the A and B ratings. As can be seen, the intercorrelations are nearly identical.

Table 5

Intercorrelation Among the Four Linguistic Skills
(Rating A Below the Diagonal, Rating B above)

	Pron	Gram	Flu	Comp
Pron	-	.724	.829	.907
Gram	.726	-	.746	.797
Flu	.821	.746	-	.884
Comp	.903	.798	.877	-

To examine dimensionality as such, we factor analyzed the matrices in Table 5. It should be noted that in doing so we ignored the covariation among raters embedded in these matrices. That is, the covariation among scores could be partitioned into a "between raters" component and a "within raters" component. For this analysis we implicitly assumed that the between-raters component was nil and that each rater was, in fact, unidimensional. That is, it was conceivable that by collapsing across raters we might create unidimensionality artifactually. Thus, the dimensionality of each rater was also analyzed. The results appear in Appendix A. It was found that while the fit of a single dimension was not equally good across raters, a single dimension was the most reasonable model. This does not guarantee that the same dimension is present in each rater, but the magnitude of the correlation between the raters points in that direction, as we shall see in Table 11.

The dimensionality of ratings A and B data was investigated by means of maximum likelihood factor analysis. A maximum likelihood estimation process is statistically most efficient and provides a measure of statistical goodness of fit, provided certain distributional assumptions are met. (Computations were performed using the SAS statistical package.) A single factor was fitted to each matrix. The results of the factor analysis, including a statistical measure of fit, appear in Table 6.

Table 6

Results of the Maximum Likelihood Factor Analysis
Extracting a Single Factor

	Rating A		Rating B	
	Loadings	Mean-sq Residuals	Loadings	Mean-sq Residuals
Pron	.917	.01	.920	.01
Gram	.812	.02	.808	.02
Flu	.894	.01	.899	.01
Comp	.983	.00	.985	.00
	Chi 7.41		Chi 7.55	
	df 2		df 2	
	p .025		p .023	

The results strongly suggest that the ratings are in fact unidimensional. The probability of the null hypothesis of a single factor is small but cannot be entirely relied upon since the data do not follow a multivariate normal distribution. More importantly, the root mean squares of the off-diagonal residuals do not show a pattern. As Table 6 shows, the magnitude of the residuals is small. In other words, it is possible to recover the original correlation matrix fairly well with the estimated loadings on a single factor. It is also worth noting that the results of the factor analyses for ratings A and B are quite similar. That is, the largest contributor to the factor is comprehensibility, followed by pronunciation, fluency, and grammar in that order.

Reliability. Having estimated the parameters of a single factor model, we could estimate the internal consistency of rating A and rating B data. This was done by following Maxwell's (1971) formulation for estimating reliabilities of composite scores. In the present case, the composite consisted of the four scores. When there is a single factor in the data, reliability is given by the following formula:

$$r = \frac{\sum_i \frac{\lambda_i^2}{1 + \lambda_i^2}}{1 + \sum_i \frac{\lambda_i^2}{1 - \lambda_i^2}} \qquad (1)$$

where λ is the loading of the ith score on the single factor. The estimate can be obtained by using the estimates of λ given in Table 6. For rating A the estimated internal consistency reliability across four scores was .976; it was .978 for rating B. These estimates suggest a high degree of consistency in the ratings but are different from estimates of interrater reliability.

Relationship between ratings A and B. Since the data collected up to this point were for two raters, we could perform an analysis shedding further light on the nature of the ratings. Specifically, we could conduct an analysis similar to the one above but incorporating both sets of ratings for each examinee. The resulting correlation matrix between ratings A and B appears in Table 7. The interrater correlations on the four scores appear in parentheses. The correlation is highest for grammar and lowest for fluency.

Table 7

Intercorrelation Matrix for Ratings A and B, Including the Factor Loadings and Residuals for a One-Factor Model

	APRON	AGRAM	AFLU	ACOMP	BPRON	BGRAM	BFLU	Loading	Residual
APRON	-							.873	.070
AGRAM	.726	-						.838	.064
AFLU	.821	.746	-					.837	.079
ACOM	.903	.798	.877	-				.901	.081
BPRON	(.741)	.673	.668	.740	-			.881	.067
BGRAM	.679	(.827)	.637	.726	.724	-		.841	.061
BFLU	.662	.653	(.669)	.680	.829	.746	-	.849	.075
BCOM	.735	.716	.670	(.753)	.907	.797	.884	.906	.078

A single factor was extracted from this correlation matrix. The loadings on that single factor appear at the extreme right of the table. From a statistical point of view, however, a single factor was not sufficient to account for the correlations, as was evidenced by the highly significant chi-square statistic (chi-square = 1196.97, df = 20, $p < .0001$). More important, the residual off-diagonal correlations were substantial. The root mean off-diagonal residuals are shown in the far right hand column of Table 7. Moreover, there was a specific pattern to those residuals. Specifically, the largest residuals were ACOM-APRON, ACON-AFLU, BCOM-BPRON, and BCOM-BFLU. One possible interpretation of this pattern is that, although for the most part the two raters shared the same perspective about proficiency, they seemed to differ somewhat with respect to how they integrated pronunciation and fluency into the comprehensibility rating. Indeeed, the variability of the differences for the pronunciation and fluency ratings (see Table 4) is larger than it is for the grammar and comprehensibility ratings.

Deviations from unidimensionality. The data presented thus far suggest that unidimensionality is a reasonable model of speaking proficiency as evaluated by raters. Therefore, an index that examines deviation from unidimensionality was investigated as a way of predicting instances in which an examinee would be rated discrepantly by two raters. The index we used was suggested by Gnanadesikan (1977) in a different context, namely, the identification of multivariate outliers. The rationale of the index can best be seen in the bivariate case. Figure 5 shows a scatter plot for two variables. The first principal component is the line that minimizes the perpendicular distance of each point to this line: the second principal component is error. Being high on that component is indicative of a peculiarity. For example, suppose the two variables under consideration are height and weight. The first principal component would be a linear combination of these two variables. If we find subjects high on the second principal component, chances are that they would be unusually heavy or light for their height.

The applicability of this rationale to the present application is justified by the fact that speaking proficiency seems to be unidimensional. Deviations from unidimensionality could thus be viewed as evidence of a rater-by-ratee interaction. If that peculiarity can predict discrepancy between two raters, we might be justified in using it in a single rater system as a quality control mechanism.

The formula for the peculiarity index is given by

$$d^2 = \sum_{j=p-q+1}^{p} [a'_j(y_i - y)]^2 \qquad (2)$$

where

y_i is a vector of ratings for the ith examinee, which in this case consists of four scores.

y is the mean vector of ratings over all examinees

a'_j is the jth principal component.

p is the number of variables, four in this case,

q refers to the last q principal components

Figure 5

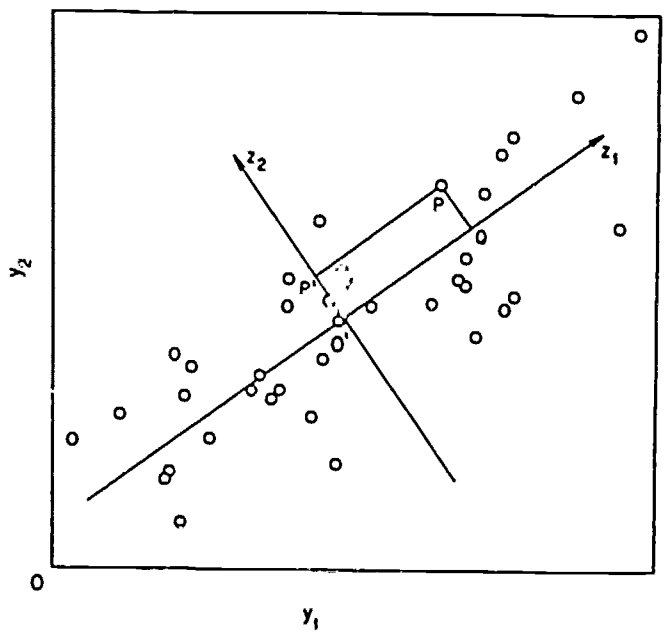


Illustration of principal components residuals.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations.

Several indices can be computed from this formula. Since we have presented evidence that unidimensionality is a reasonable model, we computed the index setting q to 3. That is, the second, third, and fourth principal components were taken to be error. In this form the index quantifies, under the assumption of unidimensionality, how dissimilar an examinee's ratings are from the mean rating obtained by all examinees.

To compute the index it is first necessary to compute the principal components. The four principal components for ratings A and B are given in Table 8.

Table 8

Principal Components for the Covariance Matrix for Rating A and B (N = 560)

	Rating A				Rating B			
	1	2	3	4	1	2	3	4
PRON	.558	-.495	-.523	-.411	.564	-.483	-.537	-.400
GRAM	.424	.860	-.231	-.165	.417	.866	-.218	-.164
FLU	.502	-.079	.819	-.266	.506	-.089	.814	-.271
COMP	.506	-.096	-.041	.856	.501	-.088	-.035	.860

The standing of each examinee on these three indices was computed by means of equation 1. A roster was prepared containing the index for each ratee as well as the ratings provided by two raters and the corresponding difference. It quickly became apparent that the magnitude of the differences between two raters could not be predicted by the index. Indeed, the correlation of the index with the absolute difference between the two raters on any of the linguistic skills was no larger than .15. In short, it appears that an approach based on a discrepancy index such as the one proposed by Gnanadesikan (1978) does not appear useful as a means of predicting disagreement between two raters.

Analysis of Raters

The second possibility we investigated for predicting rater disagreement focused on the individual raters. Table 9 shows the number of examinees a given rater was assigned and the mean rating of those examinees. The table also shows the mean rating for the same examinees given by the raters with whom a given rater had been paired. This information gives us an indication of the degree of severity applied by each rater. Note, however, that since there is no guarantee that examinees are assigned at random to raters, the mean rating awarded by a specific rater is not necessarily the best indicator of that rater's severity. The contrast with the mean rating provided by the other raters for the same examinees is a better indication of whether a rater has a tendency to overrate or underrate examinees.

Table 9

Means for Each Rater and the Paired Raters on Four Linguistic Dimensions

	Means for each rater				Means for paired raters			
Rater ID.	Pron.	Gram.	Flu.	Comp.	Pron.	Gram.	Flu.	Comp.
111	1.88	2.31	1.93	1.94	1.88	2.31	2.18	2.15
113	1.92	2.48	2.17	2.04	2.07	2.37	2.26	2.20
114	2.07	2.18	2.29	2.20	1.84	2.14	1.99	1.96
118	2.23	2.40	2.41	2.31	1.93	2.34	2.08	2.04
120	1.75	2.13	1.71	1.87	1.98	2.33	2.14	2.15
121	1.81	2.42	2.17	2.05	2.00	2.37	2.12	2.14
124	1.92	2.37	2.05	2.12	1.95	2.30	2.05	2.07
125	1.77	2.37	1.84	1.96	2.02	2.31	1.82	2.16
126	2.07	2.26	2.07	2.26	2.18	2.31	2.04	2.20
127	2.38	2.32	2.12	2.27	2.03	2.34	2.08	2.23
128	2.10	2.21	2.24	2.21	1.96	2.19	2.10	2.13
129	2.19	2.28	2.38	2.36	2.00	2.31	2.18	2.09
130	1.58	2.08	1.66	1.77	1.88	2.17	2.09	2.17
135	1.97	2.21	2.19	2.22	2.10	2.24	2.25	2.20

Table 9 clearly shows that some raters tended to give lower ratings, and they did so consistently across all four scores. Table 9 also shows that of all raters, raters 120 and 130 were the most severe. If an examinee were to be assigned to two raters who tend to underrate, it is probable that the examinee would receive a lower rating than if assigned to a different pair of raters. This also extends to the case where there is just one rater. An examinee assigned to a severe rater might receive a lower rating than some other rater might give.

It should be remembered that Table 9 depicts the distribution of TSE raters according to their severity. As in any distribution some individuals fall below the mean and some will fall above. The data in Table 9 show that raters 114, 118, and 129 were more generous than their colleagues. An examinee assigned to two lenient raters might receive a higher score than another pair of raters might give. However, the practical effect an assignment to two similarly disparate raters is not large in terms of scaled score points, and the probability of an examinee being assigned to similarly disparate raters is quite low.

The difference between a rater and his or her paired raters can easily be computed from Table 9. Figures 6-9 show the difference for pronunciation to comprehensibility, respectively, for each rater. A negative difference indicates that the paired raters gave the same examinees a higher rating. Again we see that raters 120 and 130 tended to give the lowest ratings. For comprehensibility, rater 120 tended to rate examinees lower by .28 scale points (1.87 vs. 2.15)--which is about half a standard deviation with respect to the pooled-within-rater variability of the comprehensibility rating. Similarly, for comprehensibility, raters 118 and 129 tended to rate examinees higher by 27 scale points. We also note that the smallest differences among raters are on grammar and the largest are on fluency.

With an indication of the severity of each rater, we were able to examine again the distribution of differences to see if specific raters tended to exhibit large differences more frequently. Specifically, we examined instances where the difference between raters was more than .95, the difference that triggers a third rating under the current system. The results appear in Table 10.

Out of 560 examinees, there were 32 instances of discrepancies greater than or equal to .95. Rater 120 was involved most frequently in discrepancy cases followed by raters 111, 114, 121, and 129, with counts of about 10 each.

Table 10 shows that the largest number of discrepancies greater than .95 occurred in the criterion of fluency. Of the 32 examinees involved in such discrepancies, the rating assigned to fluency was at issue in 22 of them. Having noticed that the largest number of discrepancies occurred on this criterion, TSE program staff revised the descriptive statement given raters that accompany each point on the fluency scale in November 1983. Subsequently, staff report a very considerable reduction in the number of discrepancies involving fluency. Since this study includes data produced as of June 1983, ratings given after this refinement of the fluency scale were not analyzed here.

Prior to November 1983 ratings were assigned by ESL professionals living in or near Princeton, New Jersey. However in November 1983 TSE program staff decided to utilize as raters graduate students pursuing a masters or doctorate in teaching English as a second language at the University of Delaware. TSE program staff report about a two-thirds decrease in the number of discrepancies with this new group of raters. It is believed that this improvement is due to the fact that members of this group share a common academic background (in terms of core courses), in addition to their TSE training, and because the members of the group are in almost daily contact with each other. Again, this more recent data was not included in this study. However, once it is analyzed, it may result in further gains in interrater agreement.

Figure 6

Mean Difference Between Each Rater
and the Paired Raters on the Pronunciation Score

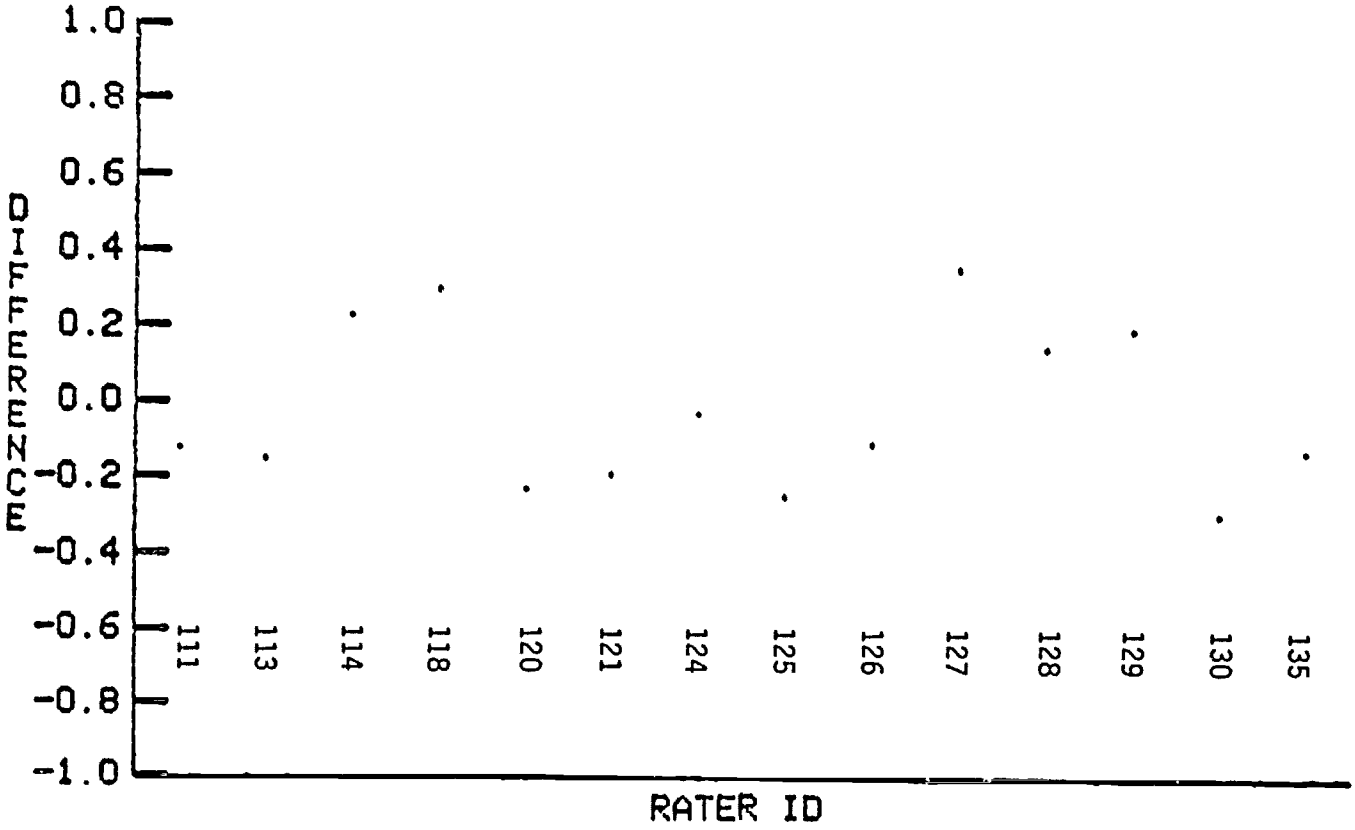


Figure 7

Mean Difference Between Each Rater
and the Paired Rater on the Grammar Score

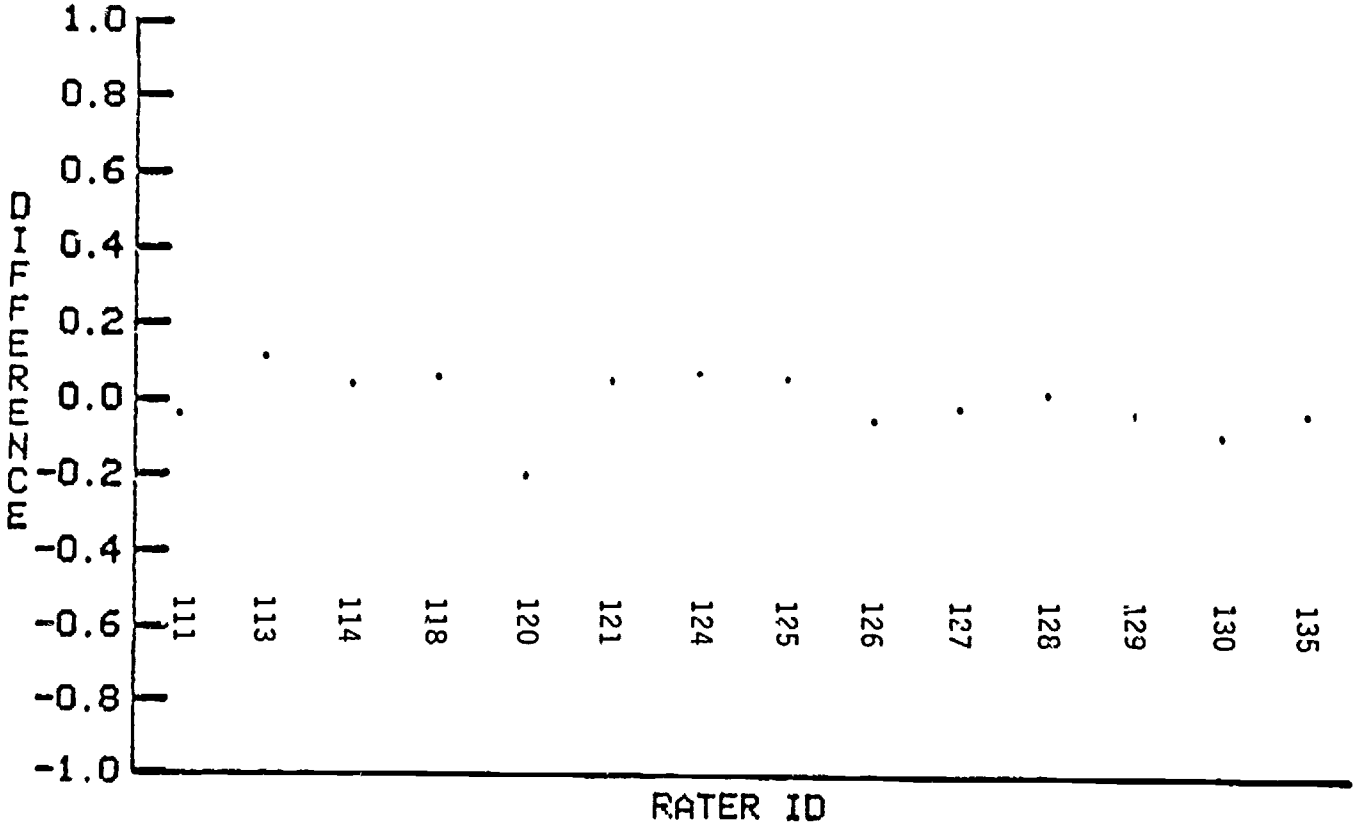


Figure 8

Mean difference Between Each Rater
and the Paired Raters on the Fluency Score

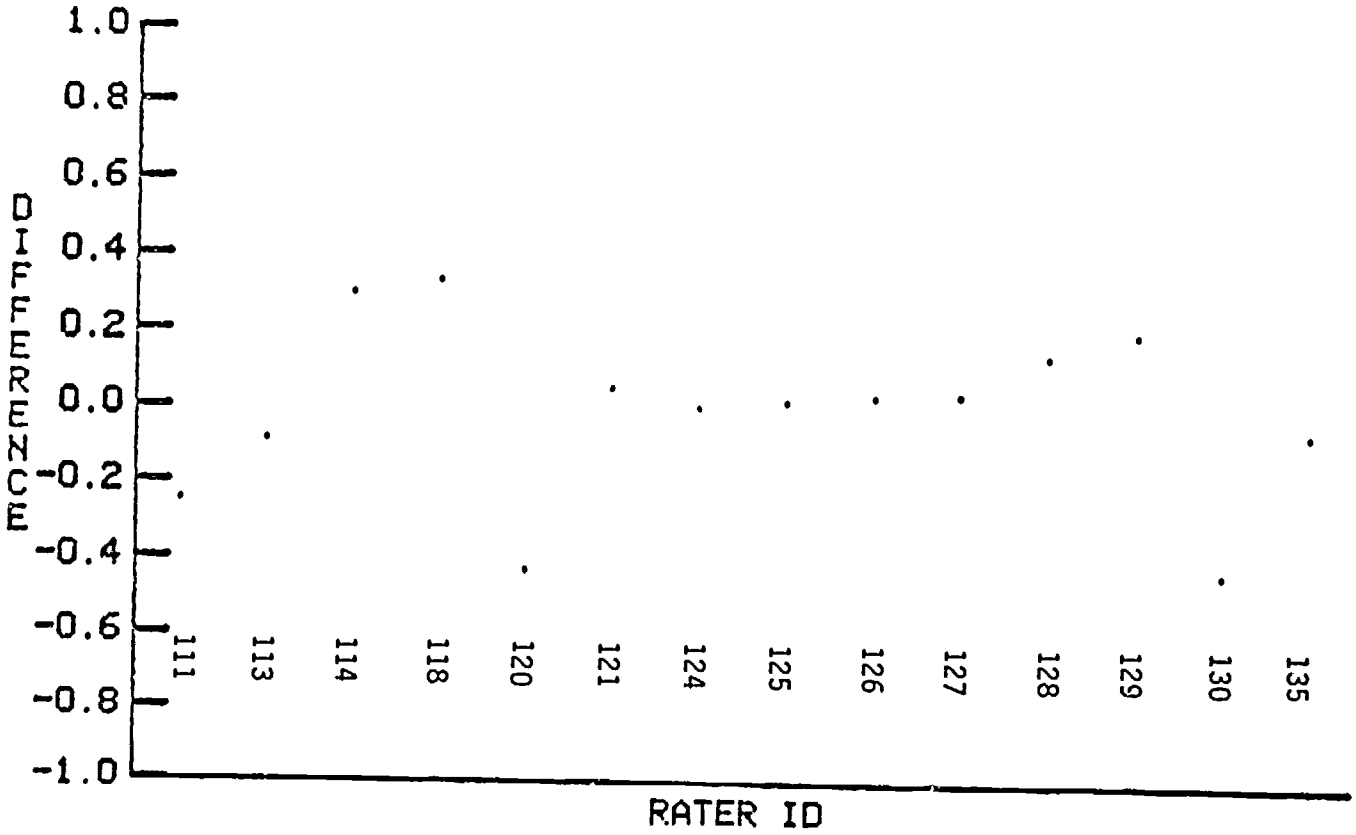


Figure 9

Mean Difference Between Each Rater
and the Paired Raters on the Comprehensibility Score

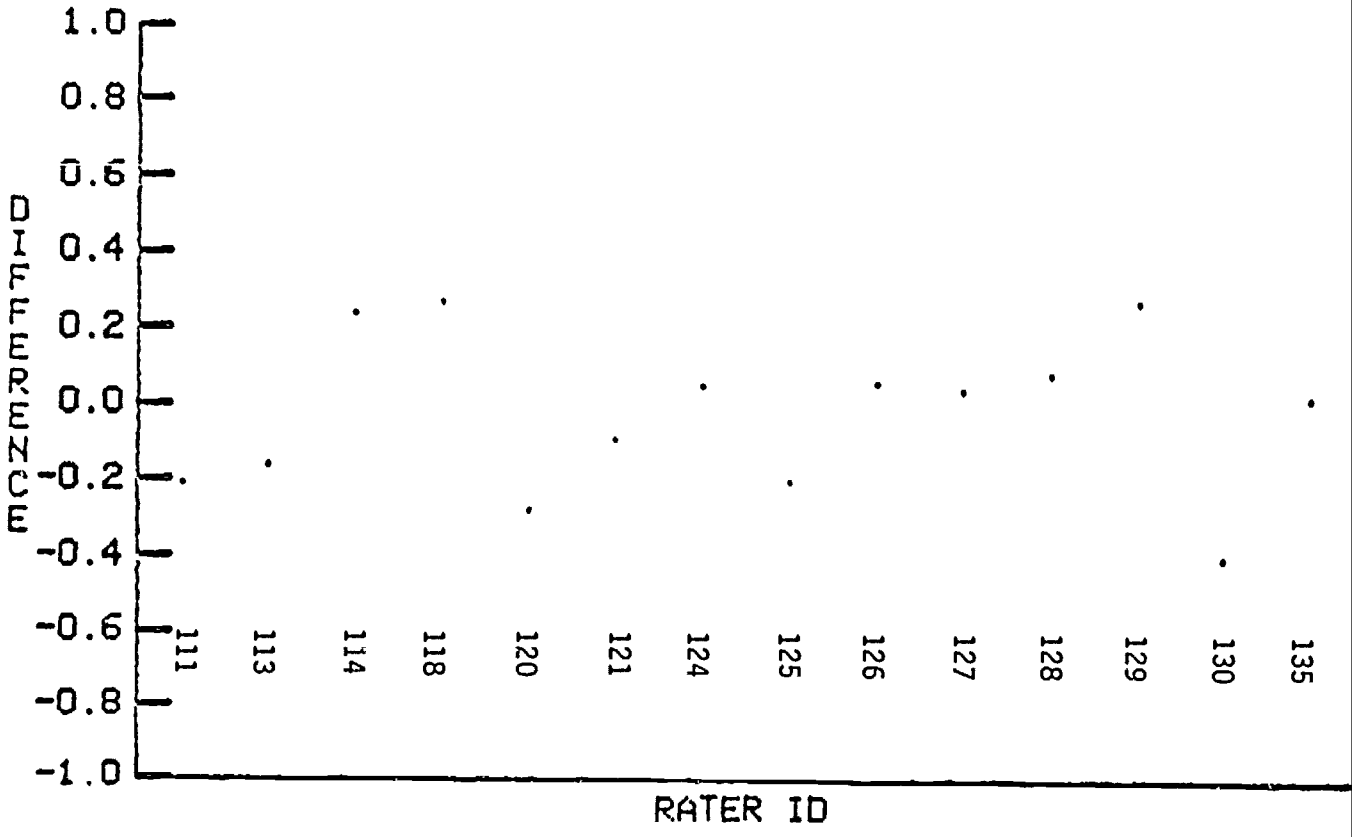


Table 10

Identification of Pairs of Raters Involved in Unusually High Discrepancies

PRON	GRAM	FLU	COMP
	120-121	120-121	120-121
			120-129
	111-114		
		120-114	
		120-129	120-129
		127-111	
	114-111		
		111-120	
111-121			
		120-114	
		120-114	114-120
		120-129	
135-121			
		120-114	
		111-121	
		111-121	
		114-120	114-120
113-118			
		120-121	
		111-121	
		118-121	
128-130		128-130	
		118-113	
128-130			
		111-128	
135-121			
118-120			
129-120			
118-120		118-120	
111-129		111-129	
		114-120	
		129-111	

Note: Although examinee response tapes in all of the above discrepancies received a third rating before the score was reported, only data from the first two ratings were included in this study.

Consistency and validity of individual raters. The question of standards is quite separate from that of consistency and validity. That is, a rater could consistently give lower ratings and yet give ratings that correlate highly with those of other raters. To obtain measures of individual raters' validity, we correlated the ratings of individual raters with the ratings of the paired raters. The results are shown in Table 11.

Table 11

Correlations of Individual Raters with Paired Raters on Each Linguistic Dimension

Rater ID.	N	PRON	GRAM	FLU	COMP	
111	93	.75	.77	.61	.81	*
113	141	.72	.81	.62	.68	
114	59	.87	.75	.74	.83	*
118	174	.77	.80	.71	.76	
120	151	.79	.85	.75	.82	*
121	119	.82	.82	.70	.77	*
124	39	.83	.89	.82	.88	
125	13	.74	.87	.88	.90	
126	22	.77	.93	.88	.89	
127	33	.89	.93	.79	.88	
128	75	.84	.91	.86	.92	
129	89	.80	.85	.81	.82	*
130	13	.23	.80	.57	.82	
135	75	.87	.90	.87	.91	
Median[1]		.79	.85	.77	.83	
Clark & Swinton		.77	.85	.79	.79	

[1]Note: The median correlations depicted here represent the interrater reliability of a TSE score based on a single rating. Official TSE scores are based on two or in some cases three ratings. Thus the reliability of official TSE scores is considerably higher.

There is a substantial range of correlation for each linguistic dimension, but the correlations tend to be in the .70s and .80s. The median correlation is reported at the bottom of the table along with the interrater reliability estimates obtained by Clark and Swinton (1980). It cannot be said that a given rater consistently correlates lower, except for rater 130, who showed a very low correlation on pronunciation and fluency. More important, the five raters identified earlier as generating most of the large discrepancies (marked by asterisks) correlate as well with other raters as anyone else did. Thus, interrater reliability does not appear to determine the likelihood of being involved in a discrepancy.

It is perhaps noteworthy that the data show that the TSE program is obtaining improvements in the degree of agreement among raters. The last two rows of Table 11 offer a comparison between this data and the interrater correlations obtained by Clark and Swinton (1980) in their research study. In general, operational data used in this study show a greater degree of agreement than that obtained in the earlier study. Recently, the program initiated new rating procedures. Two of these procedures should result in further gains in interrater agreement. The first new procedure involved referencing the descriptions utilized by raters in assigning ratings for fluency. The second involved an attempt to improve communication among raters.

Summary and Conclusions

This investigation was undertaken to provide information about the feasibility of using one rater instead of the two that are now used for the TSE. The results suggest that a single rater system would yield highly internally consistent data across scores and that the data could be described by a unidimensional model. Working from that result we examined the possibility that deviations from unidimensionality could be used as a quality-control mechanism to detect instances in which there would be large disagreement if a second rater were involved. The approach that was investigated could not be validated.

We then turned to an analysis of individual raters. The data clearly showed that at least two of the raters appeared to have considerably stricter standards. One of these, rater 130, also had substantially lower correlations on two of the four linguistic dimensions, but rated only thirteen examinees. Rater 120 was involved in a large number of unusually high disagreements; however, this rater correlated as highly with the paired rater as did any other rater.

The foregoing leads to the following conclusion: Because of the possibility of different standards among potential raters, it does not appear feasible to use a single rater as the sole determiner of speaking proficiency at this time. In the remainder of this section two possible alternatives, consistent with the original motivation for the study, will be discussed. One of these possibilities is psychometric; the other is technological.

One possible solution to the problem of different standards among raters is to exclude from the pool those raters who are too severe or too lenient. A more psychometrically oriented solution is to view raters as test forms and to equate them, much as test forms are equated to control for differences in the difficulty of test forms. Although the author is not aware of any testing programs that equates raters, the idea has at least been discussed (de Gruijter, 1980; Pilliner, 1958). Such a psychometric solution would probably require a specialized data collection design. Nevertheless, this study has shown that if we view raters as test

forms, there is reason to believe that their ratings are sufficiently reliable and valid, in the sense of rank ordering examinees in the same fashion as the other raters. Therefore, the idea of equating raters seems feasible from a psychometric point of view.

The second possibility is to use multiple raters, each of whom would be asked to rate only a part of the examinees' performance. That is, different sections of the test could be assigned to different raters. Since the examinees' performance is on tape, it would seem that technological help would be needed. To implement this idea, a system would have to be developed to efficiently create tapes containing portions of examinees' performance. A tape containing the same segment of performance from several examinees would be sent to each rater, who in effect would rate only part of the entire test. The rater would return as many scoring sheets as there were examinees on the tape, but would complete only some parts of the form. That information would, of course, have to be entered into a computer, which would pull together the ratings from several raters to produce a reportable score.

The purely psychometric solution of equating raters is likely to be less expensive. A disadvantage is that the evaluation of an examinee's performance would be based solely on the judgment of a single rater. Even after equating raters there is a possibility that a peculiar rater-examinee interaction could have an effect on the resulting score. By contrast, the second possibility, by involving several raters, would control not only for the different standards that raters might have but also for any possible rater-examinee interaction.

Whatever system is ultimately adopted, the potential vulnerability of individual raters to different criteria should be borne in mind. The present system, even though it uses two raters, is not free from the problem. The results of this investigation suggest that it is imperative to monitor individual raters on a regular basis. An operational system of monitoring, followed by immediate recalibration when necessary, has the potential to maintain rater reliability at a uniformly high level, as well as uniform standards across raters. Such monitoring could also eventually allow the use of a single rater.

References

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. An introduction to Generalizability Theory in Second Language Research. Language learning, 1982, 32(2), 245-258.

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

de Gruijter, D. N. M. The essay examination. In L. J. Th. van der Kamp, W. F. Langerak, and D. N. M. de Gruijter (Eds.), Psychometrics for educational debates. New York: Wiley, 1980.

Gnanadesikan, R. Methods for statistical data analysis of multivariate observations. New York: Wiley, 1977.

Maxwell, A. E. Estimating true scores and their reliabilities in the case of composite psychological tests. British Journal of Mathematical and Statistical Psychology, 1971, 24, 195-204.

Oller, J. W. Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller, Jr. (Ed.), Issues in language testing research. Rowley, MA: Newbury House, 1983.

Pilliner, A. E. G. The rescaling of teachers' estimates. British Journal of Statistical Psychology, 1958, 11, 191-197.

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

Vollmer, H. J., & Sang, F. Competing hypotheses about second language ability: A plea for caution. In J. W. Oller, Jr. (Ed.), Issues in language testing research. Rowley, MA: Newbury House, 1983.

Appendix A

Loadings and residuals for a one-factor model estimated for each rater

N	Rater ID	Loadings				Residuals			
		Pron.	Gram.	Flu.	Comp.	Pron.	Gram.	Flu.	Comp.
93	111				Communalities > 1.0*				
141	113	.90	.75	.79	.92	.02	.08	.08	.02
59	114				Communalities > 1.0*				
174	118	.85	.88	.84	.99	.03	.02	.03	.00
151	120	.95	.91	.96	.99	.02	.02	.01	.00
119	121	.96	.78	.91	.98	.03	.04	.03	.01
39	124	.90	.92	.89	.99	.02	.04	.03	.00
13	125	.97	.84	.97	.98	.01	.04	.03	.02
22	126	.98	.94	.90	.99	.00	.01	.01	.01
33	127	.98	.95	.95	.91	.01	.02	.03	.03
75	128				Communalities > 1.0*				
89	129	.92	.86	.91	.99	.02	.02	.02	.00
13	130	.74	.82	.94	.99	.11	.12	.02	.01
75	135	.93	.92	.96	.98	.01	.01	.01	.00

*It was not possible to estimate the parameter of the factor model for raters where one or more of the communalities were greater than 1.

TOEFL® Research Reports currently available . . .

Report 1. *The Performance of Native Speakers of English on the Test of English as a Foreign Language.* John L. D. Clark. November 1977.

Report 2. *An Evaluation of Alternative Item Formats for Testing English as a Foreign Language.* Lewis W. Pike. June 1979.

Report 3. *The Performance of Non-Native Speakers of English on TOEFL and Verbal Aptitude Tests.* Paul J. Angelis, Spencer S. Swinton, and William R. Cowell. October 1979.

Report 4. *An Exploration of Speaking Proficiency Measures in the TOEFL Context.* John L. D. Clark and Spencer S. Swinton. October 1979.

Report 5. *The Relationship between Scores on the Graduate Management Admission Test and the Test of English as a Foreign Language.* Donald E. Powers. December 1980.

Report 6. *Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.* Donald E. Powers and Spencer S. Swinton. December 1980.

Report 7. *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings.* John L. D. Clark and Spencer S. Swinton. December 1980.

Report 8. *Effects of Item Disclosure on TOEFL Performance.* Gordon A. Hale, Paul J. Angelis, and Lawrence A. Thibodeau. December 1980.

Report 9. *Item Performance Across Native Language Groups on the Test of English as a Foreign Language.* Donald L. Alderman and Paul W. Holland. August 1981.

Report 10. *Language Proficiency as a Moderator Variable in Testing Academic Aptitude.* Donald L. Alderman. November 1981

Report 11. *A Comparative Analysis of TOEFL Examinee Characteristics, 1977-1979.* Kenneth M. Wilson. July 1982

Report 12. *GMAT and GRE Aptitude Test Performance in Relation to Primary Language and Scores on TOEFL* Kenneth M. Wilson. July 1982.

Report 13. *The Test of Spoken English as a Measure of Communicative Ability in the Health Professions. Validation and Standard Setting.* Donald E. Powers and Charles W. Stansfield. January 1983.

Report 14. *A Manual for Assessing Language Growth in Instructional Settings.* Spencer S. Swinton. February 1983.

Report 15. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students* Brent Bridgeman and Sybil Carlson. September 1983.

Report 16. *Summaries of Studies Involving the Test of English as a Foreign Language, 1963-1982.* Gordon A Hale, Charles W. Stansfield, and Richard P. Duran. February 1984.

Report 17. *TOEFL from a Communicative Viewpoint on Language Proficiency. A Working Paper.* Richard P Duran, Michael Canale, Joyce Penfield, Charles W. Stansfield, and Judith E. Liskin-Gasparro. February 1985.

If you wish additional information about TOEFL research or would like to be placed on the mailing list to automatically receive order forms for newly published reports, write to:

TOEFL Program Office
CN 6155
Princeton, NJ 08541-6155
USA

5790038 • [illegible] • [illegible] • [illegible] U.S.A.