Oral Skills Testing: A Rhetorical Task Approach[1]

Anne Lazaraton
University of California, Los Angeles

Heidi Riggenbach
University of Washington, Seattle

This paper discusses the development, implementation, and evaluation of a semi-direct test of oral proficiency: the Rhetorical Task Examination (RTE). Many of the commonly used speaking instruments assess oral proficiency in terms of either discrete linguistic components--fluency, grammar, pronunciation, and vocabulary--or in terms of a single, global ability rating. The RTE proposes a compromise approach to rating oral skills by having two scales: one which ascertains the functional ability to accomplish a variety of rhetorical tasks, the other to address the linguistic competence (Canale & Swain, 1980) displayed in the performance.

On audiotape in a language laboratory setting, 52 students representing three levels of a university ESL program performed six tasks related to the rhetorical modes covered in their coursework: short questions and answers, description, narration, process (giving directions), opinion, and comparison-contrast. The construction and justification of both the instrument and the rating scales are explained; data obtained from administering the RTE across classes as well as before and after instruction are presented; and the relevant measurement characteristics of the test are discussed. Results of this study indicate that the Rhetorical Task Examination is promising as a measure of oral proficiency in terms of practicality, reliability, and validity.

INTRODUCTION

The testing of oral proficiency is one area in applied linguistics in which, until recently, practice has lagged behind theory, though speaking in a second or foreign language is arguably the most important of the traditional four skill areas (reading, writing, listening, and speaking). Unfortunately, it is also the most problematic to measure. Nevertheless, in the last decade, much effort has gone into the development, implementation, and evaluation of instruments which assess oral ability (e.g., Bachman

Issues in Applied Linguistics
ISSN 1050-4273

Vol. 1 No. 2 1990 196-217

& Palmer, 1981, 1982; Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Palmer et al., 1981; Pienemann et al., 1988; Raffaldini, 1988; Shohamy, 1988). Many oral proficiency tests, such as the Cambridge First Certificate Examination (FCE) (UCLES, 1987), the General Tests of English Language Proficiency (G-TELP) (TENEC, 1985) and the Speaking Proficiency English Assessment Kit (SPEAK) (Educational Testing Service, 1985) use discrete component-oriented rating scales to evaluate aspects of oral proficiency. They look at orally produced language as something which can be parcelled into separate categories for rating purposes, such as fluency, grammar, pronunciation, and vocabulary. While such discrete components certainly contribute to what is called 'oral proficiency', it may be desirable in a testing situation to go beyond these restrictive linguistic categories for a broader view of a student's language use in a given performance (Carroll, 1980).

Some established tests do take a more holistic perspective towards the assessment of oral proficiency, most notably the Foreign Service Institute/Interagency Language Roundtable (FSI/ILR) Interview (Lowe, 1982) and its derivative, the American Council on Teaching Foreign Languages/Educational Testing Service (ACTFL/ETS) Oral Proficiency Interview (ACTFL, 1986). These tests are not only based on a functional trisection (Clark & Clifford, 1988) of linguistic accuracy, functional breadth, and a range of topics and situations, they also define performance requirements at each proficiency level in terms of these three components. Nevertheless, their rating scales provide a single, global rating of proficiency. This practice has been criticized on the grounds that neither current linguistic theory nor language testing research supports the notion of language as a unitary ability (Bachman, 1988, 1990).

Therefore, the challenge facing us was to develop a test, complete with rating scales, that views a nonnative speaker's performance data across a variety of rhetorical tasks, not just in terms of linguistic competence (Canale & Swain, 1980; Canale, 1983) but also in terms of functional ability--that is, how well the speaker accomplishes a rhetorical task and maintains coherent, comprehensible discourse. The primary objective of this paper is to report the development of such a speaking test with its corresponding rating scales.

Ideally, a direct approach (Clark, 1979) to oral proficiency testing, in which spontaneous, face-to-face interaction is required, should be used. However, tests which most closely approximate

authentic interaction require extensive training to administer and score and can be costly as well. Therefore, in the interest of practicality and feasibility, we chose to develop a semi-direct test in which students are provided with a test booklet as well as aural stimuli and for which responses are recorded on audiotape. While semi-direct tests do not constitute true interactive conversation, they have been found to correlate highly with and can be considered acceptable alternatives to direct interviewing procedures (Clark & Clifford, 1988; Lowe & Clifford, 1980).

One goal which guided the project was to determine if our speaking instrument measured a range of ability which paralleled level placement as assessed by the examination at the university where this project was undertaken, the UCLA ESLPE (University of California, Los Angeles, English as a Second Language Placement Examination). Many university ESL/EAP placement tests have reading, writing, listening, and grammar components, since these skill areas are the most easily and efficiently tested. However, many of these same university ESL/EAP programs also include courses which teach and practice oral skills, though their placement tests may not measure these skills, directly or indirectly. This was the case at UCLA where addressing this situation became one of the objectives for developing our test.

A second goal, which also figured into the design of the instrument, was our interest in creating an oral achievement test that paralleled class activities in a low-intermediate multi-skills ESL course (ESL 33A) organized around rhetorical modes. That is, the test was also designed to assess students' functional and linguistic ability to perform various rhetorical tasks before and after ten weeks of instruction.

A final objective in the test development process was to evaluate the instrument's measurement characteristics, specifically its reliability and validity.

METHOD

Subjects

The *Rhetorical Task Examination* (*RTE*) was administered in the Fall quarter, 1986, to a total of 52 ESL/EAP students at UCLA (17 in ESL 33A, a low-intermediate course, 17 in ESL 33B, a high-intermediate course, and 18 in ESL 33C, an advanced course).

These three multi-skills courses were required of foreign students who placed below an exempt proficiency level as determined by the ESL Placement Examination (ESLPE) 1986 version.[2]

Administration

The RTE, which takes approximately 18 minutes, was administered on the second day of class in a language laboratory where an entire class could be tested and recorded in one sitting. Each student was presented with a test booklet containing written instructions, prompts, and pictures, but instructions and prompts were given aurally as well. Students, as a group, were allowed specific allotments of time (from 1 to 1 1/2 minutes) to respond to each rhetorical task on the audiotape. These tapes were then used for purposes of rating, which took place *after* the test administration (see the description of the rating procedure below). Since the students in all but the advanced course (ESL 33C) also took the RTE at the end of the ten-week quarter, there are pretest and posttest ratings on approximately two-thirds of the subjects. In view of our stated goals, the purpose of the pretest was to assess the instrument's utility in measuring oral proficiency across a range of student ability, while the main purpose of the posttest was to measure achievement in the two courses (ESL 33A and 33B) that devote a substantial amount of time to the teaching of the speaking skill.

Instrument

The RTE is composed of two sections. The task section includes several "warm-up" short-answer questions (not rated) plus five rhetorical tasks: description, narration, process (giving directions), opinion, and comparison-contrast. This section uses pictures and/or short instructions to elicit talk (see Underhill, 1987, for a discussion of the merits of using pictures to elicit oral language). The rhetorical tasks were chosen according to the following criteria: a) whether or not they parallel the rhetorical patterns taught in the university's ESL courses (though the emphasis in the required courses, especially in the advanced course, ESL 33C, is on writing in these modes); b) whether or not they encompass the kinds of speaking activities that take place in the courses where speaking is taught, thus allowing the test to serve as a measure of achievement; and c) whether or not we perceived them as

authentic or natural speaking situations which students face both in the university environment and in everyday life. Examples A and B in Appendix A show the prompts for the opinion and comparison-contrast tasks.[3]

The second section, an imitation section, was also included as an additional testing method, one that is less direct than the task section but easier to administer and rate. The imitation passage--similar to one developed by Henning (1983), but here contextualized so that the entire passage was a connected story--included ten sentences, ranging from 2 to 12 words, which students heard and then repeated.

Two "equivalent" forms (A and B) of the RTE were created so that one could be used as a posttest.[4] Schematically, the test versions were as follows:

Version 1: Tasks A, Imitation A
Version 2: Tasks B, Imitation B
Version 3: Imitation A, Tasks A
Version 4: Imitation B, Tasks B

The test versions were randomly assigned to students to control for possible ordering effects.

Rating Scales

From a practical standpoint, it was initially hoped that one rating scale could be developed that would cover both the functional skills necessary to perform the rhetorical tasks and the linguistic skills that oral proficiency tests have traditionally measured. It became apparent early on, however, that this would be impractical because many of the students did not perform consistently well or performed poorly in both these areas. For example, a student low in grammar and pronunciation skills might nevertheless perform well on a given task--that is, the student's giving of directions on how to get from one place to another might be clear and successfully comprehended (Raffaldini (1988) came to a similar conclusion: the acquisition of the various traits of communicative competence does not necessarily proceed in a parallel fashion). Therefore, two sets of 4-point scales were developed for each rhetorical task: one set for rating overall task performance (i.e., functional ability) and one set for rating linguistic skills on that same task. It should be pointed out that we are not advocating the *replacement* of the linguistic skills scale with the functional ability/task scale, but its use as a way of

attaining additional, yet fundamental, information on the subjects' oral ability.

The functional ability rating scales for the rhetorical tasks were created after considering data obtained from two native speakers, six nonnative speakers not involved in the UCLA ESL classes, and 18 ESL 33A students who had been given a pilot version of the RTE in 1985. A detailed discourse analysis of these responses (Riggenbach & Lazaraton, 1988) generated a set of important components for each task, which were extracted and built into the descriptors for each scale.

The linguistic skills rating scale is a combination of various oral proficiency scales from several sources (e.g., Harris, 1969; Oller, 1979; TENEC, 1985). As was mentioned, we were dissatisfied with the way these scales view linguistic skills separately; we hoped to evaluate linguistic ability holistically. While our scale does describe discrete skill areas which comprise oral proficiency, an attempt was made to have the scales focus on comprehensibility rather than on accuracy--in other words, how well control or lack of control of linguistic skills contributed to or detracted from the ability to accomplish the task. In addition, we thought it would be possible and practical to group linguistic skill areas (fluency, grammar, pronunciation, and vocabulary) together, rather than to rate each skill for each task. Although it might seem counterintuitive to group all linguistic skills together, it was the rare case when one particular area (e.g., pronunciation) differed more than one point from other skill areas (fluency, grammar, vocabulary). Rather than assuming an invariant relationship between skills, the approach we took was to assign the rating which best described the subject's linguistic performance on each task. Accordingly, each point on the rating scale includes these four skills. Appendix B shows the two rating scales for one of the tasks, the narrative.

In addition, for the pretest, each student's overall linguistic performance on the task section of the RTE was rated using a third scale (see Appendix C) adapted from an FSI "supplemental" rating scale (Educational Testing Service, 1970; reprinted in Oller, 1979, pp. 321-323). Fluency, grammar, pronunciation, and vocabulary were rated separately on this third scale in order at the onset to determine if our linguistic skills scale related to a more established rating scale. The ratings from this scale, however were included only in the analyses of the pretest results. The rating of the imitation task is discussed in the following section.

Rating Procedures

For this project, we served as test administrators, as raters, and as co-instructors for the low-intermediate course (ESL 33A) at the time of the test administration. However, ratings for the first administration were blind in that none of the students was known to us at the beginning of the quarter. In addition, ratings were done independently, and tapes were chosen randomly from a pool of tapes from all levels.

For the imitation section, speech samples were rated for accuracy and intelligibility (in the sense of "error-free repetition" as in Henning, 1983, p. 317). Typical errors included addition, substitution, or deletion of words, incorrect inflection of words (third person singular -s, past tense -ed), and incorrect or unintelligible pronunciation of words. Raters checked to see if they made similar judgments about correct vs. incorrect inflection and pronunciation. The total rating for this section represented the total number of words correct.

Pilot data were used for training purposes. After each task was rated independently, the ratings were compared and discussed, prompting minor adaptations to the rating scales.

It was decided that for cases of disagreement of more than three points on the imitation rating and two points on the functional ability/task and linguistic skills ratings, a third rater, trained later, would be appointed. This was necessary in only three cases, perhaps because of the raters' initial contribution to the rating scale and the rigor of the training session.

RESULTS

In this section we discuss the statistical analyses which address the questions posed at the beginning of this paper. To reiterate, we wanted to determine both if the Rhetorical Task Examination measured a range of student ability which mirrored course placement by the UCLA ESLPE and if the RTE would be suitable as a measure of oral achievement in the low-intermediate ESL 33A class. An alpha level of $<.05$ was selected for all statistical decisions.

Course Placement

To determine if the RTE would be appropriate for course placement (specifically, did the test measure a range of ability which paralleled level placement by the ESLPE?), only pretest data were used for between-class comparisons. Table 1 gives the descriptive statistics for all three classes on four measures: 1) functional ability/task rating (5 tasks, 4 possible points per task); 2) linguistic skills rating (5 tasks, 4 possible points per task); 3) FSI-adapted rating (4 areas, 5 possible points per area); and 4) imitation rating (70 possible points).

TABLE 1
Pretest Descriptive Statistics by Class

Class	N	Functional ability		Linguistic skills		FSI-adapted		Imitation	
		Mean	sd	Mean	sd	Mean	sd	Mean	sd
33A	17	10.71	2.7	10.35	3.4	10.59	3.4	45.97	14.4
33B	17	13.79	2.9	12.00	3.3	11.62	3.0	52.12	11.7
33C	18	12.89	2.6	13.69	3.8	13.56	2.8	59.75	11.0

As Table 1 shows, students in the higher levels (33B and 33C) were rated higher than students in 33A on all measures. Yet, while the lowest ratings were consistently assigned to 33A, 33C showed the highest ratings for all but the functional ability/task for which 33B was rated the highest. Due to the numerical differences in the means present in Table 1, the nonparametric Kruskal-Wallis One-way Analysis of Variance procedure was chosen to test for statistical differences in each of the pretest measures across classes.[5] These results appear in Table 2.

TABLE 2
Kruskal-Wallis Test
Pretest Measures
(N=52)

Measure	x^2	eta^2
Functional ability/task	8.73*	.17
Linguistic skills	7.13*	.14
FSI-adapted	8.73*	.17
Imitation	10.31*	.20

*$p < .05$

For the functional ability/task ratings, Table 1 showed that the 33B ratings were numerically the highest and that the 33A ratings were the lowest. Statistically, the Kruskal-Wallis test indicated a significant difference between the ratings for all three classes. A post hoc Ryan's procedure indicated that the 33A ratings were significantly lower than the 33B ratings at $p < .05$. However, the eta^2 value showed a weak strength of relationship: only 17% of the variance in functional ability/task ratings could be accounted for by the class in which a student was enrolled.

The results of testing for differences between the three classes on the remaining three pretest measures (linguistic skills, FSI-adapted, imitation) showed identical results for each measure: there were significant differences in the ratings for the three classes, and the 33A (low-intermediate) ratings were significantly lower than the 33C (advanced) ratings at $p < .05$. Although the differences in ratings between 33A and 33B and between 33B and 33C were not significant, the pattern illustrated here, higher level = higher rating, is what one would expect, given the purpose of the ESLPE. Again, the class membership did not explain a great deal of the variance in these three pretest measures (from 14% to 20%).

Therefore, the results of the RTE were similar to the results of the ESLPE in terms of level placement, with the exception of the functional ability/task ratings: in general, the higher the rating, the higher the level (the reverse also being true). Of course, level had already been determined at the time of the pretest.

Achievement

Our second question was whether low-intermediate (ESL 33A) students made measurable performance gains by the end of the ten-week quarter. To test this statistically, the nonparametric Wilcoxon Matched-Signs Ranked-Pairs test was selected. Table 3 shows the results of the Wilcoxon tests for the 33A students on the three posttests.

TABLE 3
Wilcoxon Matched-Signs Ranked Pairs Test
33A Pretest and Posttest Ratings

	Pretest Mean	sd	Posttest Mean	sd	z	N	eta²
Functional ability/task	10.75	2.8	13.34	2.9	3.35*	16	.75
Linguistic skills	10.50	3.5	11.47	3.6	2.07*	16	.28
Imitation	45.97	14.4	53.38	9.0	2.89*	17	.52

* $p < .05$
Note: The differences in means and standard deviations in Tables 1, 3, and 4 are due to student attrition in the courses.

As Table 3 shows, low-intermediate (33A) students appear to have made significant gains on all three measures, the most dramatic being in functional ability/task ratings. There is a very strong relationship between functional ability/task rating and time ($eta^2 = .75$). While the 33A students also made significant gains on the other two measures, the strength of relationship between these measures and time was not as impressive, especially for the linguistic skills ratings. Yet, because there was no control group, we cannot be sure why these gains were made. Analyses of the high-intermediate (33B) posttest data showed these students did *not* make significant gains from pretest to posttest on any of the measures, as shown in Table 4.

TABLE 4
Wilcoxon Matched-Signs Ranked Pairs Test
33B Pretest and Posttest Ratings

	Pretest Mean	sd	Posttest Mean	sd	z	N
Functional ability/task	13.93	2.5	13.70	2.8	.56	15
Linguistic skills	12.10	3.1	12.67	3.6	1.42	15
Imitation	53.57	12.3	53.25	10.8	.34	15

* $p < .05$

Measurement Characteristics

With regard to the reliability of semi-direct tests of oral proficiency, test content variation is not a problem, but the assignment of consistent ratings is a concern (Clark & Clifford, 1988). Therefore, inter-rater reliability estimates for the task section of the RTE were calculated along three dimensions: the two versions (A and B) of the test, the two times at which the test was administered (pretest and posttest), and the two scales on which students were rated (functional ability/task and linguistic skills). The eight resulting intra-class correlation coefficients ranged from .93 to .98. These high figures were undoubtedly achieved by our close attention to consistent initial ratings of the samples and by the periodic scoring checks we undertook (on this point see Clark & Clifford, 1988). Internal consistency of the RTE itself was measured by alpha coefficients for the task section. The estimates for the two versions of the test (A and B) were .96 and .97 (pretest) and .98 and .95 (posttest). The KR-21 reliability estimates for imitation total ratings were .94 (pretest) and .88 (posttest).

With respect to content validity, we feel that since the rhetorical tasks were selected, in part, to replicate the rhetorical organization patterns covered in the courses, it can be assumed that the instrument exhibits a fair degree of content validity. This could be tested by having independent experts make judgments as to the degree of match between the RTE criteria and the content being measured.

The predictive capacity of the RTE has not been fully explored because the sample size was not sufficient to permit multiple regression analysis, the procedure of choice for this type of research question. Instead, correlations of the pretest measures,[6] shown in Table 5, can be considered preliminary information for answering questions about predictive validity.

TABLE 5

Pearson Correlations of Pretest Measures

	ESLPETOT	PREFA	PRELS	PREFSI	PREIMIT
ESLPETOT*	1.000				
PREFA**	.410	1.000			
PRELS***	.367	.729	1.000		
PREFSI****	.357	.573	.904	1.000	
PREIMIT*****	.412	.452	.654	.690	1.000

$p < .01$ for all correlations

*ESLPE total score; **Functional/Ability task; ***Linguistic skills; ****FSI-Adapted; *****Imitation

Some interesting trends present themselves in Table 5. One is the relatively low correlation between any of the measures on the RTE and the ESLPE total score. The correlations range from .36 to .41, meaning that since none of these measures alone accounts for more than about 16% of the variance in the placement exam's scores, none alone would be a good predictor of a student's ESLPE score. In any case, the RTE was not designed to predict overall language proficiency (which is what the ESLPE purports to measure), but specifically oral proficiency.

Another result seen in Table 5 is the fairly high correlation between the linguistic skills ratings and the functional ability/task ratings ($r = .73$), which means that one measure accounts for 53% of the variance in the other. This correlation suggests that some, but not all, information is shared by the two measures.

Finally, tentative validation of our linguistic skills rating scale is shown by its high correlation with the FSI-adapted scale ($r = .90$), a result which indicates that the two scales share 81% variance. In addition, ratings from the linguistic skills scale show fairly high correlations with the individual FSI-adapted scale areas: fluency, $r = .85$; grammar, $r = .76$; pronunciation, $r = .81$; vocabulary, $r = .77$.

DISCUSSION

To review our findings, we found that the Rhetorical Task Examination, in general, gave results parallel to those of the ESLPE. It could be expected that 33A (low-intermediate) students would be rated lower than either 33B or 33C students (high-intermediate and

advanced, respectively), but it is not immediately clear why the 33B (and not the 33C) students received the highest functional ability/task ratings. Perhaps there was something unique about this 33B class; on the other hand, further test administrations might prove this occurrence to have been nothing more than a statistical aberration. In any case, class membership did not prove to be strongly related to ratings on the pretest measures; clearly, other factors were at work. Perhaps the answer could be found in a future analysis of demographic data (e.g., sex, major, native language, length of stay in U.S.). Another interesting question is whether placement decisions would have been the same regardless of the measure used, if both the ESLPE and the RTE had been given simultaneously.

Secondly, the ESL 33A students made significant gains in oral skills, as measured by the RTE, during ten weeks of instruction, while ESL 33B students did not show comparable gains. It is tempting to say that this was because, in contrast to the 33A course, oral skills are not routinely stressed in the 33B course, but there are other plausible explanations. One such explanation is that since the 33A students took the posttest as a part of the final exam, it is likely that they had high motivation to do well (perhaps higher than the 33B students). The 33B students, on the other hand, knew that their posttest performance had no bearing on their grades, and thus this motivation may not have existed for them. Another possibility is that lower level students (33A) are more likely to make gains that can be detected statistically than are higher level (33B) students. Unfortunately, the absence of the 33C students at the time of the posttest limits our interpretation of the inter-class differences in achievement. What *is* interesting about Tables 3 and 4 is the relative equivalence (in terms of speaking ability) of 33A and 33B students at the end of the quarter, as can be seen by examining the pretest and posttest ratings. However, whether the gains 33A students made were due to instruction, maturation, or other uncontrolled factors cannot be ascertained from this study.

Finally, evidence has been provided to suggest that the RTE is promising in terms of its measurement characteristics. Not only was very high interrater reliability obtained for both rating scales, the test showed high internal consistency as well. A fairly strong relationship between functional ability/task and linguistic skills ratings was also found, but this could be because for each tape both ratings were assigned at the same time. While even a higher correlation between the two measures might suggest that the

functional ability/task scale alone could be used to focus on *how* each task was accomplished, we would opt for continuing to use both scales, since, for reasons discussed in this paper, we feel it is not intuitively practical to combine both scales into one. Finally, the very strong relationship between the linguistic skills rating scale and a third, FSI-adapted, scale suggests that our linguistic skills scale, which measures four linguistic skills (fluency, grammar, pronunciation, vocabulary) as a group, appears to be an efficient way to assess linguistic performance.

CONCLUSION

Our primary motivation for developing the Rhetorical Task Examination was to create a practical and efficient method of assessing oral proficiency. Consideration was given to the concept of language as a complex system. Rhetorical tasks were designed so as to allow students flexibility in communicating real information and in expressing their own perspectives. Rating scales were created which looked first, and most importantly, at nonnative speakers' ability to communicate this information in a cohesive and comprehensible manner, and second, at the linguistic skills which enabled them to do this. Such an approach tried to take into account the idea that language is more than grammatical accuracy; that it is a *system* of many levels with holistic goals--communication and self-expression. Thus, the RTE tried not to isolate form from function (although both were rated separately), and the important goals of expediency and practicality were kept in mind.

In addition, the RTE (with its achievement test function) was intended to look at the *process* of language learning for students enrolled in the low-intermediate course (33A). The tasks in the test thus parallel the rhetorical patterns introduced, analyzed, and practiced by students throughout this course. However, though students made gains on the posttest as compared to the pretest, questions still remain as to why this was so: Were the instructional activities effective; did they assist students in figuring out the structures used and the components of these various modes? Would measurable gains in functional and linguistic ability have occurred anyway because of exposure to the language in environments both outside and inside the academic setting?

Several limitations of this project should be mentioned. For one, since problems can occur from teacher and researcher

expectations when the teachers, testers, raters, and researchers are the same people, we are therefore cautious in interpreting our results. Secondly, since intact classes of non-randomly selected students were used to answer our questions, it is unwise to generalize our results to other situations. Finally, while we feel that the RTE is promising in terms of reliability, validity, and practicality, it fails to tap a fundamental feature of oral proficiency: the ability to interact with interlocuters. This is a crucial difference between semi-direct tests, as ours is, and oral interviews, a difference which explains the broad appeal of interview-type testing.

A final point, made by Byrnes (1987), is that the ultimate goal of oral testing is to go beyond product assessments to process recommendations for materials and curriculum development. The product assessments made by the RTE have been described in detail in this paper, but it is worth mentioning briefly the other uses to which the RTE and the data obtained from it have been put. Since this project began, we have routinely used the test as a diagnostic instrument in many of the courses we have taught at UCLA. A useful follow-up activity for the classroom is to have students transcribe various rhetorical tasks from their tapes and then analyze their speech for phonological, lexical, grammatical, and/or discourse features. We have also had students generate written texts from the test prompts, which, in conjunction with their spoken texts, were used as a basis for discovering differences between spoken and written discourse. Related to this pedagogical activity was a larger project in which the comparison-contrast task was used as a prompt for eliciting spoken and written data from both native and non-native speakers. The resulting texts generated a database, parts of which have been used for various language analysis projects (e.g., Lazaraton, forthcoming; Riggenbach, 1989; Turner, 1989). We hope, therefore, that others with an interest in oral proficiency test development will benefit from our experience with the Rhetorical Task Examination itself and with its broader applications to the field of applied linguistics.

NOTES

[1]We would like to thank Brian Lynch, Fred Davidson, Evelyn Hatch, and various anonymous reviewers for their informative responses to earlier versions of this paper as well as Donna Brinton and Grant Henning for their assistance in implementing this project. Any errors that remain are our own.

[2]In brief, the ESLPE of 1986, composed of 150 items, had five subtests: listening, reading, grammar, vocabulary, writing error detection, and a composition task. Evidence of the reliability of this version of the ESLPE is

available in terms of internal consistency estimates such as the KR-21 which was .944 for the Fall 1986 administration, with a mean of 102.9 and a standard deviation of 22.34 (N = 798; k = 150); in terms of norm-referenced reliability, correlational and regression analyses with the TOEFL suggest that the ESLPE is a valid measure (Lynch, 1985).

[3]For complete copies of the oral skills test and the rating scales described in this paper, write to:

Anne Lazaraton
Department of TESL and Applied Linguistics
UCLA
3300 Rolfe Hall
Los Angeles, CA 90024-1531

[4]According to Henning (1987), equivalent tests must show three characteristics: a) equivalent mean scores, b) equivalent variances, and c) equivalent covariances (or equivalent correlations with a third measure). An ANOVA procedure with Bartlett-Box F for equality of variance was used to check assumptions a) and b) for both the task section ratings (linguistic skills rating + functional ability rating) and the imitation section ratings. For the task section, F (1,51) = .882, n.s.; Bartlett-Box F = 1.004, n.s. For the imitation section, F (1,51) = 1.54, n.s.; Bartlett-Box F = .153, n.s. Assumption c) was not met. The two test forms showed differing correlations with the ESLPE: imitation section, r = .26 and r = .53; task section, r =.29 and r =.50 for Forms A and B, respectively.

[5]This nonparametric test was selected instead of the more conventional parametric ANOVA procedure because we are not convinced that the research design used or the data collected in this study allow us to meet the assumptions of Analysis of Variance (e.g., normal distribution of data, equal variances, interval-level measurement). Furthermore, the study needs to be replicated before sufficient evidence can be obtained to support the utility and measurement adequacy of the instrument. Therefore, various nonparametric procedures (Wilcoxon Matched-Signs Ranked-Pairs Test, Ryan's procedure, eta^2 strength of association) will be reported. See Hatch & Lazaraton, (1991) for a thorough discussion of these and other procedures.

[6]Strictly speaking, parametric Pearson correlations may be inappropriate for these data, given the concerns voiced above. However, a nonparametric Spearman rho rank-order correlation matrix showed correlations which were virtually identical to the Pearson coefficients. Since Pearson correlations are more easily interpreted than Spearman coefficients, the former are reported.

REFERENCES

ACTFL (American Council on the Teaching of Foreign Languages) (1986). *ACTFL/ETS proficiency guidelines*. Hastings-on-Hudson, NY: Author.

Bachman, L. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*, 149-164.

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

Bachman, L.F. & Palmer, A.S. (1981). The construct validation of the FSI oral interview. *Language Learning, 31*, 167-186.

Bachman, L.F. & Palmer, A.S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*, 449-465.

Bachman, L.F. & Savignon, S. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*, 380-390.

Byrnes, H. (1987). Second language acquisition: Insights from a proficiency orientation. In H. Byrnes & M. Canale (Eds.), *Defining and developing proficiency: Guidelines, implementations, and concepts* (pp. 107-131). Lincolnwood, IL: National Textbook Company.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Carroll, B.J. (1980). *Testing communicative competence: An interim study*. Oxford: Pergamon.

Clark, J.L.D. (1979). Direct vs. indirect tests of speaking ability. In E.J. Briere & F.B. Hinofotis (Eds.), *Concepts in language testing: Some recent studies* (pp. 35-49). Washington, DC: TESOL.

Clark, J.L.D. & Clifford, R.T. (1988). The FSI/ILR/ACTFL proficiency scale and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition, 10*, 129-147.

Educational Testing Service. (1970). *Manual for peace corps language testers*. Princeton, NJ: Author.

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

Harris, D.P. (1969). *Testing English as a second language*. New York: McGraw-Hill.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.

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

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Lantolf, J.P. & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *Modern Language Journal, 69*, 337-345.

Lazaraton, A. (forthcoming). Linking ideas with AND in spoken and written discourse. *IRAL*.

Lowe, P.J. (1982). *ILR handbook on oral interview testing*. Washington, DC: DLI/LS Oral Interview Testing Project.

Lowe, P.J. & Clifford, R.T. (1980). Developing an indirect measure of overall oral proficiency. In J.R. Frith (Ed.), *Measuring spoken language* (pp. 31-39). Washington, DC: Georgetown University Press.

Lynch, B. (1985). *ESLPE manual*. Unpublished manuscript, UCLA.

Oller, J. (1979). *Language tests at school*. New York: Longman.

Palmer, A.S., Groot, P.J.M., & Trosper, G.A. (Eds.). (1981). *The construct validation of tests of communicative competence*. Washington, DC: TESOL.

Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition, 10*, 217-243.

Raffaldini, T. (1988). The use of situation tests as measures of communicative ability. *Studies in Second Language Acquisition, 10*, 197-216.

Riggenbach, H. 1989. *Nonnative fluency in dialogue versus monologue speech: A microanalytic approach*. Unpublished doctoral dissertation, UCLA.

Riggenbach, H. & Lazaraton, A. (1988). *Discourse competence: Assessing structuring of second language learner texts*. Unpublished manuscript.

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition, 10*, 165-179.

TENEC. (1985). *General tests of English language proficiency*. Los Alamitos, CA: Author.

Turner, J. (1989). *An analysis of syntax-level and discourse-level comparison in native and non-native writing*. Unpublished qualifying paper, Dept. of TESL & Applied Linguistics, UCLA.

Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.

UCLES (University of Cambridge Local Examinations Syndicate). (1987). *First certificate in English* (Paper 5: Interview). Cambridge: Author.

APPENDIX A

EXAMPLE A
Prompt for Opinion Task

Some people believe that all high school students in every country should be required to learn at least one foreign language. What do you think?

You will have 15 seconds to think about whether or not you agree with this statement, and why (or why not). Then you will have one minute to tell us your opinion. Please be sure to give reasons for your opinion.

[directions given aurally (on audiotape) and in writing (in test booklet)]

EXAMPLE B
Prompt for Comparison-Contrast Task

You all know or have heard something about the way Christmas is celebrated in the U.S. Think of a major holiday in your country, and then compare and contrast this holiday with the Christmas holiday as it is celebrated in the U.S.

You will have 30 seconds to think about your answer. Then you will have one and a half minutes to answer the question.

[directions given aurally (on audiotape) and in writing (in test booklet)]

APPENDIX B
NARRATIVE TASK RATING SCALE

Functional ability/task rating

4 Task requirements: a) General orientation to characters and setting clear. b) Steps in story ordered and cohesive--natural, appropriate use of transition signals. c) Finishes story and/or concludes it appropriately. d) Some mention of personal attributes of characters and/or their emotional state.

3 May finish story but at least 1 task requirement is lacking.

2 Doesn't finish story and/or doesn't perform up to 2 of the task requirements.

1 Description of steps or pictures only with little attempt at connecting these as a "story"--2 or more of the task requirements missing.

Linguistic skills rating

4 No unnatural pauses, almost effortless and smooth although still perceptively non-native. Always intelligible. Only occasional, minor errors in grammar. Pronunciation always intelligible. Use of language precise, appropriate to task. Vocabulary misuse is rare.

3 Fairly smooth and effortless delivery. Few unnatural pauses. Grammatical errors are usually minor; don't interfere with overall intelligibility. Accent foreign, but rarely interferes with comprehension. Occasional misuse of vocabulary words, but clear and intelligible with little hesitation.

2 Occasionally halting and fragmentary, some unnatural pauses. Problems with basic grammatical constructions may sometimes interfere with intelligibility. May sometimes be hard to understand due to pronunciation problems. Limited vocabulary requires hesitation and circumlocution. Simple terms may be used, but these are usually adequate for task.

1 Very halting and fragmentary, many unnatural pauses. Little grammatical or syntactical control except in simple structures. Interferes with intelligibility and with apparent ability to complete task. Often hard to understand due to pronunciation problems. Vocabulary limited or inadequate for accomplishing tasks.

APPENDIX C
FSI-ADAPTED SCALE*

Fluency General criteria: Overall smoothness, continuity and naturalness of speech (as opposed to pauses for rephrasing sentences, groping for words and so forth).

5 Speech is almost effortless and smooth although still perceptively nonnative. No unnatural pauses.

4 Fairly smooth and effortless. Speech is occasionally hesitant but these unnatural pauses are rare.

3 Speech is occasionally halting and fragmentary. Some unnatural pauses.

2 Speech is slow and uneven except for short or routine sentences. Many unnatural pauses.

1 Speech is so halting and fragmentary that delivery is extremely labored. Strongly affects intelligibility of speech.

Grammar General criteria: Appropriateness of grammatical constructions to task. Intelligibility due to grammatical correctness or incorrectness of utterances.

5 Only occasional, minor errors with no patterns of failure. Always intelligible, constructions used are appropriate to task.

4 Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.

3 Some errors which show a lack of control with some major patterns. Causes occasional misunderstanding.

2 Frequent errors showing control of very few major patterns. Causes frequent problems with intelligibility.

1 Very little grammatical or syntactical control except in the simplest structures. Interferes with intelligibility and with apparent ability to complete task.

Pronunciation General criteria: Overall comprehensibility/intelligibility. Phonemic accuracy, "natural" intonation.

5 No conspicuous mispronunciations, but would not be taken for a native speaker. Intonation "natural."

4 Marked "foreign accent" and occasional mispronunciations which do not interfere with understanding.

3 "Foreign accent" may require some concentrated listening. Mispronunciations lead to occasional misunderstanding.

2 Frequent serious errors require concentrated listening. Very "heavy" accent leads to misunderstandings.

1 Pronunciation frequently unintelligible.

Vocabulary General criteria: Appropriateness of choice of words as opposed to a too-simple or inadequate vocabulary according to task requirements.

5 Use of language broad and precise, words always appropriate for task. Vocabulary adequate to cope with more difficult concepts.

4 Misuse of vocabulary words is rare but may occur. Usually clear and intelligible with little hesitation.

3 Choice of words sometimes inaccurate; simple terms are primarily used. Some evidence of hesitation and circumlocution due to limited vocabulary.

2 Vocabulary limited and choice of words often inaccurate. Clear evidence of circumlocution and hesitation, affects performance on task completion.

1 Vocabulary very limited and usually inadequate for accomplishing tasks.

(adapted from an FSI supplemental rating scale: Educational Testing Service, 1970; reprinted in Oller, 1979, pp. 321-323)

*Results claimed in this article using an adaptation of an ETS testing instrument should in no way be construed as confirming or denying the validity of the original test on which it was based, or as possessing any validity of the original test.

Anne Lazaraton, a doctoral student in Applied Linguistics at UCLA, has recently co-authored *The Research Manual: Design and Statistics for Applied*

Linguistics (Newbury House, 1991) with Evelyn Hatch. She has taught ESL and science in Tonga, and in the U.S. she has taught ESL, research design and statistics, and linguistics for language teachers. Her current research interests include conversational interaction in language interviews and oral skills assessment.

Heidi Riggenbach, an assistant professor in the University of Washington's M.A. TESL program, has taught ESL and EAP in China and the U.S., and was a Fulbright lecturer in an ESL teacher education program in Malaysia. Her teaching/research interests are conversation and discourse analysis, curriculum and materials development, and oral proficiency testing.