LANGUAGE
TESTING

A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors

Language Testing
27(2) 235–260

DOI: 10.1177/0265532209349469
http://ltj.sagepub.com
SAGE

India C. Plough
University of Michigan, USA

Sarah L. Briggs
University of Michigan, USA

Sarah Van Bonn
University of Michigan, USA

Abstract

The study reported here examined the evaluation criteria used to assess the proficiency and effectiveness of the language produced in an oral performance test of English conducted in an American university context. Empirical methods were used to analyze qualitatively and quantitatively transcriptions of the Oral English Tests (OET) of 44 prospective Graduate Student Instructors (GSI). The language required to complete the tasks on the test was conceptualized from the functional perspective of transactional and interactional language use as defined by Brown and Yule (1989). Listening comprehension and pronunciation were also analyzed and scored. Stepwise logistic regression was used to determine the extent to which these linguistic features contributed to final ratings. These quantitative findings were then compared to 'real-time' written comments made by evaluators during the tests. Intuitive methods were then used to further explore those features of candidate performance attended to by evaluators: interviews were conducted with experienced evaluators to determine the features they judged necessary for communicating effectively in instructional settings. Results indicate that the three data sources converge on two main features – pronunciation and listening comprehension – that are important in describing and evaluating the proficiency of prospective GSIs.

Keywords

evaluation criteria for speaking tests, oral performance testing, rating scale, rating scale validation, second language speaking assessment, second language speaking proficiency, teaching assistants

Corresponding author:
India C. Plough, Research Associate, University of Michigan, English Language Institute, 500 East Washington Street, Ann Arbor, MI 48104–2028, USA.
E-mail: indiac@umich.edu

In general terms, the development of a test of second language oral performance requires a definition of the construct, operationalization of that construct in the form of a test task or tasks, and, finally, a method of scoring performance on the test task(s). Each stage of this development process has generated rich discussion among researchers in the field. For example, to what extent should contextual factors be incorporated into the definition of the construct and/or the design of the test tasks (Chapelle, 1999; Douglas, 2000)? And then, to what extent can performance on these specific test tasks be generalized to other contexts? Relatively recent investigations (Brown, 2003 ; Fulcher & Márquez Reiter, 2003; Lumley & O'Sullivan, 2005; Nakatsuhara, 2008; O'Loughlin, 2002; Taguchi, 2007) have increased our understanding of the multitude of interacting variables that potentially impact ratings on oral performance tests. While debate continues over the nature of the speaking construct and valid methods of assessing speaking, there is virtually no disagreement over the potential for improving validity by developing rating scales from actual performance as opposed to deriving scales from a conception of the ideal performance.

The value of data-driven methods for developing rating scales was brought to the forefront in the 1990s by Upshur and Turner (1995, 1999), Chalhoub-Deville (1995), and Fulcher (1996). Upshur and Turner (1995) were among the first to respond to the need for improved reliability and validity in the rating scales that were increasingly being used in second language assessments. These researchers developed the well-known EBB, or empirically derived, binary-choice, boundary-definition scales. The key characteristics of these scales are as follows: (1) they are based on learner performance – not a theoretical description of ability; (2) raters use learner performance to make binary decisions that distinguish ability level boundaries; (3) these scales are developed for specific tasks so that an EBB scale developed for one speaking task cannot be used for a different speaking task; and (4) these scales are developed for specific populations. In their study, two EBB scales, one for communicative effectiveness and one for grammatical accuracy, were developed and used to assess the story retelling of 101 young learners. Two different raters assessed each student using the two EBB scales. Results did indeed indicate 'that high agreement among raters can be achieved … [with the use of] EBB scales' (p. 11).

Chalhoub-Deville (1995) also employed empirical methods to investigate the criteria used by raters in assessing L2 speaking ability. Her study consisted of six students of intermediate-level Arabic as a second language who performed three prototypical speaking tasks: interview, narration and read-aloud. Holistic scores were assigned to the 18 speech samples by different native speaker rater groups. Chalhoub-Deville derived three dimensions underlying the ratings across the three tasks: dimension one included 'grammar-pronunciation,' dimension two was represented by 'creativity in presenting information,' and dimension three was identified as 'amount of detail provided.' One conclusion drawn from the study was that all 'dimensions do not have the same interpretive meaning and value across all three tasks … For example, the "grammar-pronunciation" dimension can be thought of as relevant to all three tasks, while "creativity in presenting information" and "amount of detail provided" are more meaningful for the interview and the narration than the read-aloud task' (p. 27). Importantly, variation in learner performance based on test type, in addition to variation in evaluations based on rater group, led Chalhoub-Deville to argue for 'a research approach that derives scales empirically according to the given tests and audiences, and according to the purpose of assessment' (p. 28).

Fulcher's (1996) contribution to the research on scale development resides in his investigation of the definition and operationalization of the construct of fluency. Fulcher examined the oral interviews of 21 learners of English as a second language who had been rated using band descriptors of the English Language Testing Service (ELTS). Using a Grounded Theory approach, fluency features were derived and then categorized from the transcripts and the audio-recordings of learner language in the interviews. Fulcher then used discriminant analysis to investigate the extent to which the frequency counts of the features coded in the qualitative analysis could predict membership into the particular band level where a learner had been placed. Results indicated that 'the categories discriminate well between students' (p. 223); additionally, 95% (20 of 21) of students had been correctly classified based on the fluency categories developed. Fulcher 'tentatively suggested that a data-based approach to rating scale development appears promising, and that further research should be carried out into the description and operationalization of constructs for language testing' (p. 28).

Most recently, a study by Iwashita, Brown, McNamara, and O'Hagan (2008) provides insight into the componential nature of proficiency. The researchers examined 200 speech samples of the TOEFL iBT in order to determine differences in performance at each level and the features distinguishing those performances. The researchers analyzed approximately 6.5 minutes of speech from 40 participants for linguistic resources (grammatical accuracy, grammatical complexity, and vocabulary), phonology (pronunciation, intonation, rhythm), and fluency (filled and unfilled pauses, repair, total pausing time, speech rate, and mean length of run). Findings indicated that grammatical accuracy, vocabulary, pronunciation, and fluency influenced overall score. That is, features from each of the categories (linguistic resources, phonology, and fluency) affected the score and no single feature (e.g. pronunciation) within a category determined the overall score. In addition to increasing our understanding of the 'composition' of various levels of proficiency, these results are informative not only for validating the iBT speaking scale but also for developing future rating scales.

The current investigation complements and builds on the aforementioned research in a number of ways. First, while prior studies used empirical methods to examine the nature of proficiency and to develop rating scales, the current study applies empirical methods to investigate the meaningfulness of the evaluation criteria, or features of the language sample to be rated, used in an operational scale to assess speaking proficiency. Second, the scoring procedures differ; the ratings analyzed in this study have been given in 'real-time.' Third, the contexts differ; examined here are the oral performances of a high-stakes test of English administered to prospective graduate student instructors (GSIs)[1] in an American university. Fourth, the speech samples analyzed in this study include both monologic and interactive speech events, which had been included in only some of the previous studies. Finally, the performance features under investigation in this study are conceptualized somewhat differently, as described below, and extend beyond those features attended to in the research summarized above.

The current study explores the dynamic and relativistic relationship that exists between the different components (e.g. phonological, lexical, pragmatic) of the speaking construct as represented in the evaluation criteria of a holistic rating scale. As explained in greater detail below, the language that is elicited and required by the test tasks (and,

thus, forms the basis for the evaluation criteria used for the test) has been conceptualized in functional terms as transactional and/or interactional (Brown & Yule, 1989). Motivated by observations that an evaluator's final rating is influenced, in part, by the task, the rater, and the presence or absence of various linguistic components, the study asks:

1. Which feature(s), or combination of performance features, is/are significant predictor(s) of approval for duties as a Graduate Student Instructor?
2. What is the relative weight of each feature in predicting approval?
3. Are these significant features and/or combination of features similar across disciplines in predicting rating?
4. To what extent do the qualitative and quantitative data converge and correspond with those features used in the evaluation of speaking proficiency?

Method

Materials: The Graduate Student Instructor Oral English Test (GSI OET)

The GSI OET (Briggs, 1987, 2003) is required for any prospective graduate student instructor who did not receive an undergraduate degree from an institution in which English is the language of instruction. In order to be eligible for a teaching assistantship, graduate students must pass the GSI OET. Thus, the domain of behavior relevant to the test design is language use relevant to the instructional settings and duties in the department where the candidate might be working.

Reflected in the design, scoring, and validation procedures of the GSI OET is an interactionalist perspective of second language performance, which draws on applications of sociocultural theory and systemic functional theory to second language acquisition research and assessment (Brown, 2003; Chapelle, 1998; Chalhoub-Deville, 2003; Lantolf, 2000; Lantolf & Appel, 1994). Within these frameworks, 'performance is jointly constructed and distributed across the participants. Dialogues construct cognitive and strategic processes which in turn construct student performance, information that may be invaluable in validating inferences drawn from test scores' (Swain, 2001, p. 275). Based in part on Halliday and Hasan (1989), an interactionalist view maintains that learner factors (e.g. world knowledge, language knowledge, 'fundamental' processes) and contextual factors (field, tenor, mode) interact to construct and influence performance. Test construction and revision, therefore, has involved ongoing observations of target language use situations, continual examination of the tasks that comprise the test, and analysis of test participants in terms of the discourse and language features elicited and produced.

The GSI OET is an oral performance test that evaluates proficiency at the high-intermediate to advanced level. In addition to assessing general level of language ability, the tasks and evaluation criteria of the GSI OET are used as diagnostic tools to identify strengths and weaknesses of candidates. Using both direct and indirect measures, the test consists of four tasks that require both one-on-one and group interactions (see *Format of the Test* below).Three evaluators participate in the test: two staff members of the Testing

and Certification Division of the English Language Institute and a faculty member from the academic department in which the prospective GSI would be working. Once a test is finished, the candidate leaves the room and each evaluator rates the candidate independently. Evaluators then discuss the linguistic strengths and weaknesses of the candidate to reach a consensus rating.

Target language use situations GSI teaching appointments include a wide range of contexts. GSIs may be responsible for teaching sections of introductory undergraduate courses, guiding small discussion sections that are part of large lecture courses, conducting lab sections of large lecture courses, conducting review and/or study sessions, and/or holding office hours. GSIs may interact with students whose level of education ranges from freshmen to graduate. Topics may include structured, prescribed and preplanned subject matter and impromptu problem resolution, which means GSIs must be effective in both transactional and interactional language use (see operational definitions below).

In terms of transactional competence, the ability to sequentially develop a topic or explain a procedure is crucial. GSIs also provide multiple examples of concepts, rephrase explanations and questions to students, and summarize content.

In terms of interactional competence, in those classes where the GSI actively engages students in small talk, uses humor, and is attentive and responsive to questions from students, a relaxed atmosphere and good rapport also characterize the class. The ability to negotiate, hedge, and use softeners is apparent when GSIs are discussing grades with students.

Format of the test The test takes approximately 20 to 30 minutes and, as shown in Table 1, is composed of four tasks: General Interview, Lesson Presentation, Office Hour Role Play, and Video Questions. The tasks are designed to represent the multiple contexts in which GSIs may interact.

The General Interview lasts approximately 5 minutes, serves as a warm-up, and allows all participants to get a sense of each other and the setting. During the Interview, the prospective GSI is asked a few general questions about his or her background and educational interests.

The Lesson serves as a direct measure of primarily transactional language use. During the Lesson, which takes approximately 10 minutes, the prospective GSI teaches a topic or concept that he/she has chosen in advance. Candidates are instructed to choose topics that are suitable for undergraduates in an introductory level class or lab in the particular department in which he/she may be teaching. Evaluators ask impromptu questions during the Lesson.

The Office Hour Role Play takes approximately 5 minutes and serves as a direct measure of a candidate's interactional language skills. One of the evaluators plays the role of a student visiting the GSI during office hours. The 'student' seeks advice or guidance about issues related to administrative matters or personal academic problems. The other two evaluators do not participate in this segment of the test.

The Video Questions take approximately 5 minutes and serve as a semi-direct measure of listening comprehension. Ten different, minimally contextualized questions are posed by expert speakers of varieties of English common in the university undergraduate population. Questions have been compiled from interviews with undergraduates and with GSIs

Table 1. Graduate Student Instructor Oral English Test

Task	Type	Primary focus	Time
1. General Interview	Direct	Warm-up Interactional competence	5–7 minutes
2. Lesson Presentation	Direct	Transactional competence	10 minutes
3. Office Hour Role Play	Direct	Interactional competence	3–5 minutes
4. Video Questions	Indirect	Listening comprehension	5 minutes

as well as from the MICASE (Michigan Corpus of Academic Spoken English) data base. Test questions, which are not repeated, represent typical concerns of undergraduate students (e.g. 'How much is the final exam worth?'); candidates are instructed to respond by creating or making up an answer that makes sense for the question asked. The video/DVD is played and then paused, during which time the candidate responds to the monitor.

Evaluation criteria and rating scale The evaluation criteria were originally developed from the evaluators' comments written during the tests of 350 candidates. Recall that each test is evaluated by three individuals. Thus, approximately 1050 comments were compiled and categorized to arrive at the evaluation criteria. These criteria include pronunciation, which includes fluency and intelligibility (e.g. pausing/hesitation, prosody, articulation, and voice projection); vocabulary (e.g. colloquial phrasing and idiomatic expressions, field-specific terminology, rhetorical expressions); grammar (e.g. syntax and morphology); interactional competence (e.g. responsiveness, ability to provide suggestions, and active listening); transactional competence (e.g. topic development, coherence; and paraphrasing ability); and, finally, aural comprehension.

Based on the empirically derived criteria and the rank ordering of candidate performances, a holistic 5-point rating scale (Appendix 1) was created. Inasmuch as the contexts for language use differ significantly for candidates in the College of Engineering and for those in the College of Literature, Science, and the Arts (LSA), College of Engineering (CoE) candidates are approved for teaching duties with a rating of 4− or higher, while LSA candidates must receive a rating of 4 or higher. As noted by Briggs (1994, pp. 78–79), 'using level descriptors is indeed problematic. The communicative ability necessary for working in an inorganic chemistry lab is different than that needed to explain problems in an introductory microeconomics recitation section .... The individual profile of one TA [GSI] rated 4−, the minimal approval rating [in Engineering], may differ significantly from that of another rated 4 in another discipline. Thus general descriptions of test performance appear to serve only as guidelines, not as strict criteria.' The goal of the current study is to examine these evaluation criteria in order to obtain a more accurate understanding of their contribution to varying levels of proficiency and, in turn, to the speaking construct.

Operational definitions

For the extant study, the tasks on the test were examined from the functional perspective of transactional and interactional language use as defined by Brown and Yule (1989).

Transactional is that function which language serves in the expression of content. Transactional language 'is used to convey "factual or propositional information."].... [It] is primarily "message oriented." It is important that the recipient gets the informative detail correct' (Brown & Yule, 1989, p. 2). Interactional is that function which language serves in expressing social relations and personal attitudes and in establishing and maintaining human relationships.

In order to investigate empirically the effects of transactional and interactional language on the overall rating of a candidate's proficiency, the specific features of the language used to fulfill these two functions must be identified. With respect to transactional language use, we examined the discourse in terms of organization and cohesion. Regarding the former, we turned to van Dijk's (1977, p. 166) suggestion that items will be ordered based on 'perceptual salience so that the more salient entity will be mentioned first' (cited in Brown & Yule, 1989, p. 145). According to this view, the general, for example, will be mentioned before particulars; whole before part; large before small. Levelt (1981, p. 94) 'suggests that by adopting the stereotypical pattern of the culture "the speaker facilitates the listener's comprehension" since both speaker and hearer share the same stereotype' (cited in Brown & Yule, 1989, p. 145). For text/discourse cohesion, we took a 'strong' interpretation of the approach proposed by Halliday and Hasan (1976) and assumed that it is the explicit expression – through cohesive devices – that bind a (spoken) text together and force co-interpretation. Included in transactional language use are features that are traditionally classified under grammar, such as tense and aspect.

In terms of interactional language use, utterances that encouraged a speaker to expand or elaborate on a given topic, or mitigated the apparent rigidity or hostility of an utterance through syntactic downgraders such as durative aspect, past tense, hedges, or modals (Blum-Kulka, House, & Kasper, 1989; Beebe & Waring, 2004), served various functions in maintaining and promoting the dialogue, which, in turn, established social relationships. Form alone, however, did not define these features. Therefore, the content of the utterance had to be examined. A list of the transactional and interactional features that were coded appears in Appendix 2.

Test tasks were also scored separately for listening comprehension and pronunciation. It could be argued that pronunciation should be included as a feature of transactional and of interactional language use. After all, if the words cannot be understood, the information or content material cannot be conveyed/received, and rapport between interlocutors cannot be established. However, given that one of the goals of the current study is to determine the relative importance of the various components of oral proficiency, pronunciation was treated as a separate ability and measured independently of the transactional and interactional features.

To normalize the data, Analysis of Speech Units (AS-Unit), which are defined as 'a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either' (Foster, Tonkyn, & Wigglesworth, 2000, p. 365), were used. Foster et al. provide detailed explanations for coding AS-Units, which incorporate features of spoken discourse that are key in determining a speaker's proficiency and yet heretofore have not been included into a single unit of analysis. That is, some units (e.g. idea unit, c-unit) focus mainly on semantics while others (e.g. tone unit, phonemic clause) focus primarily on intonation; still others

are syntactically driven (e.g. s-node, t-unit). The AS-Unit is syntactically based but also integrates the use of intonation and pausing in determining AS-Unit boundaries. Total clauses and total AS-Units, along with instances of transactional and interactional language, were calculated for each transcript.

The data

The data for this study consist of the following: 1) the GSI OETs of 44 candidates who took the test between May 2002 and September 2006; 2) the 'real time' written evaluations of candidate performance; and 3) interviews with experienced evaluations. A quantitative examination of the GSI OETs is presented first, followed by the qualitative analyses of the written commentary and the interviews.

Video- and audiotaped recordings of the GSI OETs were transcribed by four research assistants; each transcription was then checked for accuracy by a second transcriptionist. Table 2 provides a summary of learner variables and approval status. The first languages and degree programs of candidates were taken into account so that the data set would form a representative sample of the entire population of OET candidates. Eighty-four percent of the candidates were native speakers of either Korean or Chinese. Twenty-three candidates came from the College of Literature Science and Arts (LSA); 21 candidates came from the College of Engineering. There were approximately twice as many males as females (30M and 14F). Approximately the same number of LSA candidates were approved (10) and not approved (13) for teaching; three times more Engineering candidates were approved for teaching (16) than not approved (5).

The current study focused on the Lesson Presentation (Task 2), the Office Hour Role Play (Task 3), and the Video Questions (Task 4).[2] To summarize, the Lesson Presentation primarily involves transactional language use. This is not to say that interactional language use is not called upon. In fact, in acknowledging and responding to 'student' questions, interactional language use is key. However, the primary purpose of this portion of the GSI OET is to serve as a direct measure of a candidate's ability to convey information in a comprehensible and organized manner. The Office Hour Role Play serves as a direct measure of primarily interactional language use. Again, transactional language use is needed, but the task is intended to focus on the candidate's responsiveness to the 'student' and the candidate's ability to understand intended or implied meanings. Finally, the Video Questions represent a semi- direct measure of both transactional and interactional

Table 2. Candidate profile summary

		LSA		Engineering	
L1		Approved	Not Approved	Approved	Not Approved
Chinese	23	8	8	7	0
Korean	14	2	4	3	5
Other	7	0	1	6	0
TOTAL	44	10	13	16	5

language use. Speaking time (in minutes) for the Lesson Presentation and the Office Hour Role Play of each candidate ranged from 7:33[3] to 21:45 with a median of 11:49 and a mean speaking time of 11:47. A complete list of participants and related variables can be found in Appendix 3.

Coding protocol

Uniform procedures for coding each of the transcripts for the features described above were established through consensus among three researchers. Once inter-rater reliability (100% agreement) was established, each researcher took responsibility for coding the transcripts for particular features. That is, one researcher coded all the transcripts for AS-Units, another researcher then coded all of these transcripts for transactional and interactional language use, and the third researcher scored the audio portions of the tests for pronunciation and listening comprehension. If any questions of coding or scoring arose during the process, the researchers discussed the feature and reached a consensus.[4]

The Lesson Presentation and the Office Hour Role Play tasks were coded for AS-Units, transactional and interactional language use, and pronunciation; the Video Questions were scored for listening comprehension. Pronunciation was scored using the integrated speaking rubric created for the iBT/Next Generation TOEFL Test. Pronunciation scores range from 0–4. The Video Questions were used as a measure of listening comprehension. Of course, listening comprehension is an ability used throughout the test. By using this task alone as an indicator of listening comprehension, we are assuming that a candidate's 'listening score' on this task would be consistent on all the tasks. There is justification for this assumption: candidate performance on the Video Questions has consistently confirmed evaluations of listening comprehension made prior to this task. Additionally, no discrepancies between listening comprehension on this task and on the other tasks were identified during the analyses of the tests. For the purposes of this study, we assigned each response a score of 0, 0.5, or 1. This score was determined based on the appropriateness of the response and on the length of the response, which had to be expanded sufficiently in order to indicate that the candidate clearly understood the question posed. Responses that were expanded and appropriate received a score of 1. If a response was appropriate but short, the candidate received a score of 0.5; similarly, if a response was expanded but inappropriate, the candidate received a score of 0.5.

Task totals for these features were then combined so that each candidate obtained scores for (1) Total clauses/AS unit, (2) Total transactional language use/AS unit, (3) Total interactional language use/AS unit, (4) Pronunciation, and (5) Listening comprehension.

Results

We first present the quantitative analyses, which include both descriptive and inferential statistics. Inferential statistics are presented in response to each of the study's research questions. The qualitative analyses follow, including the written comments of evaluators during the test and commentary provided by evaluators during interview sessions. The small sample size of this study must be highlighted at the outset of the presentation of

these analyses. Interpretations and generalizations should be viewed cautiously. Nonetheless, certain patterns and/or trends can be observed. The research questions are repeated here for convenience:

1. Which feature(s), or combination of performance features, is/are significant predictor(s) of approval?
2. What is the relative weight of each feature in predicting approval?
3. Are these significant features and/or combination of features similar across disciplines in predicting rating?
4. To what extent do the qualitative and quantitative data converge and correspond with those features currently used to evaluate speaking proficiency?

Quantitative results

Descriptive statistics Descriptive statistics were compiled for all independent variables, which are listed below.

- Total clauses/analysis of speech (AS) unit
- Total transactional language use/AS unit
- Total interactional language use/AS unit
- Pronunciation
- Listening comprehension.

A comparison of the difference in means of each variable between the Approved group and the Not Approved group appears in Table 3. As can be seen, the overall mean of each variable is higher for the group approved for teaching than for the group that was not.

The Independent Samples T Test is shown in Table 4. Levene's Test indicates that the Approved and Not Approved groups have approximately equal variance on each variable. The t-test indicates that, with the exception of total interactional language use per AS Unit, which is approaching significance ($p = 0.055$), the difference between the means of the two groups is statistically significant for each variable: listening comprehension ($t = 6.264$, $df = 42$, $p < 0.05$), pronunciation ($t = 5.956$; $df = 42$, $p < 0.05$), total clauses

Table 3. Means for each variable, grouped by approval

Pass Fail	Total Clause Per AS	Total Trans Per AS	Total Inter Per AS	Listen.	Pron.
F					
Mean	1.70	0.93	0.42	0.55	2.11
N	18	18	18	18	18
SD	0.52	0.42	0.17	0.21	0.58
P					
Mean	2.04	1.5	0.52	0.87	3.27
N	26	26	26	26	26
SD	0.44	0.53	0.18	0.14	0.68

Table 4. Independent Samples T Test

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
LISTEN.	3.11	0.09	6.26	42	0.000	0.33	0.05	0.22	0.43
PRON.	2.12	0.15	5.96	42	0.000	1.16	0.19	0.77	1.55
TOTAL CLAUSE.	0.07	0.80	2.39	42	0.021	0.35	0.15	0.06	0.64
TOTAL TRANS.	1.29	0.26	3.45	42	0.001	0.52	0.15	0.22	0.82
TOTAL INTER.	0.59	0.45	1.98	42	0.055	0.11	0.05	–0.01	0.21

Equal variances assumed.

Table 5. Means for each variable, grouped by college

College	Total Clause Per AS	Total Trans. Per AS	Total Inter. Per AS	Listen.	Pron.
LSA					
Mean	1.81	1.07	0.50	0.67	2.78
N	23	23	23	23	23
SD	0.51	0.49	0.21	0.25	0.90
COE					
Mean	2.00	1.43	0.46	0.81	2.81
N	21	21	21	21	21
SD	0.48	0.55	0.14	0.19	0.814
Total					
Mean	1.90	1.24	0.48	0.74	2.80
N	44	44	44	44	44
SD	0.50	0.55	0.18	0.23	0.85

per AS unit ($t = 2.394$, $df = 42$, $p < 0.05$), and total transactional language use per AS unit ($t = 3.452$, $df = 42$, $p < 0.05$).

Table 5 presents a comparison of the difference in means of each variable between the College of Engineering (COE) and the College of Literature, Science, and the Arts (LSA). As can be seen, with the exception of interactional language use, the mean of each variable is higher for Engineering candidates than for those in LSA.

Results of the Independent Samples T Test are provided in Table 6. Levene's Test indicates that there is a significant difference ($p < 0.05$) on the variance of listening comprehension for the two colleges; that is, the two groups do not have approximately equal variance on the dependent variable as they should. Therefore, the t-test for equality of means is not considered for listening comprehension. With respect to the other variables, results of the t-test for equality of means show that the difference between the means of the two colleges is statistically significant solely for Total transactional language use ($t = 2.304$, $df = 42$, $p < 0.026$).

Inferential statistics: Evaluations of the Logistic Regression Model As no a priori assumptions had been made regarding the relationships between variables, stepwise logistic regression[5] was used to explore the independent variables and their significance as predictors of the dependent variable (Approval). Before addressing the specific research questions, we first present the necessary evaluations of the logistic regression model, which include tests of the null hypothesis, measurements of goodness of fit, and assessment of the predicted probabilities. Table 7 presents results of tests of the null hypothesis. As shown, the likelihood ratio test was equal to 40.49, which corresponds to a p-value of less than 0.0001; thus, the model with the independent variables is more effective than the null model.

Table 8 presents two descriptive measures of goodness-of-fit. The closer the values of R^2 are to 1, the better the fit of the model. As can be seen, based on the Max rescaled R-Square, predictors in the model account for 81% (Approval), 70% (Rating-LSA) and 80% (Rating-Engineering) of the variation in the dependent variables.

Table 6. Independent Sample T Test

	Levene's Test for Equality of Variances		T-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
TOTAL CLAUSE PER AS	1.04	0.31	1.26	42	0.22	0.18	0.15	–0.11	0.49
TOTAL TRANS. PER AS	0.66	0.42	2.30	42	0.03	0.36	0.16	0.05	0.68
TOTAL INTER. PER AS	1.11	0.30	–0.75	42	0.46	–0.04	0.05	–0.15	0.07
LISTEN.	4.22	0.05	2.10	42	0.04	0.14	0.07	0.01	0.28
PRON.	0.89	0.35	0.10	42	0.92	0.03	0.26	–0.50	0.55

Equal variances assumed.

Table 7. Test of null hypothesis

Null Hypothesis: Beta Ø	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.49	5	<0.0001
Score	26.33	5	<0.0001
Wald	6.24	5	0.28

Table 8. Measurements of goodness-of-fit

Dependent Variable	Cox and Snell R-Square	Nagelkerke (Max rescaled R-Square)
Approval	0.60	0.81
Rating-LSA	0.67	0.70
Rating-Engineering	0.76	0.80

Table 9. Validations of predicted probabilities

Dependent Variable	Percent Concordant	Percent Discordant	Percent Tied	Pairs	Somers' *D*	Gamma	Tau-*a*	*c*
Approval	96.2	3.8	0.0	468	0.92	0.92	0.46	0.96
Rating-LSA	87.6	11.9	0.5	202	0.76	0.76	0.61	0.88
Rating-Engin	92.8	7.2	0.0	167	0.86	0.86	0.68	0.93

Associations of predicted probabilities and observed responses for each dependent variable (approval, rating) indicate that the model correctly predicts higher probabilities (i.e. *c* statistic or percent concordant, which should be close to 100) for observations with the event outcome than the probability for non-event outcomes. Measures of association are presented in Table 9, Validations of Predicted Probabilities.

In summary, results presented in Tables 7 through 9 indicate that the null hypotheses can be rejected and that there is good model–data fit for all models.

Logistic regression Stepwise regression was used to explore the independent variables and their significance as predictors of the dependent variable (Approval). The summary of the stepwise selection presented in Table 10 indicates that listening ($p < 0.0001$) and pronunciation ($p < 0.0072$) are significant predictors of Approval. Thus, in response to Research Question 1 (Which feature(s), or combination of performance features, is/are significant predictor(s) of approval?), when all candidates are considered together, listening comprehension and pronunciation are significant predictors of being approved for a GSI appointment.

Given that Listening scores ranged from 0.15 to 1.0 and pronunciation scores ranged from 1.0 to 4.0, the Odds Ratio Estimates provided in Table 11 were calculated for a 0.1 unit change in the score in order to address Research Question 2: What is the relative weight of each feature in predicting approval? As can be seen, if other variables are held

Table 10. Summary of stepwise selection (approval as dependent variable)

Effect		Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
Step	Entered							
1	LISTEN.		1	1	21.25		<0.0001	LISTEN.
2	PRON.		1	2	7.21		0.0072	PRON.
3	TRANS.		1	3	6.04		0.0140	TRANS.
4		TRANS.	1	2		3.80	0.0512	TRANS.

Somers' *D* = 0.878; Gamma = 0.884; Tau-*a* = 0.434; *c* = 0.939

Table 11. Odds ratio estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Listen	2.01	1.15	3.50
Pron	1.25	1.05	1.50

Table 12. Classification table

	Observed	
	1	2
Predicted		
1	16 (36.4%)	4 (9.1%)
2	2 (4.5%)	22 (50.0%)

constant, a 0.1 unit increase in the Listening score results in the odds of Approval increasing by 2.01. Additionally, if other variables are held constant, a 0.1 unit increase in the pronunciation score results in the odds of Approval increasing by 1.25.

The classification results presented in Table 12 are the predicted group membership (Approval [1] or Non-Approval [2]). Using a probability cut-point of 50%, 86.4% of the cases are classified correctly; the false positive rate is 4.5% and the false negative rate is 9.1%.

A stepwise regression was also run to investigate which variables are significant in predicting rating for each College (Research Question 3). The summary of this stepwise regression is presented in Table 13. As can be seen, in LSA, pronunciation ($p < 0.0001$) is a significant variable in predicting rating. In the College of Engineering, both pronunciation ($p < 0.0004$) and interactional language use ($p < 0.0182$) are significant variables in predicting rating.[6]

Table 13. Summary of stepwise selection (Rating as dependent variable)

LSA								
Effect		Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
Step	Entered							
1	PRON.		1	1	15.19		<0.0001	PRON.
Somers' *D* = 0.693; Gamma = 0.897; Tau-*a* = 0.553; *c* = 0.847								
ENGINEERING								
1	PRON.		1	1	12.56		0.0004	PRON.
2	INTERACT.		1	2	5.57		0.0182	INTERACT.
Somers' *D* = 0.760; Gamma = 0.760; Tau-*a* = 0.605; *c* = 0.88								

Quantitative results: Summary In summary, the descriptive statistics indicate that there is a significant difference between the means of those candidates who were approved for teaching and those candidates who were not approved for all variables except interactional language use. When examined by college, the mean of transactional language use is significantly higher in the College of Engineering.

When all candidates are considered together, the regression analyses indicate that listening and pronunciation are significant predictors of Approval and that the model correctly classifies 86.4% of the cases. Of the two, the listening score appears to have greater influence on the odds of approval than the pronunciation score. When the data are grouped by College, and rating (rather than Approval) is considered the dependent variable, pronunciation is a significant predictor of rating in LSA; both pronunciation and interactional language use are significant predictors of rating in Engineering.

Qualitative results

OET reports: Evaluator comments in 'real time' Now that we have seen the quantitative data for Research Questions 1–3, we turn to the qualitative analyses in order to address Research Question 4: To what extent do the qualitative and quantitative data converge and correspond to those features currently used to evaluate speaking proficiency?

For the 44 OET candidates whose test performances were analyzed for this study, the 'real time' written comments provided by the three evaluators were reviewed for positive and negative descriptions of candidates' pronunciation, listening comprehension, and use of those features that have been categorized as transactional or interactional language. These qualitative characterizations were then compared to the quantitative measures obtained from the analyses of the transcripts to determine (1) whether the qualitative description is consistent with the quantitative measures (i.e. transactional language use, interactional language use, listening comprehension, and pronunciation) obtained for that candidate[7] and (2) whether the qualitative description includes features beyond those taken into account in the quantitative analyses.

A comparison of the quantitative and qualitative measures of the 44 candidates indicates that evaluator comments were consistent with the quantitative measures for 36 (82%) of the candidates, both in terms of the particular features commented on and measured, and in terms of the general assessment of those features. Evaluators' written remarks on eight candidates, however, did not match the quantitative measures that had been tallied from the transcripts of those candidates.

Specifically, the commentary for five of the eight candidates turned out to be more favorable on one or two of the features (for each candidate) than the 'scores' on these measures would indicate. On four of the five, descriptions and 'scores' for transactional and/or interactional language use were the most inconsistent. For example, one evaluator commented that a candidate's lesson was 'well organized' and that the candidate was 'very interactive' during both the Lesson and the Office Hour Role Play; however, this candidate's scores on transactional language use and interactional language use were below the group mean. On the fifth test, the description of the candidate's pronunciation ('clear, deliberate, and comprehensible') did not correspond to the 'score' given.

In contrast, commentary on the reports of the remaining three candidates was generally less favorable on one or two of the features (for each candidate) than were the quantitative measures. For all three of the candidates, again, descriptions of transactional language use (e.g. 'poor organizational structure') and/or interactional language use (e.g. 'not attentive to listeners') were not consistent with the 'scores' given, which were above the group means. On one report, the description of the candidate's listening comprehension was not consistent with the 'score' given.

Turning to a comparison of the overall ratings (i.e. Approved/Not Approved) and the evaluators' comments for these eight candidates, interestingly, the evaluators' descriptions of the language of six candidates are consistent with the ratings those candidates received. The reports of only two candidates did not correspond with their final ratings: the reports contained positive descriptions and yet those two candidates were not approved for teaching.

Except for the eight mismatches just discussed, the evaluators' qualitative descriptions of the language of all the candidates are consistent with the quantitative scores that were tallied. As mentioned above, the features that show inconsistency are mainly transactional and interactional language use. This observation, in conjunction with the results of the regression analyses showing that transactional and interactional language use are not significant predictors of approval, indicates that the meaningfulness of the features and/or the operational definitions and descriptions of these two functions must be reviewed.

The 'real time' comments made by evaluators were also examined to determine whether the qualitative descriptions included features beyond those taken into account in the quantitative analyses of the current study. Evaluators highlighted expanded responses, lexical range, and grammatical errors as significant. To a certain extent, these linguistic and discourse features are included in the functional categories (i.e. transactional and interactional language) that were created for this study. For example, recall that appropriate tense/aspect and cohesive devices were counted as instances of transactional language use, as they contribute to a coherent explanation and to conveying information clearly. However, given their saliency to evaluators, these features may represent discrete components of the speaking construct and,

therefore, need to be explicitly included in the descriptions of proficiency at each level. Although these features cannot be quantified, their general frequency, based on evaluators' commentary, can be explicitly described for each rating level. Examples of evaluator comments on these features include: 'ability to paraphrase is weak,' 'needs to use cohesive devices to provide organization,' 'noticeable gaps in vocabulary,' and 'clear, coherent explanations.'

Interviews with experienced evaluators: Summary To complement the quantitative analysis of OET transcripts and the qualitative analysis of evaluators 'real time' commentary, intuitive methodology was employed. Five experienced faculty evaluators from Mathematics, Molecular Cellular and Developmental Biology, Ecology and Evolutionary Biology, Chemistry, and Economics were interviewed to discover: (1) the features they focus on when evaluating prospective GSIs; (2) the features that distinguish a candidate who is approved for teaching duties from a candidate who is not approved for teaching duties; and (3) whether the evaluators place more weight on any of the four tasks that constitute the OET.

In general, faculty evaluators agree on the features they primarily focus on when evaluating prospective GSIs: listening comprehension, pronunciation, and responding to questions. Responses of the experienced faculty evaluators correspond in significant ways with the results of the logistic regression, in which the listening comprehension and pronunciation are significant predictors of approval for GSI duties. Evaluators also noted that lack of interactional ability would not 'fail' a candidate, but a borderline candidate showing strength in interactional ability could be pushed from Not Approved to Approved. Finally, although no evaluator weights one task more than another, some differ slightly on the information that they glean from a particular task. For example, one evaluator finds Task 1 (General Interview) to be a good predictor of the outcome; another uses Task 4 (Video Questions) as a discriminator. More than suggesting additional features that should be included in the level descriptors or features that should be excluded, these interviews validate the use of current evaluation criteria.

Discussion and Conclusion

To summarize the results of this study:

RQ#1: Which feature(s), or combination of features, is/are significant predictor(s) of approval?

Results of the logistic regression analyses indicate that listening comprehension and pronunciation are significant predictors of being approved for a GSI appointment. The model correctly classified 86.4% of the cases. Recall the findings reached by Chalhoub-Deville (1995) and Iwashita et al. (2008) indicating that a combination of

features influence ratings and these may vary by task type and rater. That is, Chalhoub-Deville's (1995) results indicated that 'amount of detail' provided, for example, was a meaningful feature in certain tasks (e.g. narrative) that were evaluated by particular raters (e.g. different native speaker groups). On the other hand, Iwashita et al. (2008) concluded that grammatical accuracy, vocabulary, pronunciation, and fluency all influenced overall score on the iBT. In the study presented here, it appears that only listening comprehension and pronunciation are significant predictors of approval; importantly, they are meaningful features on *this* specific test for *these* particular evaluators. However, it is notable that findings of the current study support those of the Iwashita et al. study: both indicate that pronunciation[8] is a key factor in the evaluation of proficiency level.

RQ#2: What is the relative weight of each feature in predicting approval?

Results of the logistic regression analyses indicate that listening comprehension has more of an impact on the likelihood of being approved than does pronunciation. This finding may be somewhat surprising, especially because the latter skill has traditionally received somewhat more attention from the field of graduate student instructor testing and training. However, the importance of listening comprehension and the need for GSIs to be able 'to follow students' often complex and confusing questions' (Myers, 1994, p. 100) has continually been highlighted in the research in this area as well. Studies have found the frequency of student questions in the discourse of many GSI-conducted courses (e.g. lab sessions) to be relatively high (Rounds, 1990 cited in Madden & Myers, 1994) and that one of the most common complaints of undergraduate students is the inability of their GSI to understand their questions (Plakans, 1997). Indeed, the development of effective listening can now be found as one of the top five priorities of virtually all GSI training courses (Huang, 2005).

It is also informative to consider features that 'fall out' of a logistic regression model. In the present case, total clauses, which can be used as an indicator of linguistic complexity (Bygate, 1999; Foster, Tonkyn, & Wigglesworth, 2000; Skehan, 2001), did not appear to be a significant variable in the current study. This result is similar to the findings of Iwashita et al. that indicate linguistic complexity is not a significant feature that distinguishes different proficiency levels. This finding is not surprising in the current context given that teachers often use short, simple sentences in order to reduce the processing demands placed on students and to facilitate comprehension.

RQ#3: Are these significant features and/or combination of features similar across disciplines in predicting rating?

Results of the logistic regression analyses indicate that there is a difference between the colleges of Engineering and LSA in terms of those features that predict approval. In LSA, pronunciation is a significant predictor of rating; in the College of

Engineering, pronunciation and interactional language use are significant predictors of rating. That there are differences in the discourse of various disciplines is well-accepted. Is it the case then that the difference in the language of candidates at various levels of proficiency and from various disciplines can be meaningfully described in terms of transactional and interactional language? And, if so, what are the implications for establishing equitable evaluation criteria to assess proficiency across disciplines in an academic setting? These are certainly empirical questions that can be addressed by subsequent research. It is worth noting that the difference in means of transactional language use prompted a closer comparison of the two colleges, though we must keep in mind the small n-sizes. A comparison of the means for each feature of only the approved groups revealed no significant difference between the two colleges. This is certainly a welcomed trend, indicating that the evaluation criteria are being applied similarly by evaluators across disciplines.

RQ#4: To what extent do the qualitative and quantitative data converge and correspond with those features currently used to evaluate speaking proficiency?

In order to address Research Question 4, we must note that (1) the 'Real Time' evaluator commentary was consistent with the quantitative measures for 36 (82%) of the 44 candidates, and (2) interviews with experienced faculty evaluators confirmed that listening comprehension and pronunciation are salient features used in assessing candidate performance. In other words, the three data sources converge on two features that are of primary importance in representing and evaluating the proficiency of prospective GSIs: pronunciation and listening comprehension. Equally important is the fact that there was no disagreement or divergence among the data sources.

While the qualitative measures and the inferential statistics confirm the significance of pronunciation and listening comprehension as indicators of proficiency, a more complete response to Research Question 4 can be provided by returning to the results of the descriptive statistics of the comparison of means. The fact that there is a significant difference in transactional language use between the means of those candidates who were approved for teaching and those who were not approved suggests that this variable may also be a useful indicator of oral proficiency, as defined for current purposes. The importance of transactional language use in determining the critical cut-point decision must not be underestimated. Framing learner production in terms of transactional language use (e.g. 'use of grammar and vocabulary to provide coherent explanations') may be a more accurate description of learner performance on this particular test, for this specific purpose, and thus provide a useful measure for evaluators, test takers, and test users.

While this empirical study has shown us ways in which we can revise the OET rating scale in terms of the descriptions of evaluation criteria at different levels of proficiency, there are several drawbacks to the investigation. As mentioned, the number of

candidates is small, which prevents certain analyses and limits the discussion to tentative trends. A comparison of those LSA candidates who received a 4− with those Engineering candidates who received a rating of 4− is critical in order to determine the relative contributions of different features at various levels of proficiency. Unfortunately, the current sample size of those receiving a 4− is not large enough to perform inferential statistics. Additionally, since the range of proficiency levels examined in this study is very narrow, no conclusions or generalizations can be made regarding features that are unique to particular levels of proficiency. That is, features that may be significant at 'low-intermediate' or 'beginner' levels of proficiency were not revealed. Ultimately one would want to define levels of proficiency in terms of degree of mastery of these features.

The quantitative and qualitative findings of the current study, however, do make meaningful contributions to our understanding of what constitutes oral proficiency of graduate student instructors teaching at an American university. Results indicate that, when rating candidates, evaluators attend primarily to pronunciation and listening comprehension. We suggest that these two features represent essential linguistic components that provide the foundation on which the 'larger' functional categories of interactional and transactional language are built. The quantitative measures indicate a trend for an emerging significance of transactional language use while the qualitative analyses point to the importance of interactional language use. With a refinement in the operational definitions of these functional categories, we maintain that analyzing learner performance in terms of transactional and interactional language may yield meaningful results in our understanding and assessment of the language required for instructional purposes.

Acknowledgements

Versions of this paper were presented by the first two authors as Works in Progress at the 27th Annual Language Testing Research Colloquium, Ottawa and at the 28th Annual Language Testing Research Colloquium, Melbourne. We would like to thank LTRC participants for insightful comments on the study. We would also like to extend our gratitude to Liane Patsula for assistance with the statistical analysis, Eric Frey, Sarah Goodwin, Theresa Koehler, and Sunny Park for the hours spent transcribing, and Roann Altman, Christine B. Feak, and Spiros Papageorgiou for their meticulous reading of the manuscript and invaluable suggestions, which helped us to improve both clarity and content. Finally, John Swales continues to be a source of inspiration and wisdom. All shortcomings remain our own.

Notes

1. At many institutions, Graduate Student Instructors are referred to as Teaching Assistants.
2. The General Interview (Task 1) is the least structured task and thus contributes the greatest variation in the test. Discussion topics may range from descriptions of personal hobbies to theoretical explanations of current research projects. Because of this variability and because Task 1 serves mainly as a warm-up, this task was excluded from the current analysis.

3. For those candidates who show Very Strong Proficiency (4+), tasks may be shortened as soon as a speech sample that indicates Very Strong Proficiency has been obtained. This accounts for the large range in the time of tests.
4. Coding and scoring of the features was conducted in a linear fashion so that the third researcher was able to check the coding accuracy of the second researcher who in turn had checked the coding of the first researcher who had checked the scoring of the third researcher.
5. All assumptions (independent variables do not demonstrate multicollinearity; observations are independent; independent variable is linearly related to logit of dependent variable) were met.
6. A stepwise regression by college with approval as the dependent variable resulted in a warning message indicating that there is a cell with 0 observations for pronunciation.
7. To determine a candidate's strength or weakness for a particular feature, his/her 'score' for that feature (e.g. transactional language) was compared to the group mean.
8. Iwashita et al. created an overarching category of Phonology, in which fluency and pronunciation were subcategories. The current study merged these two features into a single pronunciation score.

References

Beebe, L. & Waring, H. Z. (2004). The linguistic encoding of pragmatic tone: Adverbials as words that work. In Boxer, D. & Cohen, A. (Eds.), *Studying speaking to inform second language learning* (pp. 228–249). Clevedon, UK: Multilingual Matters.

Blum-Kulka, S., House, J., & Kasper, G. (Eds.) (1989). *Cross-cultural pragmatics: Requests and apologies.* Norwood, NJ: Ablex.

Briggs, S. L. (Developer). (1987, 1999, 2003). *Graduate Student Instructor Oral English Test*. English Language Institute, University of Michigan. Ann Arbor, MI: English Language Institute.

Briggs, S. L. (1994). Using performance assessment methods to screen ITAs. In Madden, C. G. & C. L. Myers (Eds.), *Discourse and performance of international teaching assistants*. Alexandria: TESOL.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*, 1–25.

Brown, G. & Yule, G. (1989). *Discourse analysis.* Cambridge: Cambridge University Press.

Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Testing Research, 3*, 185–214.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tasks and rater groups. *Language Testing, 12*, 16–33.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*, 369–383.

Chapelle, C. 1998: Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Second language acquisition and language testing interface* (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254–272.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238.

Fulcher, G. (2003). *Testing second language speaking.* London: Pearson.

Fulcher, G., & Márquez Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing, 20*, 321–344.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context, and text: A social semiotic perspective.* Oxford: Oxford University Press.

Huang, Li-Shih. (2005). Fine-tuning the craft of teaching by discussion. *Business Communication Quarterly, 68*, 492–500.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 24–49.

Lantolf, J.P., & G. Appel. (Eds.). (1994). *Theoretical framework: An introduction to Vygotskian perspectives on second language research.* Westport, CT: Ablex.

Levelt, W. J. M. (1981). The speaker's linearisation problem. *Philosophical Transactions of the Royal Society of London*, *295*, 305–315.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience, and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*, 415–437.

Madden, C. G., & C. L. Myers (Eds.) (1994). *Discourse and performance of international teaching assistants*. Alexandria: TESOL.

Myers, C. L. (1994). Question-based discourse in science labs: Issues for ITAs. In Madden, C. G., & C. L. Myers (Eds.), *Discourse and performance of international teaching assistants*. Alexandria: TESOL.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal, 62*, 266–275.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*, 169–192.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly, 31*, 95–119.

Rounds, P. L. (1990). Student questions: What do we know about them? Paper presented at the Preconference Symposium on ITA Training, 24th Annual TESOL Convention, San Francisco, CA.

Skehan, P. (2001). Tasks and language performance assessment. In Bygate, M., P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks second language learning, teaching, and testing.* London: Pearson.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275–302.

Taguchi, N. (2007). Task difficulty in oral speech act production. *Applied Linguistics, 28*, 113–135.

van Dijk, T. A. (1977). *Text and context.* London: Longman.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3–12.

Upshur, J., & Turner, C. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*, 82–111.

Appendix 1: Rating Scale (Briggs, 1999)

5 Superior Acceptability for teaching duties	Expert user of English as a second language
4+	
4 Acceptable for a range of teaching duties in most departments in LSA or Engineering	Communicates well in spoken English Typical characteristics: Range and control of linguistic repertoire: uses field-specific vocabulary that promotes clear expression of concepts; uses some colloquial and idiomatic terms and expressions; uses terms and expressions to link concepts and highlight key points; may show some word choice variation but it does not inhibit communication of concepts; grammatical deviations, when present, are minor and are not particularly distracting. Speech production: fluent and understandable; may have phonological variation or some variation in rhythm or rate but is intelligible; speech is clear and projected adequately. Language use and instructional context awareness: is appropriately concise or elaborated depending on context; frame or previews concept to link to prior knowledge, conveys a coherent explanation of a concept, offers relevant examples or analogies, defines terms, summarizes or rephrases points, aware of student perspective, provides relevant suggestions and guidance. Interactive communication: gestures, eye contact, and bodily stance promote intended communication; blackboard use and other visuals promote communication of concept; anticipates what might not be understood, attentive to communication and monitors the communication; understands spoken English well.
4– Acceptability Restricted to certain College of Engineering teaching assignments – relatively structured with built-in support, certain lab settings, and/or courses for upper level or advanced students.	Some limitations in effectiveness in fluency and/or clarity especially in more extended discourse; or occasional speech intelligibility problems with single lexical items but shows effective compensatory skills. Linguistic limitations are compensated for by communication strategies or pedagogical skills.
3+	Evidence of compensatory strategies but use not sufficiently effective in overcoming linguistic barriers.
3	Limitations in language use and speech production. Pronunciation and/or lexical/grammatical inaccuracies impede effective communication and/or quite apparent limitations in understanding spoken English.
3–	
2+	Very weak performance in all four test tasks, seems to struggle with expression and comprehension, discourse skills limited throughout by apparent limitations in linguistic repertoire and range.
2	
1	

Appendix 2: Features Examined

Transactional language

Operational Definition: Transactional is that function which language serves in the expression of content. Transactional language 'is used to convey 'factual or propositional information …[it] is primarily 'message oriented.' It is important that the recipient gets the informative detail correct' (Brown & Yule, 1989, p. 2). Syntactic/lexical realization of transactional language use are logical and temporal markers which serve a cohesive function (Halliday & Hasan, 1976). Examples of forms coded in the current study:

- Articles
- Conjunctions and connectors
- Demonstrative pronouns and adjectives
- Partial and full repetition
- Reference forms (e.g. such + plural noun)
- Tense/aspect

Interactional language use

Operational Definition: Utterances that encourage a speaker to expand or elaborate on a given topic, or mitigate the apparent rigidity or hostility of an utterance; collaborative moves and signs of empathy that are used to build rapport. Form alone, however, does not define these features. Therefore, the content of the utterance has to be examined.

Examples of features coded in the current study:

- Collaborative moves during the Office Hour Role Play: 'Does this seem ok with you?'
- Forms that serve to mitigate apparent rigidity or hostility: Durative aspect, past tense, hedges, modals.
- Forms that build rapport: Comprehension checks (e.g. 'Is that clear?' 'Can I move on?'); expressions of sympathy/empathy (e.g. 'sorry to hear that,' 'oh that's too bad').

Appendix 3: Candidate Profiles

Candidate n=44	L1	College	Gender	Time US	(Not) Approved
1	Chn	LSA	F	1 month	Approved (4+)
2	Chn	LSA	F	1 month	Approved (4+)
3	Chn	LSA	F	2 years	Approved (4+)
4	Chn	CoE	F	3 years	Approved (4+)
5	Farsi	CoE	M	1 year	Approved (4+)
6	Chn	CoE	M	2.5 months	Approved (4+)
7	Chn	LSA	F	1.5 years	Approved (4)
8	Chn	LSA	F	4 months	Approved (4)
9	Kor	LSA	M	2 weeks	Approved (4)
10	Chn	LSA	M	3 weeks	Approved (4)
11	Kor	LSA	M	2.5 years	Approved (4)
12	Chn	LSA	M	2 years	Approved (4)
13	Chn	LSA	M	1 month	Approved (4)
14	Frn	CoE	M	1 year	Approved (4)
15	Grk	CoE	M	1 year	Approved (4)
16	Dtch	CoE	M	4 months	Approved (4)
17	Chn	CoE	M	7 years	Approved (4)
18	Chn	CoE	M	1.5 years	Approved (4)
19	Chn	CoE	F	1 year	Approved (4)
20	Chn	LSA	M	3 weeks	Not Approved (4-)
21	Chn	LSA	M	4 years	Not Approved (4-)
22	Kor	LSA	M	1 year	Not Approved (4-)
23	Kor	LSA	M	1 year	Not Approved (4-)
24	Chn	LSA	M	3 weeks	Not Approved (4-)
25	Kor	LSA	F	1 year	Not Approved (4-)
26	Kor	LSA	F	9 months	Not Approved (4-)
27	Kor	CoE	M	1.5 years	Approved (4-)
28	Chn	CoE	F	6 months	Approved (4-)
29	Chn	CoE	M	3 months	Approved (4-)
30	Kor	CoE	M	4 months	Approved (4-)
31	Kor	CoE	M	3 years	Approved (4-)
32	Thai	CoE	M	9 months	Approved (4-)
33	Viet	CoE	M	2 years	Approved (4-)
34	Kor	CoE	F	1 year	Not Approved (3+)
35	Kor	CoE	M	9 months	Not Approved (3+)
36	Kor	CoE	M	1.5 years	Not Approved (3+)
37	Chn	LSA	F	1 year	Not Approved (3+)
38	Chn	LSA	F	1 year	Not Approved (3+)
39	Chn	LSA	F	9 months	Not Approved (3+)
40	Chn	LSA	M	1 month	Not Approved (3)
41	Chn	LSA	M	1.5 years	Not Approved (3)
42	Thai	LSA	M	1 month	Not Approved (3)
43	Kor	CoE	M	1.8 years	Not Approved (3)
44	Kor	CoE	M	8 months	Not Approved (3)

Considering all candidates by rating, six received a rating of 4+, 13 received a rating of 4, 14 received a rating of 4-, six received a rating of 3+, and five received a rating of 3. The majority of ratings are thus grouped around the cut-scores. The length of time candidates had been in the USA ranged from 2 weeks to 7 years. The median time in the USA was one year.[1]

[1] No correlation has been found between length of time in the USA and final rating.