Providing Validity Evidence for a Speaking Test Using FACETS

Hyun-Joo Kim[1]
Teachers College, Columbia University

ABSTRACT

Speaking has been considered one of the most important skills in second language teaching and assessment. However, to date there has not been a clear definition of speaking ability. The present study attempts to examine what it means to be able to speak a second language using a newly designed computer-delivered speaking test. The test was created based on a theoretical definition of speaking ability, and then administered to 95 ESL students. A variety of statistical analyses were employed to examine the validity of the test, including reliability analyses, correlation analyses, and a many-facet Rasch measurement (Linacre, 1989) analysis. Results seem to provide some evidence for the validation of the test.

INTRODUCTION

Speaking has traditionally been viewed as one of four language skills along with listening, reading, and writing. This skill-based distinction has served to provide L2 researchers with a common conceptual framework for categorizing different aspects of communicative language use. However, when it comes to defining what speaking skills actually consist of, there is no widely accepted theoretical model. In fact, most measures of speaking ability are task-centered, which, according to Bachman (2002), is not concerned with speaking ability—the construct of interest—but rather performance on tasks. For example, one of the most frequently used oral proficiency measures, the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI), has been criticized for its lack of a theoretical foundation. Many researchers (e.g., Bachman, 1988, 1990; Bachman & Clark, 1987, Bachman & Savignon, 1986; Lantolf & Frawley, 1985, 1988; Salaberry, 2000; Van Lier, 1989) have pointed out that oral proficiency is not defined in the OPI. Instead, the instrument merely lists the real-life language use situations it intends to measure. Subsequently, the validity of ACTFL-like tests is established through a comparison between the features of performance elicited via the elicitation method (i.e., interview) and features of real-life situations, primarily based on discourse analytic tools. The results to date indicate that features of performance in an OPI are different from those in natural conversation, leading some researchers (e.g., Johnson, 2001) to conclude that OPIs measure a specific type of speaking event, that is, speaking in the context of an interview. As a result, inferences based on the scores are only predictions about future performance on the task

[1] Hyun-Joo Kim is a doctoral student in Applied Linguistics at Teachers College, Columbia University. Her area of research interest is language assessment, with a special emphasis on the testing of second language speaking for adult learners. Correspondence should be sent to Hyun-Joo Kim, 1230 Amsterdam Ave., #802, New York, NY 10027. E-mail: hk312@columbia.edu.

itself, which limits generalization and extrapolation of the scores to the target domain of interest (i.e., speaking ability). In other words, the lack of theory underlying a test design confounds the method (i.e., interview) with the construct of interest (Bachman, 2002), thus confusing the observed performance with performance as a vehicle of assessment of ability being measured. Therefore, it is critical to define what speaking ability is when designing and validating a speaking test.

Beyond the issue of construct definition, performance assessments including a speaking test bring unique challenges to the validation procedure as they involve human judgment in scoring and a rating scale. Rater behavior, for instance, has been considered one of the crucial sources of variation in a performance assessment, and a major issue of concern (Lynch & McNamara, 1998; McNamara, 1996; Shohamy, Gordon, & Kraemer, 1992; Tyndall & Kenyon, 1996; Upshur & Turner, 1999; Wigglesworth, 1993). Given the importance of rater variation, the degree to which raters are consistent within themselves and across different raters should be examined and included in validation efforts. Furthermore, to ensure the appropriateness of inferences about test-takers' ability, one must make certain that the criteria used to score responses have been applied as intended. In other words, the rating scale also needs to be investigated. For these purposes, many-facet Rasch measurement (MFRM) has proven useful. MFRM is an extension of the one parameter Rasch model, which calculates examinee ability along with other variables, such as raters and tasks.

In sum, multiple sources of evidence should be considered in the development and validation of a speaking test, and in the present study, validity evidence for a newly designed speaking test is collected. This speaking test was developed as part of a comprehensive placement test battery for an English language program for adult learners learning English as an additional language. The test intends to measure students' communicative language ability in a variety of oral communication situations.

In this paper, I will first briefly review how researchers have defined the construct of speaking ability. Then, evidence for test validity, including the relationship among the components of the construct, rater behavior, and functionality of rating scales, will be presented. The study specifically addresses the following research questions:

1. What is the nature of communicative language ability as measured by the current speaking test?
2. What is the nature of the task-dependent measure of task completion as measured by the current speaking test?
3. To what extent does the test separate examinees into distinct levels of speaking ability?
4. How appropriately are the rating scales functioning?
5. To what extent do the raters vary in terms of severity?

CONTEXT OF THE STUDY

The Community English Program (CEP) at Teachers College for which the speaking test was developed provides instruction for learners of English as a second language. It caters to an extremely diverse student population in terms of age, native language, socio-economic and immigration status, educational background, and purpose for learning English. Furthermore, the program is a learning environment for developing teachers who are students enrolled in the

TESOL or Applied Linguistics programs at Teachers College. Although the diverse population of students has enriched the program greatly, it has also been a challenge to place them into appropriate levels. Some of the students, who have spent most of their adult lives in the United States, tend to be much more proficient in speaking and listening than in writing or reading. Other students are highly proficient in reading and/or grammar, but limited in speaking or writing. In an attempt to make accurate placement decisions, a comprehensive placement test battery was created, consisting of grammar, listening, reading, writing, and speaking sections.

The speaking test was originally conducted in a paired interview format. However, some concerns associated with potential sources of construct-irrelevant variation were noted, including interviewer variability, the effects of the partner's proficiency level, gender, and nationality, and examinees' personalities. Because these issues could not be researched systematically due to practical constraints, it was necessary to devise a new test. For the format of the new test, a computer-delivered, semi-direct test integrating various types of media, such as audio and video clips, to simulate face-to-face interaction appeared most appropriate. It was considered ideal to maintain some interactive, conversational features that were present in the previous face-to-face format interviews, while minimizing examiner-related variations. Furthermore, considering the diverse nature of the student population in the CEP, and the purpose of the test (i.e., placement), it seemed imperative to include a variety of task types that could spread students out along their ability levels. Using computers as a delivery means was considered the most efficient and appropriate way to do so. Another considerable advantage of a computer-delivered test is its capability to record examinees' responses. This way, raters can listen to responses more than once, if needed, and judge them more reliably. Furthermore, responses can be rated by more than one rater, reducing the effects of rater-related variations.

In the next section, a review of the literature will be presented in order to define the construct under investigation.

REVIEW OF THE LITERATURE

Defining Speaking Ability: The Interactionalist Perspective

The widely accepted deconstruction of language ability into four separate skills has been questioned because the distinction based on channel and mode is rather simplistic. Bachman and Palmer (1996) have criticized the skills view of language ability, claiming that it fails to recognize differences within the supposedly same skill (e.g., casual conversation and formal presentation), and similarities among different skills (e.g., conversation and e-mail exchange). In other words, similarities or differences in the contextual features that activate speakers' (meta-) cognitive processes and language ability are not taken into account in the skills view of language ability. Bachman and Palmer further argue that language skills are not part of language ability, but rather of "the contextualized realization of the ability to use language in the performance of specific language use tasks" (pp. 75-76). This clearly suggests that speaking ability (or any other skill) should be defined as an interaction between language ability and the context in which language is used. Similarly, Chapelle, Grabe, and Berns (1997), and Chapelle (1999) argue that language ability must be described in relation to the characteristics of the situation in which communication takes place.

Defining speaking ability in terms of the interaction between language ability and the specific contextual variables elicited by the task, in essence, reflects an *interactional* approach to construct definition (Chapelle, 1999; Messick, 1989). Interactionalists advocate that contextual variables interact with the internal traits of the test-taker, and the scores obtained from the test should be interpreted as indicators of ability in a given context. Within this perspective, relevant aspects of both trait (i.e., communicative language ability) and context (i.e., a variety of oral communication situations) need to be identified. Following this approach, speaking ability in this study is defined as communicative language ability realized in different contexts.

Communicative Language Ability

The definition of communicative language ability has been debated over many years. In the early 1960s, language proficiency was defined very narrowly as consisting of linguistic components, such as phonology, structure, and the lexicon (Lado, 1961). It has since been recognized that while elements such as these are necessary to speaking, they cannot be detached from the context in which language is used.

In addition to linguistic accuracy, Hymes (1972, 1974), in his notion of communicative competence, proposed two other components: sociolinguistic appropriacy and psycholinguistic feasibility. Heavily influenced by Hymes, Canale and Swain (1980), and later Canale (1983) put forward a framework of communicative competence consisting of grammatical competence, sociolinguistic competence, discourse competence, and strategic competence. The addition of sociolinguistic and discourse competencies in their model indicated the need to account for a speaker's ability to communicate more than a single de-contextualized sentence.

Building on Canale and Swain's (1980) notion of communicative competence, Bachman (1990), and Bachman and Palmer (1996) proposed the most comprehensive model of language ability to date, called communicative language ability (CLA). The model of CLA consists of organizational knowledge (i.e., how individuals control language structure to produce grammatically correct utterances or sentences and texts), and pragmatic knowledge (i.e., how individuals communicate meaning and how they produce contextually-appropriate utterances, sentences, or texts). The former includes grammatical and textual knowledge, and the latter consists of sociolinguistic and functional knowledge.

While Bachman and Palmer (1996) substantially expanded the notion of communicative language ability, their definition of grammatical knowledge is still limited to grammatical form. In other words, situations where "students might know the form, but be unclear about the meaning" cannot be easily explained in the model (Purpura, 2004, p. 70). Purpura asserts that given the central role of meaning in communicative language use, a more explicit depiction of this aspect of grammatical knowledge would be helpful. Thus, he proposed a model of grammatical ability consisting of form and meaning. Grammatical form in Purpura's model of grammatical ability refers to the knowledge of linguistic forms at phonological, lexical, morphosyntactic, cohesive, information management, and interactional levels. Grammatical meaning embodies the literal and intended meanings of an utterance derived both from the meaning of the words arranged in syntax (literal meaning), and the way in which the words are used to convey the speaker's intention (intended meaning).

In addition to the rules operating at the sentential level, the ability to communicate includes rules that govern language use at the discourse level. Canale (1983) included this aspect in his revised framework of communicative competence and defined it as "mastery of how to

combine and interpret meanings and forms to achieve unified text in different modes (e.g., conversation, argumentative essay, or recipe)" (p. 339). Bachman (1990), and Bachman and Palmer (1996) also considered knowledge of cohesion (i.e., pronouns and lexical repetition), rhetorical organization (i.e., logical connectors), and conversational organization (e.g., turn-taking strategies and topic nomination) as an integral part of communicative language ability.

Besides grammatical and discourse competence, sociolinguistic competence is another crucial component in L2 communication. Sociolinguistic competence refers to a set of internalized rules concerning how to use language in socioculturally appropriate ways. In other words, to be a competent language user of a target language, a speaker needs to have the ability to choose appropriate language for a given situation and perform language functions or speech acts as intended in that context (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980; Celce-Murcia & Olshtain, 2000; Crystal, 1997; Larsen-Freeman, 1991; Leech, 1983; Purpura, 2004; Ranney, 1992).

What, then, does the ability to use language appropriately in an oral communication situation entail? According to Thomas (1983), it consists of two components: pragmalinguistic and sociopragmatic competence. The former refers to language users' knowledge of linguistic items used to realize speech acts (e.g., performative verbs) and their intended pragmatic force. The second component, sociopragmatic competence, encompasses knowledge of sociocultural factors such as the size of imposition, cost/benefit, social distance, and relative rights and obligations. Based on this knowledge, speakers assess contextual features embedded in a given interactional situation, select appropriate levels of politeness and formality, and encode such factors linguistically (Brown & Levinson, 1987; Faerch & Kasper, 1984; Hudson, Detmer, & Brown, 1995). Misinterpretation of the social context due to culturally different perceptions of what constitutes appropriate linguistic behavior may yield an inappropriate utterance.

This aspect of knowledge is similar to what Bachman and Palmer (1996) identify as sociolinguistic competence. In their model, sociolinguistic competence is defined as the appropriate use of registers, dialects or varieties, cultural references, and figures of speech. Additionally, the ability to use natural or idiomatic expressions is considered part of sociolinguistic competence. In summary, sociolinguistic competence encompasses the ability to use appropriate and natural language in a particular context containing various sociocultural features. Sociolinguistic competence is an important part of learners' communicative language ability, and should be addressed in an assessment.

As shown in the above review, communicative language ability is multi-dimensional, comprised of grammatical form and meaning, discourse competence, and sociolinguistic competence. When defining language ability, several researchers (e.g., Bachman, 2002; Bachman & Palmer, 1996; Chapelle, 1999; Chapelle et al., 1997; Douglas, 2000) have also argued that the construct should be described in relation to the characteristics of the situation (i.e., context) in which language use takes place. While there is a general consensus that context does play a significant role in defining a construct and interpreting test scores, the problem lies in a common understanding of the nature of context and what constitutes it. Thus, in the next section, an attempt will be made to provide a definition of context.

Context

Communicative language ability consists of interrelated knowledge components, activated by various features of the context in which communication happens. Contextual factors

are claimed to play a much more significant role in spoken interaction than in written communication since most oral exchanges are spontaneous (Celce-Murcia & Olshtain, 2000). According to Hymes (1972, 1974), context can be described in terms of the situation (setting/scene), participants, ends/purposes, act sequence, key/tone, instrumentalities (i.e., channel), genre, and norms of interaction.

The first feature, *situation* refers to physical and abstract psychological settings, or the speech event. The physical setting is the specific place where the interaction occurs, such as an employer's office or a classroom. Occasionally, these physical places might be less relevant than psychological settings established among the participants. For example, the place in which a caller makes a phone call may be irrelevant when he or she has called to complain about a defective product. Instead, the caller's and the respondent's roles (i.e., customer and customer service representative, respectively) and their goals (i.e., to get a refund and to please the unsatisfied customer) will likely determine how they will interact.

Participants refers to the sociocultural features implicitly or explicitly attached to the interlocutors in a communication situation. These features include the roles of each participant, the social distance (familiar vs. unfamiliar), the power relationships (i.e., professor vs. student, employer vs. employee, and so forth), age, and social identity. Information about the participants and their relationships is important in deciding appropriate norms of interaction in a communicative situation. Because the effects of participant roles and characteristics can vary across cultures, learners of a second language must be able to assess them appropriately and apply proper social norms (Celce-Murcia & Olshtain, 2000).

The third component in Hymes' framework, *ends/purposes*, refers to the goal or conventionally recognized function of a communicative event. In the example mentioned above, the customer's goal in calling the company about a defective product is to complain and ask for some type of compensation. According to the goal in this situation, the speaker would know which language forms to use in order to achieve the desired effects. Therefore, knowing the intended goal of a communicative event is critical in defining its context.

The *key* or *tone* of the context is also likely to influence language use. Tone can be scholarly or casual, serious or light, or formal or informal. For example, presenting research at a conference automatically sets the tone to be scholarly and formal. Differences in the tone of an interaction usually mean differences in registers or styles. Therefore, speakers need to be aware of the tone of the context and use language that is appropriate to that situation.

Instrumentalities refers to the channel of communication. Communication might take place in spoken or written form, or a combination of these forms. In addition to the linguistic channel, communication may also involve a nonverbal channel, such as pictures and graphs.

Genre refers to clearly defined types of communicative speech acts that are culturally and linguistically distinct. Each genre of communication contains unique conversational, textual, and structural characteristics, and these characteristics distinguish one genre from another. For instance, academics communicate knowledge through journal articles, book reviews, and the like that conform to particular sets of discourse features that have evolved over time. To become an accepted member of this community, one is expected to conform to the norms of academic writing (Swales, 1990).

The last component, *norms of interaction,* is heavily dependent upon other contextual features discussed above. In other words, appropriate norms of interaction in a given context are sensitive to the setting, participants, key, instruments, and genre of the particular interaction. In order to avoid violating the proper norms of interaction in a communicative event, second

language learners should assess the situation and integrate these variables when communicating. For example, speakers need to be able to use different forms of address, topics, and levels of politeness when the interlocutor is a professor than when he or she is a classmate.

In addition to encompassing the appropriate level of politeness and formality shared by members in a speech community, norms of interaction also govern turn-taking rules. Within a spontaneous, fast-paced conversation, speakers are expected to follow the norms of interaction by adhering to appropriate turn-taking rules (Celce-Murcia & Olshtain, 2000). These norms make it possible for the speaker and listener to constantly change speaking roles and construct shared meaning by maintaining the flow of talk with relatively little overlap and pausing.

In summary, context can be depicted in terms of its setting, participants, ends, act sequence, key, instrumentation, genre, and norms of interaction.

Quite differently from the interactionalist perspective, the behaviorists interpret test scores as indicators of performance specific to the task, not necessarily indicators of underlying traits. These researchers also consider context as a crucial variable in influencing test performance, but the type of inferences made are different. The behaviorist perspective is widely adopted in speaking assessment, particularly for tests designed based on the principles of task-based language assessment (see Brown, Hudson, Norris, & Bonk, 2002; Norris, Brown, Hudson, & Bonk, 2002; Norris, Brown, Hudson, & Yoshioka, 1998). Thus, in the next section, the behaviorist framework will be discussed in more detail.

Defining Task-Specific Performance: The Behaviorist Perspective

The fundamental difference between interactionalist and behaviorist perspectives is that, according to the behaviorist view, the context and the construct being measured cannot be separated and each construct can only be defined within a particular context. Advocates of this perspective (e.g., Brown et al., 2002; Norris, et al., 1998; Young & He, 1998) argue that "the learner's underlying knowledge is considered too elusive to define and so construct definition becomes a matter of defining the context in which language is used" (Read & Chapelle, 2001, p. 8). Task-based language performance assessment (Brown et al., 2002; Norris et al., 1998; Norris et al., 2002) is a typical example rooted in this approach.

Proponents of task-based language assessment are mainly concerned with performance observed during a particular task or within a particular context. In other words, the construct of interest in a task-based assessment is performance on the task itself (Brown et al., 2002). Therefore, consistencies observed in responses are explained as *samples of response classes* (Messick, 1989) rather than underlying traits, and hence, test scores are highly task-dependent and tied to the particular context in which the performance was elicited and observed. Consequently, several second language testers have argued that task-based performances should not serve as the focus of language performance assessment because this approach does not provide any basis for making interpretations beyond the particular task and test context. In other words, generalizability and extrapolation of test scores become a serious issue.

Despite these criticisms, the types of inferences that can be made based on task-based assessment may produce important information about test-takers. For instance, inferences regarding the degree to which a learner can utilize language to accomplish specific communication tasks can be useful to complement separately assessed dimensions of language ability. Advocates of task-based assessment further assert that the task-based language approach achieves vital assessment aims, such as fostering students' abilities to achieve communicative

goals, and do things with the knowledge they have acquired beyond the simple display of that knowledge on tests (Norris et al., 2002). Thus, it might prove valuable to assess the degree to which a task is completed, along with the underlying traits of communicative language ability. In the next section, theoretical definitions of the construct to be measured in the current study are provided.

OPERATIONAL DEFINITION OF SPEAKING ABILITY

Based on the above review, speaking ability is defined as communicative language ability realized in a variety of oral communication situations. Communicative language ability consists of grammatical form, grammatical meaning, discourse competence, and sociolinguistic competence.

Primarily based on Purpura's (2004) model of grammatical ability, grammatical form is defined as the ability to employ linguistic forms (i.e., grammar and vocabulary) at both sentential and suprasentential levels. Reflecting this definition, grammatical form in this test was operationalized as grammatical competence encompassing the accuracy, complexity, and range of linguistic resources (i.e., grammar and vocabulary). It pertains to how accurately the speaker uses his/her language in each situation. Additionally, the use of complex and various structures was included in the definition of grammatical competence.

Grammatical meaning in Purpura's model refers to the ability to produce and understand literal and intended meaning associated with an utterance. This entails phonological meaning, lexical meaning, and the morphosyntactic meaning of individual words and syntactic structures (literal meaning) as well as the speaker's intention encoded in the utterance (intended meaning). For this test, grammatical meaning was operationalized as meaningfulness, indicating how completely and clearly the speaker conveys what he or she means. It is mainly concerned with how meaningful the utterances are to the interlocutor. As long as the speaker's message and intention are understood, it can be said the response was meaningful to some degree. Another aspect to consider in the dimension of meaningfulness is the degree of completeness and sophistication of the message encoded in grammar—how much information is provided and how complex the information is.

Sociolinguistic competence includes the ability to use appropriate and natural language in a given context. It entails knowledge of language variation that consists of register variation and knowledge of naturalness (e.g., archaic word vs. contemporary expression). These four components of communicative language ability are believed to play an integral role when learners try to use the language effectively and appropriately in various social and cultural contexts. The operational definition of sociolinguistic competence is appropriate and natural language use—appropriate in terms of contextual variables in a given setting. The degree of politeness and formality in a communicative situation is expected to vary according to the contextual variables. Sociolinguistic competence is also concerned with naturalness of linguistic resources (e.g., grammatical structures, expressions, and vocabulary)—how natural and idiomatic the response is. In short, the operational definition of sociolinguistic competence encompasses both appropriate and natural use of language.

Discourse competence refers to the ability to organize information in a coherent way, using conventions to join utterances together (e.g., transitional words and phrases, repetition of key words). In an oral communicative situation, discourse competence includes the application

of conversational rules, such as adjacency pairs and turn-taking conventions. The operational definition of discourse competence is clear, coherent organization of responses. Organization pertains to the overall structure and order of content in a text. For instance, in the situation of leaving a telephone message, the speaker needs to identify herself, explain the reason for calling, and close the discourse by asking for future action (i.e., "call me back at…"). Cohesion across utterances may also be achieved by the use of pronouns, ellipsis, lexical repetition, and cohesive devices like conjunctions. In short, discourse competence was operationalized as the degree to which the speaker organizes a response in a clear, coherent way.

In addition to the four dimensions of communicative language ability, a measure of task completion was included to assess overall accomplishment on a given task. The operational definition of task completion refers to how successfully the speaker addresses the task given. In order to complete a task successfully, examinees would need to understand what the task entails and respond to it as instructed. In this sense, task completion is strongly tied to characteristics of context and its characteristics.

METHOD

Participants

The participants in this study were 95 students registered in the Community English Program in the spring of 2004. The background information is summarized in Table 1.

TABLE 1
Background Information

Languages	Percent	Status	Percent	Length of Residence	Percent
Spanish	40.2%	Immigrant	41.8%	>6 yr	12%
Korean	17.4%	Spouse (F-2/J-2)	23.1%	3 - 6 yr	7.6%
Japanese	13.0%	Student	19.8%	1 - 3 yr	21.7%
Chinese	7.6%	Visiting scholar	6.6%	6 mo-1 yr	20.7%
Polish	7.6%	Other	8.7%	<6 mo	38%
Russian	4.3%				
Italian	3.3%				
Other	6.6%				
Total	100%	Total	100%	Total	100%

TABLE 1
Background Information (Continued)

Reason for studying English	Percent	Educational background	Percent
Academic	36.7%	Doctorate	4.4%
Job	28.9%	Master's	21.2%
Both	24.4%	Bachelor's	52.2%
Other	10%	High school	17.8%
		Middle school	4.4%
Total	100%	Total	100%

As it can be seen in the table, the vast majority of the students reported Spanish as their first language (40%), followed by Korean (17%). About 42 % of the students were immigrants and formed the largest group. The second largest group consisted of spouses of international students or visiting scholars. In terms of length of residence, most students in the program have lived in the United States for less than three years. With regard to the reasons for attending the CEP, some students reported that they wanted to improve English for academic purposes and others for job-related reasons. About 24 % of the participants noted both academic and professional development as reasons for studying English. The average level of education for the students in the program was quite high. The majority of students had a college-level degree or above.

Instrument

The Test

The test consisted of ten tasks, which intended to assess grammatical competence, meaningfulness, discourse competence, and sociolinguistic competence as well as task completion in various oral communication situations. The test was semi-direct and delivered via computer. Every task included 20 to 30 seconds of planning time. The response time varied from 30 to 60 seconds.

The description of the tasks is summarized in Table 2. Click on Task 7, Office, for an example of the task from the test.

TABLE 2
A Description of the Tasks

TLU Domain	*Selected Contextual Features*	*Task Description*	*Task Description*
Service encounters	Setting	Home	Office
	Participants	❶Customer—Catering Service	❹Customer—Realtor
	Ends	Complaining about services	Complaining about a missed appointment
	Channel of input	Listening	Listening
One-on-one meeting	Setting	Professor's office	Office
	Participants	❺Professor—Student	❼Employer—Employee
	Ends	An advising session: To discuss the student's performance	Performance review session: To discuss the employee's sales report
	Channel of input	Test scores	Line graph representing the employee's sales report
Leaving a voice message	Setting	N/A	N/A
	Participants	❻Customer—Car dealer	❽Teacher—Student
	Ends	Refusing a suggestion by a sales person	Refusing a request for deadline extension
	Channel of input	Listening: voice message (Version1)	Listening: voice message (Version2)
		Reading: e-mail (Version2)	Reading: e-mail (Version1)
Narrating	Setting	N/A	N/A
	Participants	❷Between friends	❸Employer—Employee
	Ends	To tell a story about a movie	To tell a story based on pictures
	Channel of input	None	Visual: a set of pictures
Summariz-ing	Setting	N/A	School—an art history class
	Participants	❾Between friends	❿ Between classmates
	Ends	To summarize information for a friend	To summarize a lecture for a friend who missed the lecture
	Channel of input	Listening: radio commentary (Version1)	Listening: lecture (Version2)
		Reading: magazine article (Version2)	Reading: a part of a book (Version1)

Note. The underlined participant is the role that the test-taker is supposed to play.

The Rubric

Analytic rating scales ranging from 0 to 5 were used to score the responses. Rubrics for each language dimension and task completion are presented in Appendix A.

Test Administration

The test was administered in a computer lab. As examinees entered the lab, they were randomly assigned to a computer. Each student had their own computer and a headset to listen to the tasks and record their responses. One half of the computer stations in the lab were networked to Console 1, and the other half to Console 2. The test administrators at each console sent out tasks to individual stations and collected responses; the examinees did not need to operate the machines.

Before testing began, the students were asked to fill out the background questionnaire. Then, they were given a practice task to familiarize themselves with the testing format. Just like with an actual item, they listened to the question and recorded their responses. The students were given a chance to ask questions after that. The whole test administration lasted an hour.

Scoring Procedures

The teachers in the CEP also served as raters in this study. CEP teaching is part of the requirement for the TESOL program at Teachers College. The total number of raters was 19, and all of them were required to attend a norming session conducted on a separate day. During the training, raters were also given time to score sample responses, followed by discussions to clear up any ambiguities and questions.

After the test-takers finished their test, raters came into the lab and retrieved recorded responses in MP3 format. They scored responses based on grammatical competence, meaningfulness, discourse competence, sociolinguistic competence, and task completion. Each rater was given a rating sheet containing information regarding which dimensions and examinees to score. Instead of scoring for all five dimensions, raters were assigned to score for two dimensions. In that way, raters were specialized for the specific dimensions assigned to them.

Missing Information

While scoring, raters were asked to indicate inaudible responses. Two possible reasons can account for these non-responses. One is technical problems, and the other - the examinee's limited proficiency. Five examinees were identified with such missing information, and were subsequently excluded from the analyses.

Analyses

Statistical analyses were performed using SPSS for Windows, version 11.0. First, descriptive statistics for each task were calculated in order to examine the central tendencies, variability, and distribution of the scores.

Then, a series of reliability analyses were performed. First, the reliability of the whole test with ten items was calculated using Cronbach's alpha to examine how the test is functioning

as a whole. Then, the reliability of each dimension of communicative language ability was estimated. In addition, inter-rater reliability was estimated, using Pearson product-moment correlations.

To examine rater variation and rating scale functionality, many-facet Rasch measurement (MFRM) was employed. MFRM is a special model of 1-parameter item response theory. It calculates item difficulty and person ability simultaneously, and produces estimates for each on an interval scale, known as logits (Wright & Masters, 1982). Separability of test-takers, that is, how spread the test-takers are in terms of their ability levels, can be specified in the results. Considering the purpose of the test, which is to make placement decisions, separability of students is crucial.

The many-facet Rasch analysis also includes other facets that contribute to test score variation. For instance, raters can be specified as an additional facet, and their differences in severity can be taken into account, and the differences can be compensated for across facets specified (Linacre, 1989; Lynch & McNamara, 1998). Given the fact that raters vary in terms of their severity, the many-facet Rasch analysis is useful in analyzing performance data like ratings on a speaking test. Many-facet Rasch measurement also allows us to identify particular elements within a facet that are problematic or *misfitting*. This may involve a rater who is inconsistent in his or her ratings. MFRM also provides information regarding how well the rating scales are applied.

For the current study, the MFRM analysis was performed using FACETS 3.22 for the IBM (Linacre, 1999). The model used for the analysis was the Partial Credit model (Wright & Masters, 1982), in which rating scales applied by raters were treated as items and the structure of the rating scale for one rater was assumed to be different. The assumption that each rater had his or her own scale structure and the scoring criteria for each item were different would allow for a detailed investigation of rater behavior and rating scales. The Partial Credit model used in this study was:

$$\text{Log}\,(P_{nijk}/P_{nijk-1}) = B_n - C_j - D_i - F_{ik}$$

P_{nijk} = the probability of examinee *n* being awarded a rating of *k* when rated by rater *j* on item *i*
P_{nijk-1} = the probability of examinee *n* being awarded a rating of *k-1* when rated by rater *j* on item *i*
B_n = the ability of examinee *n*
C_j = the severity of rater *j*
D_i = the difficulty of item *i*
F_{ik} = the difficulty of achieving a score within a particular score category *k* on a particular item *i*

FINDINGS

Descriptive Statistics

The means of the ten tasks ranged from 2.04 to 2.81 out of a possible 5 points. All of the ten tasks except Task 10 produced means around 2.5. The lowest mean was 2.04 for Task 10, and the highest was 2.81 for Task 4, producing a range of .74. Standard deviations of each task ranged from 1.13 to 1.38. Task 10 produced the highest standard deviation, showing that a wider range of abilities was invoked by this task.

In an attempt to determine if the score distributions were approximately normal, I also examined the skewness and kurtosis of the ten tasks. All values for skewness and kurtosis were within the acceptable range (i.e., ± 3.0), indicating that the items appeared to be univariately normal. The descriptive statistics of the tasks are summarized in Table 3.

TABLE 3
Descriptive Statistics for the Ten Tasks (N=90)

	Task	Mean	SD	Skewness	Kurtosis
1	Catering services	2.64	1.16	-.77	.17
2	Movie	2.49	1.21	-.26	-.39
3	Fly in soup	2.52	1.13	-.33	-.27
4	Realtors	2.81	1.15	-.83	.34
5	Advising session	2.57	1.27	-.53	-.27
6	Car dealership	2.61	1.22	-.72	-.01
7	Employee review	2.43	1.15	-.39	-.36
8	Deadline extension	2.77	1.27	-.80	.11
9	Electric cars	2.74	1.33	-.65	-.39
10	Barbizon School	2.04	1.38	-.21	-1.14

As shown in Table 3, the similar range of means across the tasks, except Task 10, indicates that the examinees in general performed similarly across the tasks. In other words, the various types of tasks included in the test might not have had much of an impact on scores.

Task 10, which was found to have the lowest mean among the ten tasks, required students to summarize a text to an imaginary interlocutor, classmate. While Task 9, which also asked examinees to summarize a text, produced a relatively high mean, Task 10 yielded a low mean. In an attempt to account for the discrepancy in means between Task 9 and 10, the characteristics of each task were examined. The setting of Task 9 was an everyday situation, while Task 10 was an academic one. The interlocutor of Task 9 was a friend who wanted to buy an electric car. Similarly, the equivalent listener in Task 10 was a classmate who missed a class. The purpose set up for Task 9 was to summarize information about electric cars to persuade the friend not to buy one. For Task 10, the examinee was supposed to summarize a lecture in an art history class to the classmate. Task 9 was 269 words long and Task 10 was 320 words.

Examining the characteristics of the two tasks, one apparent difference is the length of the text. The shorter text (Task 9) was found to produce a higher mean than the longer one (Task 10), as one might expect. Another difference between the tasks is their settings. Task 9 was in an everyday situation, while Task 10 was set in an academic situation, a lecture about a group of French artists. A daily situation might be more accessible to test-takers than an academic setting, and hence easier. In addition to the setting, the topic also seemed to differ in terms of familiarity. In the test survey, although most examinees indicated they did not have prior knowledge about electric cars or the Barbizon school artists, the topic of cars could have been more relatable to the examinees than that of art and painting. Furthermore, familiar words and concrete examples mentioned in Task 9 text might have helped the examinees to activate their schemata and fill any gaps in comprehension. Preliminary analyses of the responses indeed showed that most examinees including those with limited language appeared to understand the examples and tried to cite them in their responses.

In contrast, the topic of Task 10, art, might have been more abstract and conceptual. In other words, the text used for Task 10 was context-reduced, requiring the test-takers to rely more heavily on their knowledge of the language code and genre types. Complete comprehension of the text, thus, seemed necessary in order to carry out the task successfully, which may have increased the level of difficulty. The difficulty in processing context-reduced texts such as this one and its effects on test performance needs to be explored further.

In addition to the means of each task, I calculated descriptive statistics for each dimension of communicative language ability as well as task completion for the ten tasks to examine their relative difficulties. Means ranged from 1.93 for task completion on Task 10 to 3.14 for task completion on Task 4. The smallest standard deviation was 1.13 for grammatical competence on Task 1 and the largest was 1.53 for meaningfulness on Task 10. All skewness and kurtosis were within the acceptable range. In Table 4, the means are reported.

TABLE 4
Means for Each Dimension of Language Ability (N=90)

	Task	Meaning	Grammar	Sociolinguistics	Discourse	Task Completion	Average
1	Catering services	2.71	2.54	2.49	2.71	2.74	2.64
2	Movie	2.69	2.32	N/A	2.33	2.62	2.49
3	Fly in soup	2.71	2.56	N/A	2.37	2.46	2.52
4	Realtors	2.78	2.72	2.47	2.94	3.14	2.81
5	Advising session	2.72	2.56	2.31	2.59	2.66	2.57
6	Car dealership	2.59	2.58	2.51	2.67	2.71	2.61
7	Employee review	2.72	2.58	1.96	2.54	2.59	2.43
8	Deadline extension	2.79	2.72	2.73	2.79	2.83	2.77
9	Electric cars	2.77	2.73	N/A	2.61	2.86	2.74
10	Barbizon School	2.15	2.14	N/A	1.94	1.93	2.04

As seen in the table, the range of means for the meaningfulness dimension was from 2.15 to 2.79. The means for the dimension of grammatical competence, ranging from 2.14 to 2.73, are slightly lower than those produced for meaningfulness. The differences are very small and yet consistent throughout the ten tasks, suggesting that the examinees managed to convey meaning before mastering form.

Another notable point in the grammatical competence dimension is the range of means. The small range of .59 indicates grammatical performance did not vary much from one context to another. This may be considered counterevidence to the studies reporting variations in grammatical features due to contextual features. For example, Tarone (1985) examined the use of morphemes (i.e., the third person singular *–s*) in three different situations (i.e., a grammar test, an oral interview, and an oral narrative) and found different accuracy rates. In another study, Tarone and Liu (1995) reported the subject in their study attempted more complex syntactic structures with peers and the researcher, while resorting to very simple English in interaction with the teacher. However, in the current study employing a quantitative approach, differences in task features including the power status and/or social distance of the supposed interlocutors (employee, professor, and friend) did not seem to influence test scores much, as reflected in similar means. One reason for the inconsistent finding may be related to the definition of variable

(i.e., grammatical competence). The scale of grammatical competence used in this study is communicative in nature and mostly concerned with meaning-interfering errors, such as tenses, while Tarone's morpheme study focused on the use of the third person singular *-s* which does not interfere with communication. Minor errors like the third person singular *-s* might not have been penalized by the raters who scored the responses for the present study. For more conclusive results, an additional analysis regarding how the raters interpreted and applied the rating scale needs to be undertaken.

In terms of sociolinguistic competence, all but Task 7 produced means that were similar in range. Task 7 produced the lowest mean (1.96). In that task, examinees were supposed to play an employer concerned with past sales records of an employee. Several raters noticed that one recurring pattern of responses was the examinee sounding too harsh and mean, which they considered inappropriate. Based on the responses alone, however, it is hard to pin down what caused examinees to come off as mean employers. It could have been due to their misunderstanding of the norms of interaction, or they might have wanted to play a difficult employer. Information regarding how the test-takers interpreted the situation and what they intended to express will provide a better understanding. With the exception of Task 7, other tasks appeared to have means that are more or less the same, which indicates that the different contextual features across the tasks did not affect scores on sociolinguistic competence.

With regard to discourse competence, the lowest mean was found for Task 10 (1.94), and the highest for Task 4 (2.94), producing a range of one point. The low mean of 1.94 falling below the "Fair" category (Point 2) on the rubric indicates that test-takers had a hard time organizing their responses in a coherent manner. The lack of understanding of the input text provided for Task 10 might be responsible for the incoherent text that resulted in the low mean.

For task completion, the lowest mean was 1.93 for Task 10, and the highest was 3.14 for Task 4, producing a range wider than 1 point. As with discourse competence, Task 4 produced the highest mean, while Task 10 yielded the lowest. Here again, the difficulty associated with the input text might account for the low degree of task completion for Task 10.

The one-point difference found for the dimensions of discourse competence and task completion seems large, considering the scale of 0 to 5 used in this study. To some extent, the wide range might indicate that examinees' performances on these two measures varied highly across different tasks. In other words, characteristics of the tasks might have had a greater impact on the way examinees organized their responses and the degree to which they completed the tasks. In contrast, for the dimensions of meaningfulness, and grammatical and sociolinguistic competence, the discrepancies were less than 1 point, demonstrating the levels of performance on these dimensions tended to be stable across different tasks.

Reliability Analyses

As shown in Table 5, the internal consistency estimate using Cronbach's alpha for the tasks was very high (α=.98), which provides evidence that the examinees performed consistently across the tasks within the test. Consistent performance can be explained in terms of a single underlying factor (i.e., speaking ability).

TABLE 5
Reliability Analyses of Ten Tasks (N=90)

Task	Corrected item-total correlation	Alpha if item deleted
1	.794	.975
2	.910	.972
3	.905	.972
4	.867	.973
5	.913	.971
6	.881	.972
7	.892	.912
8	.914	.971
9	.906	.971
10	.867	.973
N of Items = 10		Alpha =.975

In Table 5, Task 1 had the lowest item-total correlation, which may be due to the fact that it was the first item administered. Although the examinees were given a practice question before the actual test began, they might not have been completely familiarized with the testing situation.

I then performed a series of reliability analyses for each dimension of language competence across the ten items. The results are present in Table 6. According to the model, the construct of language competence consists of four observable variables: meaningfulness, grammatical competence, discourse competence, and sociolinguistic competence. A reliability coefficient for the task-dependent measure, task completion, was also calculated. The large values of internal consistency estimates indicate that there is a high degree of homogeneity in ratings given for each dimension.

TABLE 6
Reliability Analyses of Each Dimension of Language Competence (N=90)

Meaning		Grammar		Sociolinguistics		Discourse		Task Completion	
Item	Item-total correlation	Item	Item-total correlation	Item	Item-total correlation	Item	Item-total correlation	Item	Item-total correlation
1M	.737	1G	.752	1S	.775	1D	.715	1T	.717
4M	.825	4G	.834	4S	.780	4D	.819	4T	.753
6M	.833	6G	.833	6S	.804	6D	.811	6T	.747
8M	.887	8G	.842	8S	.859	8D	.863	8T	.849
5M	.872	5G	.906	5S	.756	5D	.867	5T	.821
7M	.865	7G	.843	7S	.743	7D	.804	7T	.845
2M	.867	2G	.859		N/A	2D	.862	2T	.815
3M	.900	3G	.891		N/A	3D	.820	3T	.771
9M	.895	9G	.877		N/A	9D	.853	9T	.846
10M	.853	10G	.858		N/A	10D	.829	10T	.794
Alpha =	.968	Alpha =	.967	Alpha =	.926	Alpha =	.960	Alpha =	.952

In an attempt to examine the degree to which the 19 raters were consistent in rating, correlation coefficients were calculated between the first and second ratings of each measure, using the Pearson product-moment correlation. The coefficients ranged from .92 to .95, indicating a high level of consistency between the raters. All estimates were found statistically significant at the .01 level. This appears to provide some evidence for consistency among the raters in scoring the responses. Table 7 presents a summary of inter-rater reliability estimates.

TABLE 7
Inter-rater Reliability

	Inter-rater reliability
Meaningfulness	.92
Grammatical Competence	.95
Discourse Competence	.93
Sociolinguistic Competence	.92
Task Completion	.93

Correlation Analyses

To examine the relationships among the dimensions in detail, I performed correlation analyses. The results are presented in Table 8.

TABLE 8
Correlation Analyses of Task 10 (N=90)

	Grammar	Meaning	Discourse	Socio	Task
Grammar	1				
Meaning	.970	1			
Discourse	.974	.962	1		
Socio	.912	.930	.905	1	
Task	.946	.945	.963	.909	1

Note. All correlations were significant at the 0.01 level (2-tailed).

Theoretically, the variables should be highly correlated with one another since they are hypothesized components of the same language competence, and yet the magnitude of correlation should not be too high since they are also supposed to be distinct. The correlation coefficients among the variables ranged from .91 to .97. As speculated, the variables are found to be highly correlated. More specifically, correlations among the three variables of meaningfulness, grammatical competence, and discourse competence are found to be very high, ranging from .96 to .97, while their correlations with sociolinguistic dimension were a little lower, ranging from .91 to .93. One possible reason for this is that the ability to speak appropriately in a given context might be a distinct dimension. However, in general, these values appear to be too high to conclude that any of the variables are indeed separate from one another.

Correlation coefficients, as an index of the *togetherness* of variables indicate how two specified variables vary together, but it does not explain what causes the variables to be correlated or uncorrelated. The reasons can only be speculated. In this case, one of the possible accounts for the high correlations is that the variables may not be separable, contrary to the model. Said differently, separate dimensions were specified in the model, but the raters might not have been able to separate one from another and produced similar scores. This is a problem associated with test design, particularly with regard to the rubric and raters. An alternative reason for the high correlations greater than .9 might be attributed to some extent to the low proficiency test-takers. Those who have limited English proficiency are very likely to score zero (not enough evidence) or one (limited) across all dimensions for all tasks. In other words, if an examinee only managed to speak one or two words, he or she would not receive different ratings. For the different dimensions of language competencies to be observed and scored reliably, more evidence should be presented to the raters. Including the scores of the limited proficiency students might have inflated the degree of correlations among the variables. However, this would not be true for the high proficiency group. As noted earlier, in the model of language ability, meaningfulness and grammatical, discourse, and sociolinguistic competence are speculated to be separate. That is, if a speaker produces a clear, meaningful message and receives a high score on meaning, it does not automatically mean that the response would be grammatically correct, sociolinguistically appropriate, and coherent, scoring high on all these dimensions. In other words, the high proficiency group theoretically should not inflate the correlations among the variables.

If the first account related to the test design issues (i.e., inadequate descriptors of each dimension in the rubric and/or insufficient rater training) is found to be responsible for the high correlations, it should be addressed carefully when revising the instrument. Therefore, to see which hypothesis is more plausible, a series of additional correlation analyses were conducted.

First, a group of examinees whose average scores were less than 1 point, indicating *not enough evidence* were removed from the data pool to see whether their scores inflated the correlation magnitude. Ten examinees were identified as such. Secondly, separate correlation analyses were performed at each level of proficiency (i.e., low—0 and 1, intermediate—2 and 3, and high—4 and 5).

Table 9 shows the correlation matrix obtained with 80 students after the ten examinees scoring zero on average were taken out. Although the correlation coefficients are still generally high, they became smaller than in the previous calculation. The differences in magnitude from the previous matrix with all 90 participants do not seem large. However, they may be still meaningful because no such drop in correlation values was found when the top ten participants were removed. This indicates to some extent that the participants with limited proficiency might have inflated the results.

TABLE 9
Correlation Analyses after taking out 10 Students (N=80)

	Grammar	Meaning	Discourse	Socio	Task
Grammar	1				
Meaning	.957	1			
Discourse	.962	.947	1		
Socio	.918	.921	.941	1	
Task	.873	.902	.854	.833	1

Note. All correlations were found significant at .05 level (2-tailed)

This claim is further supported by the correlation coefficients obtained for each proficiency level group. Separate group correlation analyses are reported in Tables 10, 11, and 12.

TABLE 10
Correlation Analyses for the Low Proficiency Group (N=15)

	Grammar	Meaning	Discourse	Socio	Task
Grammar	1				
Meaning	.917	1			
Discourse	.955	.894	1		
Socio	.958	.897	.964	1	
Task	.799	.836	.833	.851	1

Note. All correlations were found significant at .05 level (2-tailed)

TABLE 11
Correlation Analyses for the Intermediate Proficiency Group (N=53)

	Grammar	Meaning	Discourse	Socio	Task
Grammar	1				
Meaning	.896	1			
Discourse	.909	.858	1		
Socio	.797	.812	.862	1	
Task	.688	.737	.624	.648	1

Note. All correlations were found significant at .05 level (2-tailed)

TABLE 12
Correlation Analyses for the High Proficiency Group (N=22)

	Grammar	Meaning	Discourse	Socio	Task
Grammar	1				
Meaning	.801	1			
Discourse	.824	.753	1		
Socio	.513	.583	.467	1	
Task	.681	.332	.686	.449	1

Note. All correlations were found significant at .05 level (2-tailed)

As seen in Table 10, the degrees of correlations were the largest in the low proficiency group. At that level, different dimensions might not have emerged due to significantly limited linguistic resources. Another noteworthy point is that here again the three measures (meaningfulness, grammar, and discourse) exhibited relatively higher degrees of correlation than with the sociolinguistic and the task completion dimensions, regardless of proficiency levels. The coefficients of the sociolinguistic measure with the rest of the dimensions were lowest in the high proficiency level students. This implies that, for the high score group (4s and 5s), the ability to use language appropriately is not necessarily be related to other dimensions of language competencies but may be a distinct dimension of language proficiency, as speculated in the model. The dimension of task completion at the higher level also exhibited low correlations, indicating its distinctiveness along with sociolinguistic competence.

Although the correlation analyses provided some evidence regarding the nature of variables measured by the test, they are limited in that correlations, unlike more sophisticated statistical analyses such as structural equation modeling (SEM), include error variance as well as true variance. In other words, errors are inherently included in correlation coefficients obtained, and hence the relationships between the variables may be inflated or underestimated.

The FACETS Analysis

In this section, the results from FACETS analysis will be discussed in relation to examinee separability, item difficulty, scale functionality, and rater severity and consistency. Figure 1 shows graphically the measures for examinee ability, item difficulty, and rater severity. The first column in the figure displays the logit scale. The logit scale is a true interval scale in

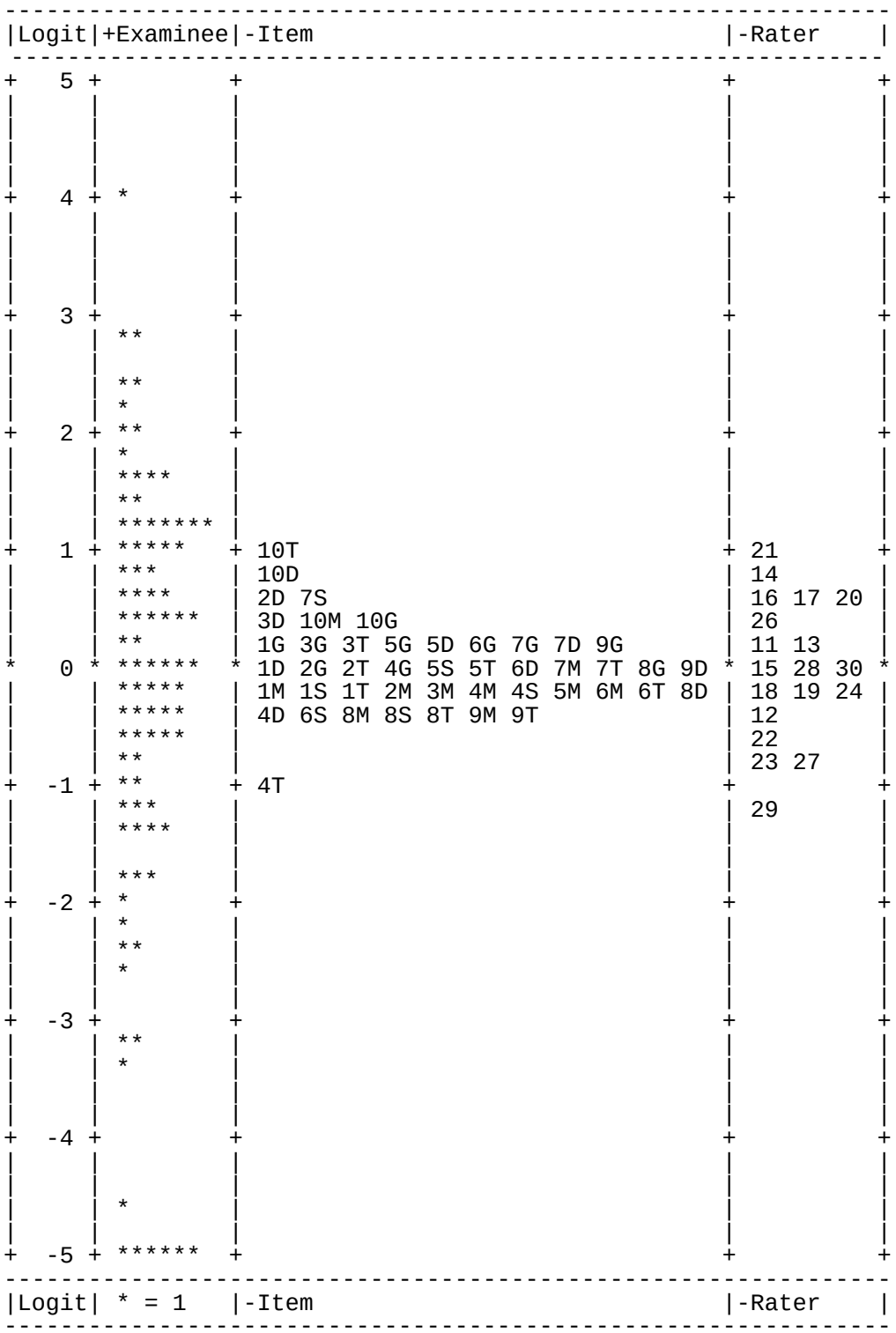
which the same distances are assumed between intervals. The FACETS program calibrates the examinees, raters, tasks, and ratings scales simultaneously so that all facets specified are positioned on the same logit scale. The second column displays estimates of examinee ability. More able examinees appear at the top of the column with higher logit values, while less able examinees appear at the bottom with lower logit values. The third column displays estimates of task difficulty. For this facet, ratings for each dimension (i.e., meaningfulness, grammatical, sociolinguistic, and discourse competence, and task completion) on each task (10 tasks) were treated as items. Since Tasks 2, 3, 9 and 10 were not evaluated for sociolinguistic competence, there were 46 items in total. Items appearing higher on the logit scale with higher logits were more difficult for examinees to receive high ratings on than on items appearing lower on the scale.

Examinees

As seen in Figure 1, the mean ability estimate for the present test-taker group was .1 logits (SD=1.58), ranging from -8.26 to 4.02 logits. The standard errors for the candidate estimates were acceptable (M=.12, SD=.02). The examinee separation reliability was .99. This is a measure of the extent to which the instrument could successfully separate candidates of varying ability. Like Cronbach's alpha or the Kuder-Richardson 21 index of reliability, the coefficient represents the ratio of variance attributable to the construct being measured to the observed variance (McNamara, 1996; Pollitt & Hutchinson, 1987). The value of .99 indicates that the current instrument was reliably separating examinees into different levels of ability. The overall difference between the ability level of the examinees was significant, $\chi^2(85) = 10034.1$, p=.00, and the hypothesis that all examinees were equally able was rejected. In short, the test appeared to spread out students successfully, which is an important feature for a placement test.

In order to identify examinees who exhibited unusual profiles of ratings, the infit mean-square statistics were examined. The infit mean-square measures indicate the degree of fit between the observed ratings and the ratings expected by the model. Some researchers (e.g., Englehard, 1994) have suggested an acceptable range between .6 and 1.5, and more conservatively between .7 and 1.3 logits. However, it has been argued that any individual infit mean-square value needs to be interpreted against the mean and the standard deviation of the set of infit-mean square values for the facet concerned (see Pollitt & Hutchinson, 1987). Using this criterion, a value lower than the mean minus twice the standard deviation would indicate too little variation, or *overfit*, while a value greater than the mean plus twice the standard deviation would indicate too much unpredictability, or *misfit*. For the examinee facet, the infit mean was 1.0, with a standard deviation of .4, producing fit criteria of .2 for overfit, and 1.8 for misfit. Three examinees were identified as misfitting, which represented 3% of the total participants. This figure is slightly higher than the acceptable percentage (i.e., 2%), suggested by Pollitt and Hutchinson. Examination of the scoring patterns for these examinees is necessary to determine the cause for the misfit.

FIGURE 1
FACETS Summary (Examinee Ability, Item Difficulty, Rater Severity)



Tasks

As seen in Figure 1, item difficulties ranged from -.94 to .91 logits, a range of 1.85 logits. Compared to the normally expected range of -3.0 to 3.0 logits (Myford & Wolfe, 2002), the range produced for the item facet seems somewhat restricted. However, the overall difference between the item difficulty estimates were significant, as indicated by $\chi^2(45) = 699.1$, p=.00, with a separation reliability of .93. These indices show that the items can be reliably separated into different difficulty levels. Task completion on Task 10 (10T) was found to be the most difficult item to get a high score, while task completion on Task 4 (10T) was the easiest. This finding corresponds to the results reported in the descriptive statistics section. The standard errors for all items were acceptable (M=.09, SD=.01).

The infit mean-square values indicating fit of individual items ranged from .7 to 1.8 (M=1.0, SD=.3). Following Pollitt and Hutchinson's (1987) criteria, the acceptable range of infit mean-square values was .4 to 1.6 (i.e., two standard deviations around the means). Three items out of 46 (6%) were identified as misfitting, all of which were ratings on task completion (1T, 6T, and 8T). One possible reason for this is the problem associated with the scale used to score task completion. The description of task completion for these tasks might not have been clearly defined in the rubric, and hence caused difficulty in scoring. Another possible reason is related to the nature of task completion. Misfitting items may signal multidimensionality, meaning that the items do not belong to the same measure as the others (McNamara, 1996). In other words, task completion might be a dimension distinctive from the rest of the variables. However, only the three tasks on task completion (Tasks 1, 6, and 8) were identified as misfitting, and thus the first account related to the rating scales sounds more plausible. Investigation of rating scale functionality might provide additional information regarding what could have been problematic and possibly how the scales could be improved. Therefore, the average examinee ability measure and outfit mean-square index for each rating category were examined, and will be discussed in the next section.

Rating Scale

Rating scale analyses using FACETS can be extremely useful to examine if the 6-point rating categories (0 to 5) used to score for each item are appropriately ordered and clearly distinguishable (Linacre, 1999). The average examinee ability measure for each rating point (0 through 5) is calculated by taking the average ability measure of all examinees receiving a rating in that particular category. If the rating scales are functioning correctly, it is expected that the average candidate ability will increase with each rating category, suggesting that examinees receiving higher ratings possess a higher level of ability. Another indicator used to examine rating scale functionality is the outfit mean-square index, which refers to the discrepancy between the average examinee ability measure (i.e., the observed measure) and an expected examinee ability measure predicted by the model. When the observed and expected examinee ability measures are close, then the outfit mean-square index for the rating category will be close to the expected value of 1.0. As the discrepancy between the observed and the expected measures increases, the mean-square index will be larger. An outfit mean-square index greater than 2.0 for a given rating category suggests that a rating in that category for one or more examinees may not be contributing to meaningful measurement of the variable (Linacre, 1999).

A review of the average ability measure for each scoring category for each item reveals that one category (i.e., 6T) displayed a reversed order between the rating category of 1 and 2, as shown in Table 13. The reversed rating categories 1 and 2 mean that examinees who received rating 1 were more proficient examinees than those who received rating 2 on this particular item. This again lends further evidence indicating a problem with the task completion scale.

TABLE 13
Category Statistics for Task Completion on Task 6

Rating Category	Average Measures	Expected Measures	Outfit Mean Square
0	-2.43	-2.35	1.0
1	-.06	-1.09	2.9
2	-.11	-.27	1.3
3	.38	.42	1.9
4	.96	1.19	1.8
5	1.96	2.22	1.3

The other two items that were identified as problematic in the previous analysis (1T and 8T) did not produce any reversely ordered categories. However, both of them contained rating categories displaying an outfit value bigger than 2.0, which indicates a large discrepancy between the observed ability estimate and the expected estimates. Two other items (1G and 1D) also produced large outfit mean squares on one of the rating categories. The results are presented in Table 14. Outfit values exceeding the acceptable level of 2.0 were underlined.

As explained, a high mean-square value indicates that this category has been used in contexts in which the expected category is far different. The finding that the problematic rating categories for all four cases were either 0 or 1 seems to indicate that a clearer depiction of 0 (No Evidence of Control) and 1 (Limited Evidence) is needed, particularly for Task 1.

TABLE 14
Average Examinee Ability Measures and Outfit Mean-square Indices from the FACETS Output

Rating Category	Task 1 Grammar		Task 1 Discourse		Task 1 Task Completion		Task 8 Task Completion	
	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq	Average Measures	Outfit MnSq
0	-2.27	2.6	-1.96	3.7	-1.67	3.7	-2.00	1.3
1	-1.39	1.1	-1.40	.8	-1.16	.8	-.68	2.4
2	-.59	.9	-.51	.8	-.38	1.6	-.02	.8
3	.35	1.2	.40	1.3	.41	1.5	.63	1.4
4	1.53	.8	1.30	1.1	1.03	1.3	.87	1.9
5	2.50	1.0	2.30	1.1	2.17	1.0	2.05	1.5

Raters

In a performance test, the effects of rater variation are a major concern. The inter-rater reliability analysis using correlational analysis provided some information about how consistent the raters were in scoring the responses. While inter-rater agreement indicated by correlation coefficients is informative, it is only a minimum step to evaluate rater behaviors and to establish a reliable and valid assessment of performance (Hamp-Lyons, 1990). Therefore, additional analyses employing FACETS were performed. Table 15 provides a summary of selected statistics on the rater facet.

In Table 15, rater IDs, rater severity, error, and infit mean-square values are reported. The second column shows estimates of severity or leniency. Rater 20 was the most severe rater while Rater 29 was the most lenient. The difference between the most severe and lenient was about 2 logits. The reliability of separation index, which indicates the likelihood to which raters consistently differ from one another in overall severity, was high at .99. The high separation reliability index indicates that the raters differed significantly in the severity estimates. Errors ranged from .04 to .08, which would be considered small. This indicates accuracy of the estimates obtained. The last column indicates infit mean-squares, indicating the degree of fit. Fit values for all raters except Rater 30 were within the range of two standard deviations around the mean of the infit measure. In this case, the mean of infit mean-square was 1, with a standard deviation of 0.2. Infit mean-square values greater than 1.4 would be considered misfitting. A misfitting rater, Rater 30 in this case, needs further training. All other raters seemed self-consistent in scoring.

The FACETS analysis, in sum, produced detailed information regarding rater behavior in terms of their severity and consistency. As noted earlier, consequences of different levels of severity and inconsistency in rating can be serious, especially when raw scores are used to make inferences about test-takers' abilities. This underscores the importance of providing good training and feedback to raters so that they become more consistent and fall within a similar range of severity. Positive effects of feedback in rater training are well-documented in empirical studies (e.g., Tyndall & Kenyon, 1996; Wigglesworth, 1993). Therefore, individual reports showing how the raters applied the scales were produced. Like in the rating scale analysis, the average examinee ability measure per category and outfit mean-square values were examined to see whether there were any reversely ordered categories and how large the discrepancies between the observed and expected ability estimates were.

TABLE 15
Rater Measurement Report

Rater ID	Rater Severity (logits)	Standard Error	Infit Mean-Square Index
20	1.05	.07	1.2
16	.85	.07	0.9
14	.71	.07	1.2
21	.71	.05	0.9
17	.46	.07	1.1
26	.22	.05	0.8
11	.20	.07	0.8
13	.17	.07	1.0
28	.09	.06	0.7
18	.01	.07	0.9
30	.00	.05	1.7
15	-.01	.04	0.9
19	-.08	.07	0.9
24	-.20	.05	1.0
22	-.57	.05	0.9
23	-.76	.05	1.0
12	-.80	.08	1.1
27	-.85	.05	0.9
29	-1.19	.05	1.2
Mean	.00	.06	1.0
SD	.60	.01	.02

Separation: 9.74; Reliability of separation index= .99; Fixed (all same) chi-square: 1955.7 d.f.: 18 significance: .00

The analyses revealed that two raters inappropriately applied the rating scales. The results are presented in Table 16 and 17.

TABLE 16
Selected Category Statistics For Rater 12

Rating Category	Average Measures	Expected Measures	Outfit MnSq
0	-1.21	-2.63	1.0
1	-2.83	-1.96	2.9
2	.13	.09	1.3
3	1.12	1.22	1.9
4	2.19	2.06	1.8
5	3.30	3.42	1.3

TABLE 17
Selected Category Statistics For Rater 30

Rating Category	Average Measures	Expected Measures	Outfit MnSq
0	-2.32	-2.37	1.3
1	-.92	-1.37	2.3
2	-.50	-.68	1.5
3	.22	.11	1.4
4	.51	1.00	2.9
5	1.39	1.84	1.4

Rater 12 showed a rating pattern in which category 0 and 1 were reversed. The average measure values for categories 0 and 1 are disordered, and category 1 is exhibiting a large outfit

mean-square value (2.9). This means that the examinees who received rating 0 by Rater 12 were in general more proficient than examinees who received 1. In other words, Rater 12 was reversely scoring for 0 and 1. This rater might not have a clear distinction between 0 and 1, resulting in the aberrant rating pattern. Another rater (Rater 30), who was identified as misfitting, did not exhibit reversely ordered categories, but had two large outfit mean-square values exceeding the acceptable level of 2. The finding illustrates that Rater 30 was not able to distinguish one level of performance from another.

The graphs (Figure 2 and 3) visually represent how the raters applied the rating scale. The horizontal axis represents the examinee proficiency scale (in logits), and the vertical axis represents probability (from 0 to 1). There is a probability curve printed for each of the scale categories. In the graphs, a separate peak for each scale category probability curve denotes that each rating point was clearly distinguished from one another. The low peak for rating category 1 in Figure 2 signals a problem associated with that particular rating point. Except for that category, other scale categories appear to be well applied by Rater 12, with separate, distinguished peaks. The curves in Figure 3 do not show clearly distinguishable peaks for rating categories, particularly 1 and 2. The probability curves as well as the outfit mean-square of Rater 30 clearly indicate a need for further training

FIGURE 2
Scale Category Probability Curves for Rater 12

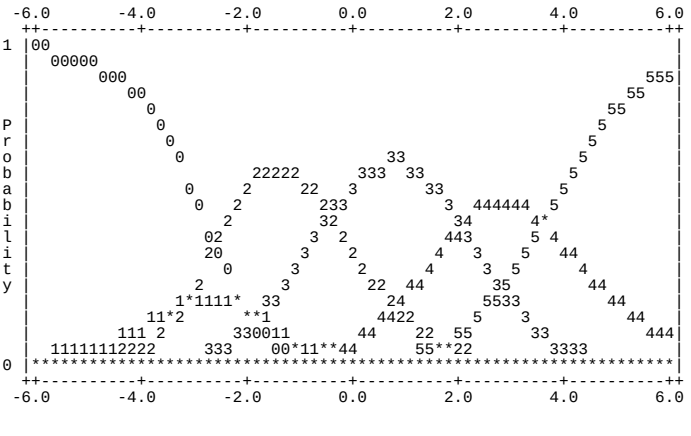
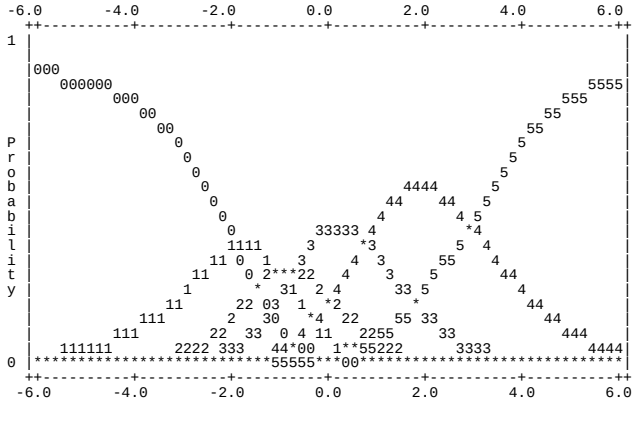


FIGURE 3
Scale Category Probability Curves for Rater 30



In sum, the results produced from the FACETS analysis provide much richer information regarding items, rating scales, and raters. The test appears to spread out the examinees well. Some items were found misfitting, possibly due to inadequate descriptors used to score. Rating scale analyses revealed how the rating scales for each item were functioning. Some scales, particularly task completion scales, need to be revised. Finally, in terms of the rater facet, variability was found in severity among the raters, in addition to one misfitting rater. This finding raised some concerns, and needs to be addressed in future training.

DISCUSSION AND CONCLUSION

With regard to Research Question 1, "What is the nature of communicative language ability measured by the current speaking test?," the results seemed to indicate highly related components of the construct. Correlation analyses revealed that the hypothesized language domains were greatly correlated with one another. Highly correlated scores generated by the test, thus, provided limited evidence for the construct validity of the four-dimension model of language ability measured by the speaking test. Although the four dimensions failed to emerge, it was found that the strengths of the relationships among the hypothesized components varied. Sociolinguistic competence, for example, appeared to have relatively weaker relationships with other components, indicating its distinctiveness to some degree.

The potentially distinct nature of the dimensions was also supported by the results from the additional correlation analyses performed at the three different proficiency levels. At the very limited proficiency level, it was fairly predictable that the scores on each of the four language dimensions would be more highly related and indistinguishable because of the lack of language proficiency. However, at the advanced proficiency level, correlations among the variables became smaller, providing evidence for the related and yet separate nature of the dimensions. In other words, high grammatical ability does not necessarily mean a comparable level of proficiency in other dimensions, such as in sociolinguistic competence.

Regarding the second research question, "What is the nature of the task-dependent measure of task completion as measured by the current speaking test?," the task-dependent measure, task completion, exhibited a close relationship with the language ability dimensions. However, when separate analyses were conducted at each proficiency level, task completion had significantly weaker correlations with other linguistic measures, especially in the advanced group of students. This finding indicates that task completion might be a distinct dimension when speakers have a full array of language ability, and should not be used as an indicator of language ability. Furthermore, the result that the range of means as well as the difficulty estimates produced by the FACETS analysis for task completion varied notably from one task to another seems to suggest performance variation due to task characteristics. This finding may serve as counterevidence to Norris et al.'s (2002) argument for the use of one task-dependent measure (i.e., task completion) to predict test-takers' abilities to perform other tasks. The nature of the task-dependent measure, including the abilities and processes involved in completing a task, needs to be investigated first. Without detailed analyses of the measure, generalization and extrapolation as well as interpretation of the observed scores would be greatly limited.

With regard to Research Question 3, "To what extent has the test separated examinees into distinct levels of speaking ability?," the investigation of the examinee facet showed that the CEP speaking test reliably separated test-takers into distinct levels of proficiency. A few test-takers were identified as misfitting, and so their score reports and responses need to be examined in detail to find out the cause.

With regard to Research Question 4, "How appropriately are the rating scales functioning?," the FACETS analysis provides evidence that the ratings scales were functioning well, except for task completion for Task 6. Task 6 had a reversely ordered rating category. Since task completion is a task-specific measure, descriptors specific to the task may need to be provided in detail. There were also three other rating categories with outfit-mean square measures exceeding the acceptable value. These results imply that the rating scales for some measures need to be revised.

With regard to Research Question 5, "To what extent do the raters vary in terms of severity and consistency?," the examination of the rater facet showed that there were variations in severity among the raters. Furthermore, one rater was found misfitting, which means he or she was not consistently scoring. Given the importance of rating quality in making inferences about test-takers' ability, rater variations should be examined and monitored thoroughly and continuously. To conclude, this paper has explored the development and construct validation of a speaking test. The test was developed based on theories of communicative language ability in relation to various communicative situations. Although the test did not produce separate language ability dimensions as hypothesized in the model of speaking ability, it appeared to be measuring one underlying ability. Additionally, FACETS analyses yielded valuable information regarding the instrument, raters and rating scales. Overall, the study provided some evidence suggesting the newly developed measure seemed to be functioning as intended.

REFERENCES

Bachman, L. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*, 149-164.

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

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.

Bachman, L., & Clark, J. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science, 490*, 20-33.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

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

Brown, J. D., Hudson, T., Norris, J., & Bonk, W. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawai'i Press.

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge, UK: Cambridge University Press.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333-342). Rowley, MA: Newbury House.

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

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language teaching*. Cambridge, UK: Cambridge University Press.

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

Chapelle, C., Grabe, W., & Berns, M. (1997). Communicative language proficiency: Definition and implications for TOEFL 2000. *TOEFL Monograph Series, 10.* Princeton, NJ: Educational Testing Service

Crystal, D. (1997). *A dictionary of linguistics and phonetics*. Oxford, UK: Basil Blackwell.

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

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Faerch, C., & Kasper, G. (1984). Two ways of defining communication strategies. *Language Learning, 34*, 46-63.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing* (pp. 69-87). Cambridge, UK: Cambridge University Press.

Hudson, T., Detmer, E., & Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics* (Tech. Rep. No. 7). Honolulu: University of Hawai'i Press.

Hymes, D. (1972). On communicative competence. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics* (pp. 35-71). New York: Holt, Reinhart, & Winston.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: Pennsylvania University Press.

Johnson, M. (2001). *The art of non-conversation*. New Haven, CT: Yale University Press.

Lado, R. (1961). *Language testing*. London: Longman.

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

Lantolf, J., & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*, 181-195.

Larsen-Freeman, D. (1991). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 299-296). Boston: Heinle & Heinle.

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Linacre, J. (1999). Investigating rating scale category unity. *Journal of Outcome Measurement, 3*, 103-122.

Linacre, M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Lynch, B., & McNamara, T. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-104). Phoenix, AZ: Oryx Press.

Myford, C., & Wolfe, E. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (Research Rep. No. 65). Princeton, NJ: Educational Testing Service.

Norris, J., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*, 395-418.

Norris, J., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessments* (Tech. Rep. No. 18). Honolulu: University of Hawai'i.

Pollitt, A., & Hutchinson, G. (1987). Calibrated graded assessment. *Language Testing, 4*, 72-92.

Purpura, J. (2004). *Assessing second language grammar ability*. Cambridge, UK: Cambridge University Press.

Ranney, S. (1992). Learning a new script: An exploration of sociolinguistic competence. *Applied Linguistics, 13*, 25-50.

Read, J., & Chapelle, C. (2001). A framework for L2 vocabulary assessment. *Language Testing, 18*, 1-32.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17*, 289 - 310.

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*, 27-33.

Swales, J. (1990). *Genre analysis: English in Academic and research settings*. Cambridge, UK: Cambridge University Press.

Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning, 35*, 373-404.

Tarone, E., & Liu, G. (1995). Situational context, variation and second language acquisition theory. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in the study of language and learning* (pp. 107-124). Oxford, UK: Oxford University Press.

Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics, 4*, 91-112.

Tyndall, B., & Kenyon, D. (1996). Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39-57). Clevedon, UK: Multilingual Matters.

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.

Van Lier, L. (1989). Reeling, writhing, drawing, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*, 305-335.

Wright, B., & Masters, G. (1982). *Rating scale analysis*. Chicago: MESA Press.

Young, R., & He, A. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Philadelphia: John Benjamins.

APPENDIX A

Scoring Rubrics

Meaningfulness

5 Excellent	4 Good	3 Adequate	2 Fair	1 Little	0 No
The response • is completely meaningful—What the speaker wants to convey is completely clear and easy to understand.	The response • is generally meaningful—in general, what the speaker wants to convey is clear and easy to understand.	The response • occasionally displays obscure points; however, main points are still conveyed.	The response • often displays obscure points, leaving the listener confused.	The response • is generally unclear and extremely hard to understand.	The response • is incomprehensible. • not enough evidence to evaluate.
• is fully elaborated.	• is well elaborated.	• includes some elaboration.	• includes little elaboration.	• is not well elaborated.	
• delivers sophisticated ideas.	• delivers generally sophisticated ideas.	• delivers somewhat simple ideas.	• delivers simple ideas.	• delivers extremely simple, limited ideas.	

Grammatical Competence: Accuracy, Complexity and Range

5 Excellent	4 Good	3 Adequate	2 Fair	1 Limited	0 No
The response • is grammatically accurate.	The response • is generally grammatically accurate without any major errors (e.g., article usage, subject/verb agreement, etc.) that obscure meaning.	The response • rarely displays major errors that obscure meaning and a few minor errors (but what the speaker wants to say can be understood).	The response • displays several major errors as well as frequent minor errors, causing confusion sometimes.	The response • is almost always grammatically inaccurate, which causes difficulty in understanding what the speaker wants to say.	The response • displays no grammatical control.
• displays a wide range of syntactic structures and lexical form.	• displays a relatively wide range of syntactic structures and lexical form.	• displays a somewhat narrow range of syntactic structures; too many simple sentences.	• displays a narrow range of syntactic structures, limited to simple sentences.	• displays lack of basic sentence structure knowledge.	• displays severely limited or no range and sophistication of grammatical structure and lexical form.
• displays complex syntactic structures (relative clause, embedded clause, passive voice, etc.) and lexical form.	• displays relatively complex syntactic structures and lexical form.	• displays somewhat simple syntactic structures. • displays use of somewhat simple or inaccurate lexical form.	• displays use of simple and inaccurate lexical form.	• displays generally basic lexical form.	• not enough evidence to evaluate.

Sociolinguistic Competence: Appropriateness and Naturalness

5 Excellent	4 Good	3 Adequate	2 Fair	1 Little	0 No
The response • is appropriate socioculturally given the context (i.e., degree of politeness and/or formality according to contextual features like power status and distance).	The response • is in general appropriate socioculturally given the context without any serious face-threatening violations.	The response • is at times inappropriate socioculturally, but generally as a whole appropriate (e.g., too direct, too polite, etc.); generally shows awareness of the contextual features.	The response • is often inappropriate, except sporadic use of appropriate language that are fixed expressions; displays an inconsistent level of politeness and formality.	The response • is generally inappropriate; displays little awareness and/or misunderstanding of contextual features, resulting in inappropriate language.	The response displays • No control over appropriate language use. • Not enough evidence to evaluate.
• sounds completely natural and idiomatic.	• sounds generally natural and idiomatic with very few awkward expressions.	• sounds at times awkward.	• sounds often awkward.	• generally sounds awkward.	

Discourse Competence: Organization and Cohesion

5 Excellent	4 Good	3 Adequate	2 Fair	1 Little	0 No
The response • is completely coherent.	The response • is generally coherent.	The response • is occasionally incoherent.	The response • is loosely organized, resulting in generally disjointed discourse.	The response • is generally incoherent.	The response • is incoherent.
• is logically structured—logical openings and closures; logical development of ideas.	• displays generally logical structure.	• contains parts that display somewhat illogical or unclear organization; however, as a whole, it is in general logically structured.	• often displays illogical or unclear organization, causing some confusion.	• displays illogical or unclear organization, causing great confusion.	• displays virtually non-existent organization.
• fully displays smooth connection of ideas with sophisticate use of various cohesive devices (transitional words & phrases, a controlling theme, repetition of key words, etc.).	• displays good use of cohesive devices that generally connect ideas smoothly.	• at times displays somewhat loose connection of ideas. • displays use of simple cohesive devices.	• displays repetitive use of simple cohesive devices; use of cohesive devices are not always effective.	• displays attempts to use cohesive devices, but they are either quite mechanical or inaccurate leaving the listener confused.	• contains not enough evidence to evaluate.

Task Completion

5 **Excellent Understanding**	4 **Good Understanding**	3 **Adequate Understanding**	2 **Fair Understanding**	1 **Limited Understanding**	0 **No Understanding**
The response • fully addresses the task. • displays completely accurate understanding of the prompt without any misunderstood points. • completely covers all main points with complete details discussed in the prompt.	The response • addresses the task well. • includes no noticeably misunderstood points. • completely covers all main points with a good amount of details discussed in the prompt. (e.g.,) Car Dealer: car being back-ordered, discount offer for the alternative color Deadline: student's problem with partner and working full time Electric Cars: two problems with the current technology (battery running out quickly and inconvenience in recharging) Barbizon School: 2 characteristics of the school and one example (painted nature and established landscaping as an independent genre, and the Forest in the sunset example)	The response • adequately addresses the task. • includes minor misunderstanding(s) that does not interfere with task fulfillment. OR • touches upon all main points, but leaves out details. OR • completely covers one (or two) main points with details, but leaves the rest out.	The response • insufficiently addresses the task. • displays some major incomprehension/ misunderstanding(s) that interferes with successful task completion. OR • touches upon bits and pieces of the prompts.	The response • barely addresses the task. • displays major incomprehension/ misunderstanding(s) that interferes with addressing the task.	• The response shows no understanding of the prompt. • The response does not provide enough evidence to evaluate.