RATER GROUP BIAS IN THE SPEAKING ASSESSMENT OF FOUR L1 JAPANESE ESL STUDENTS

HEATHER L. CABAN

University of Hawai'i

ABSTRACT

The purpose of this study, modeled after Kobayashi's (1982) investigation of writing evaluations, is to determine whether the factors like language background and educational training affect raters' assessments of four Japanese medical students in a controlled oral interview. Rater groups include ESL-trained L1 English and L1 Japanese speakers, a peer group, and native speakers with no ESL background and negligible contact with Japanese L1 speakers. The four interviewees are first rated according to seven categories: grammar, fluency, content, pronunciation, pragmatics, compensation techniques, and overall intelligibility, and are then ranked from "most able" to "least able". Group means per category are calculated and a FACETS (Linacre, 1996) analysis is used to look at the interaction between rater groups and rating categories. The findings show variation, though not seemingly as a result of the factors being investigated. Some general conclusions about the rater groups are made from the bias analysis. The limitations of the study are presented, and suggestions for further research are provided.

INTRODUCTION

Oral interviews are required in a growing number of language programs and formal language proficiency measures because they provide insights about L2 language skills that may not be measured on paper-and-pencil tests. Oral interviews are appealing in that they seem to reinforce the beliefs that form the basis of communicative language teaching (with its emphasis on "natural language"), which is now being used in increasing numbers of language classrooms all over the world. Oral interviews are also useful because they have a relatively high degree of content and face validity (Bachman & Palmer, 1990; Weir, 1990).

However, several cautions must be taken into consideration in interpreting the results of such measures. Since interviews require ratings by human observers, it is difficult to

Second Language Studies, 21(2), Spring 2003, pp. 1-44.

escape some degree of subjectivity. Raters may be biased by such factors as age, L1 background, sex, and educational level. There is also no guarantee that raters will ask the same questions in the same manner, and the range of situations/topics that they cover may be limited. Finally, the rating scale used, holistic or analytic, may also prove problematic. This paper attempts to examine rater differences in oral assessment focusing on two factors: L1 background and educational training. The goals of the paper are to evaluate teacher beliefs about oral assessment measures and explore the issue of designing "fair and ethical" speaking tests.

Communicative Competence

At the heart of L2 speaking assessment, lies the idea of *communicative competence*. Canale and Swain defined *communicative competence* as "the underlying systems of knowledge and skill required for communication" (1980), including such aspects as vocabulary knowledge or sociolinguistic conventions. They contrasted communicative competence with what they termed "*actual communication*," sometimes referred to as performance. Actual communication is the manifestation of this underlying competence "under limiting psychological and environmental conditions such as memory and perceptual constraints, fatigue, nervousness, distractions, and interfering background noises." Communicative competence is an integral part of actual communication. However, it is demonstrated indirectly and often imperfectly due to the constraints mentioned above.

Although there are a variety of frameworks for communicative competence, I will focus on an early one by Canale (1983). Canale's framework lists four main components: grammatical competence, sociolinguistic competence, discourse competence, and strategic competence. According to his definition, grammatical competence includes features that were traditionally associated with language tests, focusing on such areas as rules of the language, vocabulary, word formation, pronunciation, sentence formation, etc. In the communicative language framework, this component makes up only a fraction of the knowledge that students require and, therefore, is "not the important concern for any language program."

The second component, sociolinguistic competence, takes into account the contextual relationships of language. Speakers should be able to tailor their language appropriately according to the age and social status of the interlocutor and the purpose of interaction following the norms and conventions of the language. Until fairly recently this aspect has been ignored, and it is seldom explicitly included in language tests or language programs.

Discourse competence, the third component according to Canale, concerns "mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres." These genres refer to the type of text. In terms of speaking this might include: debate, narrative, formal meeting, casual conversation, complaint, etc. For the text to be unified it must be cohesive in form and coherent in meaning. Cohesion here refers to the structural linking of utterances using devices such as synonyms, pronouns, ellipsis, and conjunctions. Coherence deals with relationships among different meanings in the text, which may or may not appear outwardly cohesive. The meanings may be literal, implied, or strictly communicative functions.

The last component, strategic competence, appears to be one of the most important in spoken communication. This area seems to encompass the "dynamic" nature of communication, which Savignon (1983) introduced and several others have alluded to since. Strategic competence involves the skills needed to maneuver through breakdowns in communication due to such factors as misunderstandings and low proficiency. It allows speakers to get their message across despite hindrances in language ability or environment. Those who are proficient in this area will be able to exploit their language knowledge to the fullest and will ultimately be the most able to demonstrate their knowledge to those who are assessing them.

Canale stresses that these four components of communicative competence, or communicative ability, are "levels of analysis" that interact with each other in some way and only provide a starting point for creating a working model. Such a modular framework was also held by Lado (1961, 1964) and Carroll (1972) and has been adapted and adopted by others, and enhanced by a large amount of empirical evidence. Bachman and Palmer (1990), for example, found that their testing data could be accounted for by a multidimensional model. Several factor analytic studies have been carried out to test the multidimensionality of competence, with varying results. For a complete review of such studies, along with empirical evidence for the opposing unitary language factor view, please refer to Villmer and Sang (1983). This paper will adhere to the framework proposed by Canale.

The competence factors discussed in Canale have major implications for language teaching and language testing. Since observed proficiency appears to be associated with willingness and ability to adapt one's language and circumvent hindrances in language over grammatical accuracy, students should be presented with "naturalistic" situations, that is, tasks should mimic real-life encounters as much as possible. The more genuine the communicative interaction encountered by learners, the more they will be able to develop and apply the above competencies.

Testing and Teaching Communicative Language

The communicative teaching approach is now widely employed in language classrooms in the United States and is being transitioned into many countries. Japan, South Korea, and Taiwan have all been phasing it into their curriculum; however, for many teachers, the concept remains poorly defined and the goals are unclear. Added to this are high-stakes national exam systems rooted in more traditional concepts, such as grammar-translation. These tests have particularly destructive effects on classroom situations and limit opportunities for change. Writing about Japan, Buck says, "There is a natural tendency for both teachers and students to tailor their classroom activities to the demands of the test, especially when the test is very important to the future of the students, and pass rates are used as a measure of teacher success" (1988). He credits this with being one of the main reasons that Japanese high school graduates cannot use English in even basic situations. Though this claim is supported by little empirical evidence, it is widely believed that changes in the tests must occur before successful change can take place in curriculum. This effect of the test on teaching, whether it is positive or negative, is known in language testing as *washback.*

Communicative language tests (CLTs) should work toward positive washback. Bailey (1996) offers some useful ways to realize this goal. First, tests should be developed according to clear and authentic language learning goals. "By making our tests more reflective of the kinds of situations, language context and purposes for which second language speakers will need these skills, we will be able to make accurate predictions about how they will be able to function using the target language in 'real life'." Rather than demonstrating ability to memorize vocabulary and answer discrete grammar questions, exams should seek to empower students by testing skills which they can directly apply in the future. Otherwise, one wonders what the point of studying the language is to begin with.

CLTs should also highlight learner autonomy in the form of self-assessment. Bailey suggests that a questionnaire be included in the examination information booklet. She states: "Some…contend that self-assessment will promote learner involvement and autonomy. …Learners should have a large amount to say about what, how and how fast they learn. It incorporates principles of choice, intrinsic motivation, attention focus and personal evaluation (Bailey, 1996)." Rather than view themselves as unwilling victims of the test, learners should see themselves as active participants who have responsibility for their own learning outcomes.

Finally, and perhaps most importantly, standardized CLTs must be accountable for providing assessment information that is "detailed, innovative, relevant, and diagnostic and addresses a variety of dimensions" (Shohamy, 1992, p. 515). Spolsky warns that ensuring that tests do not lend themselves to too quick an interpretation is a moral responsibility akin to accuracy in scoring (1985). Rather than a holistic score which can prove meaningless to students, score reports should be broken down as much as possible. Wall suggests including the type of skill and test item reported for the country as a whole, each region, each district and each school (1996). The examining authorities might provide this information to teachers, as well as a wide range of parties including school officials at the local and government levels. Such an extensive reporting system would allow teachers and officials to easily pinpoint areas of weakness.

According to Canale, "Language tests should ideally reflect the properties of communication: contextuality, productivity, process orientation, interactivity, and adaptivity" (1983). Tests should be made up of context-dependent tasks which are interesting and pertinent to the test takers. In order to do this, the subject areas that students are studying should be identified and tests should be tailored around them. In this way the necessary "contextualization clues" will be provided, granting students the best opportunity to demonstrate their competence. This supports the claims made for tests of language for specific purposes (LSP) since, in an LSP test, the student is using the language relating to the actual field in which they intend to apply the language. Because ethical tests should present the optimum possibility for success, such contextualization is advisable and necessary.

The Oral Interview and Rater Reliability

As part of the fervor created by the communicative language movement, many language tests are now requiring or considering a speaking component. In most cases, this speaking component takes the form of a (controlled) oral interview. This method is convenient in that it allows for easy comparison and has a higher degree of content and face validity, as well as reliability, than other methods. However, there are several cautions when it comes to administering these types of tests and interpreting the scores (Weir, 1990; Salaberry, 2000).

Perhaps the greatest concern rests with the raters. There is no guarantee that all raters will ask the same questions or will ask them in the same manner (Weir, 1990). This can make a crucial difference in the output of students and, ultimately, in the scores they are given. Raters may also arrive at their evaluations in different ways. "While one rater may focus on pronunciation accuracy, another may find vocabulary to be the most salient feature. Or one rater may assign a rating as a percentage, while another might rate on a scale from zero to five" (as cited in Salaberry, 2000). Rater training with a detailed set of assessment criteria can do much to alleviate some of these inconsistencies. However, it "cannot easily eliminate a rater's overall tendency for severity or leniency." Raters can be biased toward specific candidates and tasks and affected by such factors as L1 and academic backgrounds. Such biases can particularly affect those students who are in-between levels in cases where cut-point decisions have to be made. For this reason, inter- and intra-rater consistencies should be investigated.

Multifaceted Rasch measurements, implemented using programs such as FACETS (Linacre & Wright, 1992), offer one way to examine rater bias. Through this approach, the ability measure for each candidate is can be determined through analysis of the interaction between three main "facets": candidate, item, and rater, along with various other optional "facets", including test type, candidate's country of origin, candidate's language background etc. The FACETS program also offers a unique form of analysis called *bias analysis*, which can be used to examine rater characteristics with respect to their evaluation patterns of specific items and candidates (Kondo-Brown, 2002; Lumley & McNamara, 1995; McNamara, 1996). FACETS analysis operates on the premise that, although one particular rater may be scoring differently from another rater for a particular examinee, as long as they are internally consistent in their rating patterns, they "fit" the model (Wigglesworth, 1993).

A number of studies have been done using FACETS analysis in L2 oral assessments, the majority of which are in relation to intra-rater differences and rater training (Bachman *et al*. 1995; Lumley & McNamara, 1995; Wigglesworth, 1993). These studies show that although rater training may have some evident short-term effects (Wigglesworth, 1993), these changes do hold over time (Lumley & McNamara, 1995). Brown (1995), in a departure from the previous type of studies, used the FACETS program to investigate group differences based on L1 and occupational training, in a test of spoken Japanese for future Australian tour guides. She found that there was no support for the belief that native speakers were more suitable than non-native speakers, or that raters with a teaching background were more suitable than those with a background in tourism. She did, however, note that there were differences in the manner in which the groups perceived certain items, and that, if each group were to develop its own assessment scale rather than using the one provided, evaluations of the candidates' abilities might have been quite different.

Purpose

The present study is an adaptation of Kobayashi's 1992 study on "Native and Nonnative Reactions to ESL Compositions", incorporating a FACETS analysis approach. Whereas Kobayashi's study dealt with writing assessment, the present study analyzes group rater differences in controlled oral interviews using seven rating categories (grammar, fluency, content, pronunciation, pragmatics, compensation techniques, & overall intelligibility) and four rater groups (ESL-trained L1 English speakers, EFL-trained L1 Japanese speakers, a peer group of Japanese speakers, and native speakers of

English with no ESL background and negligible contact with Japanese speakers). The research questions were the following:

1. What are the relative difficulties of the seven rating categories?
2. How do the four rater groups differ in their ratings of the four speakers on the seven rating categories?
3. Will the four rater groups differ in their overall ranking of the four L1 Japanese English speakers?
4. Of the four rater groups, which is the harshest and which is the most lenient in overall evaluations of the four speakers?
5. Do any of the four rater groups exhibit bias toward any of the four speakers?
6. Do any of the four rater groups exhibit bias toward any of the rating categories?
7. What are the implications for communicative language teaching and testing?

The researcher believed that the two L1 Japanese speaker rater groups would score the more traditional categories more severely than the native speaker raters. Since grammar constitutes a large part of the English section on the university entrance exam, it is the major focus of Japanese EFL classes; therefore, these raters might put greater emphasis on and be more sensitive to grammaticality. Also, pronunciation, although not explicitly taught, is also an area in which many ESL/EFL students are sensitive. It was also hypothesized that the two trained Master's degree groups would rank speakers similarly since they have received similar training.

METHODOLOGY

Participants

A total of 83 rater participants were asked to rate four non-native speakers (i.e., four 23-year old female Japanese medical students in a controlled oral interview). They were organized by L1 (English & Japanese) and academic rank/professional status, with cell sizes ranging from 15 to 29. The participants came from three main backgrounds. The first consisted of MA students in the SLS department at the University of Hawaii at Manoa (age M= 29.8). Of these, there were two groups: 16 English L1 speakers (NMESL) and 19 Japanese L1 speakers (JMESL). The third group (NT) consisted of L1 speakers (age M= 40.2) , who were teachers at a middle school in rural Connecticut, none of whom had any sustained contact with Japanese L1 speakers. In creating this group, the researcher hoped to see whether or not there was an effect due to lack of exposure to Japanese L1 speakers of English. The fourth group consisted of peers (PEER), whose mean age was 27.1. Peers were defined as Japanese L1 students enrolled in equivalent level ESL programs at two language institutes in Hawaii.

Materials

The four 23-year old female L1 Japanese medical students involved in this study, who were enrolled in a one-month language exchange program at a language institute in Hawaii, were asked to participate in two speaking tasks given by the researcher. The tasks included a controlled interview and a role-play, which combined, lasted about 15 minutes each. Both tasks were recorded using an MD recorder. Since the students were studying to be future doctors and were interning in American hospitals during their stay, the researcher decided to make the controlled interview portion relevant to their field. Student participants were asked such questions as “Why they began studying medicine?” and “What are the differences between American hospitals and Japanese hospitals?" For the role-play portion, student participants were asked to give advice about their city in Japan to a friend (the interlocutor) who is visiting there soon (for full directions and a list of the questions, see Appendix A).

Though the researcher originally wanted to investigate differences between scores and rankings reported for the interview versus the role-play, only the controlled interviews were ultimately chosen due to the unreasonable amount of time it would have taken raters to listen to both. The controlled interview portion was chosen because it is the most common form of oral assessment in standardized tests and language programs at the present. The interviews were unmodified and positioned in the same random order on multiple copies of a compact disk (CD). From that point on, the speakers were labeled according to their numerical position on the CD so their identities could be kept anonymous.

Procedures

Copies of the CDs were handed to willing rater participants along with a questionnaire and assessment sheet (Appendix B). Complete directions were written on the first page and at the top of each attached page, and directions were also given orally when the CD was distributed. For the L1 English group in Connecticut, materials were sent and received via mail to a contact at the participating school. Materials for the peer group were written in Japanese. All other groups received materials in English. Rater participants were asked to complete the following steps.

First, the rater participants were asked to complete the questionnaire, which included basic background information and teaching experience. It also included questions about views on oral language assessment, self-rating, and peer rating. The questionnaires differed in minor ways according to group (see Appendix B for complete basic questionnaire).

Second, the rater participants were asked to begin listening to the CD. There was a ten second pause between each speaker on the CD. When the listener heard this break in the CD, they were asked to manually pause the CD player and take as long as they needed to rate the previous speaker. Then, they were to listen to the next speaker following the same procedure, and so on. The rater directions (see Appendix C) required rater participants to rate speakers in seven categories: fluency, grammar, pronunciation, compensation techniques, content of utterance, language appropriateness, and overall intelligibility. Brief descriptions/definitions were given after each point, with the exception of grammar and pronunciation, which seemed less ambiguous. Evaluations were given using a Likert scale ranging from one (very poor) to fifteen (excellent). The researcher investigated several rating measures before choosing this one due to its relative ease of application. It seemed unreasonable to expect the peer group and those untrained in ESL to be familiar with field specific terminology and classifications found on the majority of holistic and descriptive scales. A range of fifteen was chosen in hopes of enhancing differences that might not be as noticeable on a smaller scale. The researcher feared that the relatively homogeneous proficiency levels of the student participants might cause scores to lack variance. Though the researcher believed that there was definitely one high and one low student participant, differences between two of the speakers seemed minimal. After rating each student participant on each point, the rater participants were asked to answer two questions: "What are the good points about this speaker?" and "What do you think this speaker needs to improve on?"

Finally, on the last page of the rater directions, rater participants were asked to rank the four speakers from first (most able) to fourth (least able). When all parts were completed, the rater participants handed them back to the researcher, along with the CD and a signed consent form.

Analysis

The ratings of the seven categories using the 15-point likert scales were summed in order to compute means and standard deviations for all four rater groups. Results are displayed graphically for each speaker in Figures 1-4. Rankings were calculated by percentage of ratings received at each level within each group. Finally, a FACETS analysis was performed in order to investigate rating severities between groups and bias toward speakers and categories. FACETS was chosen because of its capacity for showing rater differences and consistencies. A number of studies have used this form of analysis to look at rater bias with a smaller number of raters and a larger number of candidates (e.g., see Kondo-Brown, 2002). However, the present study uses the analysis differently from most of the other studies in the sense that it examines how a large number of raters assigned to one of four groups evaluate four somewhat similar speakers.

RESULTS

Seven Evaluative Criteria

Group means and standard deviations are reported in Tables 1-16.

Table 1

NMESL Category Statistics for Speaker 1

Scoring categories	Mean	*SD*	High	Low
Grammar	7.75	2.54	12	4
Fluency	4.50	1.80	8	1
Pronunciation	9.63	3.41	13	4
Content	9.63	2.98	14	4
Compensation tech.	7.69	3.29	15	1
Appropriateness	8.13	2.15	11	1
Intelligibility	10.00	2.94	13	3

Table 2

NMESL Category Statistics for Speaker 2

Scoring categories	Mean	*SD*	High	Low
Grammar	7.13	2.45	11	3
Fluency	6.38	2.76	11	2
Pronunciation	7.25	2.93	13	3
Content	8.88	3.46	14	2
Compensation tech.	6.94	2.61	12	3
Appropriateness	7.06	2.63	13	2
Intelligibility	8.13	2.50	14	2

Table 3

NMESL Category Statistics for Speaker 3

Scoring categories	Mean	*SD*	High	Low
Grammar	9.69	2.02	13	6
Fluency	9.75	2.77	14	3
Pronunciation	10.31	2.20	14	7
Content	11.88	2.09	15	9
Compensation tech.	10.81	2.38	14	4
Appropriateness	10.88	1.73	14	8
Intelligibility	11.00	2.32	15	6

Table 4

NMESL Category Statistics for Speaker 4

Scoring categories	Mean	*SD*	High	Low
Grammar	7.69	2.08	12	3
Fluency	6.50	2.12	10	3
Pronunciation	8.25	2.68	13	4
Content	9.38	2.78	13	4
Compensation tech.	8.31	2.59	13	2
Appropriateness	8.13	2.32	11	3
Intelligibility	9.38	2.29	12	5

Table 5

JMESL Category Statistics for Speaker 1

Scoring categories	Mean	*SD*	High	Low
Grammar	6.71	2.49	12	4
Fluency	4.57	2.01	10	1
Pronunciation	7.67	3.34	15	2
Content	10.67	3.34	15	4
Compensation tech.	6.95	3.02	13	2
Appropriateness	8.71	3.24	14	4
Intelligibility	10.57	2.57	14	4

Table 6

JMESL Category Statistics for Speaker 2

Scoring categories	Mean	*SD*	High	Low
Grammar	5.81	2.06	10	3
Fluency	5.76	2.37	12	2
Pronunciation	5.67	1.98	10	2
Content	9.19	2.82	15	5
Compensation tech.	6.71	2.43	11	2
Appropriateness	8.86	2.59	13	3
Intelligibility	9.38	2.01	13	5

Table 7

JMESL Category Statistics for Speaker 3

Scoring categories	Mean	*SD*	High	Low
Grammar	10.14	2.25	14	5
Fluency	10.62	1.68	14	8
Pronunciation	10.71	2.37	15	6
Content	12.05	2.06	15	7
Compensation tech.	11.24	2.02	14	6
Appropriateness	11.67	1.81	14	8
Intelligibility	12.19	1.79	15	8

Table 8

JMESL Category Statistics for Speaker 4

Scoring categories	Mean	*SD*	High	Low
Grammar	6.43	2.22	11	3
Fluency	6.86	2.36	13	3
Pronunciation	6.90	1.90	12	3
Content	9.29	2.80	13	3
Compensation tech	9.29	2.80	13	3
Appropriateness	8.71	2.57	13	4
Intelligibility	9.62	2.19	13	4

Table 9

NT Category Statistics for Speaker 1

Scoring categories	Mean	*SD*	High	Low
Grammar	6.63	2.91	14	3
Fluency	3.56	1.46	6	1
Pronunciation	8.25	3.01	12	3
Content	7.63	3.81	15	1
Compensation tech.	6.31	3.18	15	2
Appropriateness	6.63	3.22	15	2
Intelligibility	7.63	3.67	14	3

Table 10

NT Category Statistics for Speaker 2

Scoring categories	Mean	*SD*	High	Low
Grammar	5.19	2.81	13	2
Fluency	4.25	1.82	9	2
Pronunciation	4.63	2.74	12	2
Content	6.81	3.84	15	2
Compensation tech.	4.25	2.30	10	1
Appropriateness	6.00	2.78	13	2
Intelligibility	4.50	2.03	10	2

Table 11

NT Category Statistics for Speaker 3

Scoring categories	Mean	*SD*	High	Low
Grammar	9.50	3.04	14	2
Fluency	8.19	3.23	13	2
Pronunciation	8.81	2.98	14	3
Content	9.94	3.45	15	4
Compensation tech.	8.50	3.82	15	1
Appropriateness	9.44	2.96	15	3
Intelligibility	10.50	3.14	14	3

Table 12

NT Category Statistics for Speaker 4

Scoring categories	Mean	*SD*	High	Low
Grammar	7.44	2.91	13	4
Fluency	5.56	2.00	9	3
Pronunciation	8.50	2.69	14	5
Content	8.38	2.71	13	5
Compensation tech.	7.50	2.29	10	1
Appropriateness	8.13	2.39	13	3
Intelligibility	9.00	2.52	12	4

Table 13

PEER Category Statistics for Speaker 1

Scoring categories	Mean	*SD*	High	Low
Grammar	7.24	3.19	15	3
Fluency	5.21	2.60	12	1
Pronunciation	7.30	3.34	14	1
Content	8.94	3.42	14	2
Compensation tech.	5.85	2.66	12	1
Appropriateness	6.85	2.54	12	2
Intelligibility	8.79	2.97	13	2

Table 14

PEER Category Statistics for Speaker 2

Scoring categories	Mean	*SD*	High	Low
Grammar	7.76	3.07	13	1
Fluency	7.42	2.62	12	2
Pronunciation	6.24	3.17	12	1
Content	8.61	2.99	14	2
Compensation tech.	7.42	2.91	13	2
Appropriateness	7.94	2.67	12	2
Intelligibility	8.76	3.08	14	2

Table 15

PEER Category Statistics for Speaker 3

Scoring categories	Mean	*SD*	High	Low
Grammar	10.12	3.17	15	3
Fluency	10.45	2.66	14	3
Pronunciation	10.42	3.02	14	3
Content	10.70	2.74	15	4
Compensation tech.	9.97	3.36	15	4
Appropriateness	10.39	2.55	14	5
Intelligibility	10.97	2.59	15	6

Table 16

PEER Category Statistics for Speaker 4

Scoring categories	Mean	*SD*	High	Low
Grammar	7.64	2.93	13	2
Fluency	6.94	2.67	12	1
Pronunciation	7.18	2.79	13	2
Content	8.52	3.08	13	2
Compensation tech.	6.97	2.43	12	2
Appropriateness	7.48	2.68	13	3
Intelligibility	8.48	2.97	13	3

Grammar

Scores for grammar show that all four rater groups judged grammar with average difficulty in comparison with the other categories. The overall average low score was 5.19 (NT group, Speaker 2) with an average high score of 10.62 (JMESL group, Speaker 3). Three of the rater groups, the NMESL, the JMESL, and the NT, demonstrated the same pattern in terms of which speaker rated best to worst. In all four groups, Speaker 3 was rated most able in grammar. The NMESL, JMESL, and NT groups rated Speaker 1 as the next best, followed by Speaker 4, and then Speaker 2 as the least able. The PEER group, however placed Speaker 2 in second place, followed by Speaker 3. They rated speaker 4 as least able.

Fluency

All four groups rated this category the most severely with an overall mean high of 10.62 (JMESL group, speaker 3) and a low of 3.56 (NT group, Speaker 1). The NMESL, JMESL, and NT groups again rated in the same pattern, with Speaker 3 the most able, followed by Speaker 4, Speaker 2, and Speaker 1, respectively. The PEER group agreed on most and least able, however, placed Speaker 2 ahead of Speaker 4.

Pronunciation

Like grammar, pronunciation was rated with average difficulty. The overall mean high for this category was 10.71 (JMESL group, Speaker 3). The mean low was 4.63 (NT group, Speaker 2). The NMESL, JMESL, and PEER groups rated in the same pattern from most able to least able: Speaker 3, Speaker 1, Speaker 4, and Speaker 2. The NT Group, however, thought Speaker 4 was slightly better than Speaker 1. In fact, for the NT group there were very minor differences in scoring for Speakers 1, 3, and 4 with means of 8.25, 8.81, and 8.50, respectively.

Content

Scores for content show that, overall, judges rated this category with relative ease. The combined group mean high was 12.05 (JMESL group, Speaker 3) and the low was 6.81(NT group, Speaker 2). In this category there was some variation in rating pattern. Only the JMESL and NT group rated speakers in the same order: Speaker 3 the best, followed by Speaker 4, Speaker 1 and Speaker 2, respectively. The NMESL group showed a similar arrangement, but placed Speaker 1 slightly ahead of Speaker 4. The PEER group, like the others, chose Speaker 3 as clearly having the most ability in this category; however, they placed Speaker 1 in second place and rated speaker 2 a bit ahead of Speaker 4.

Compensation Techniques

Group means in this category revealed it to be of average difficulty, at about the same level as grammar and pronunciation. The overall mean high was 11.24 (JMESL group, Speaker 3) with a low of 4.25 (NT group, Speaker 2). The NMESL, JMESL, and NT groups exhibited the same rating pattern, placing Speaker 3 in the most able position, followed by Speaker 4, Speaker 1, and Speaker 2, respectively. The JMESL group showed little difference in mean scores for Speaker 1 and 2. The PEER group differed from the others, by rating speakers in the following order from most to least able: Speaker 3, Speaker 2, Speaker 4, and Speaker 1.

Appropriateness

Means for language appropriateness show it to be a category of easy to average difficulty. The overall mean high was 11.67 (JMESL group, Speaker 3), and the mean low was 6.00 (NT group, Speaker 2). The NMESL and JMESL rated in the same manner, with Speaker 3 as the most able, followed by a tie for second place in both cases between Speaker 1 and Speaker 4. Speaker 2 received the lowest score. The NT group demonstrated the same pattern for most and least able, but clearly found Speaker 4 better than Speaker 1. Finally, the PEER group agreed that Speaker 3 was the most able, but found little difference between Speaker 2, placed at the second position, and Speaker 4, in the third position. They gave Speaker 1 the lowest score.

Overall Intelligibility

The means for this category showed that raters judged it leniently in comparison to other categories. The overall mean high was 12.19 (JMESL group, Speaker 3) and the low was 4.50 (NT group, Speaker 2). The NMESL and JMESL groups rated in a similar pattern. As in all previous cases, Speaker 3 was found to be the best. This was followed by Speaker 1, Speaker 4, and Speaker 2. The NT group agreed that Speaker 3 was the best, but placed Speaker 4 clearly in second place, followed by Speaker 1, and in last position, Speaker 2. Like the other groups, the PEER group rated Speaker 3 the best, but produced a minute (.03) mean difference between Speaker 1 and Speaker 2. Speaker 4 was rated a bit lower, in last position.

Speaker Rankings by Group

Tables 17-20 provide rankings by group. Numbers of ratings for each speaker are calculated into percentages.

Table 17

Speaker Ranking by NMESL Group

	Speaker 1	Speaker 2	Speaker 3	Speaker 4
1st	4 (25%)	2 (12.5%)	10* (62.5%)	0
2nd	2 (12.5%)	3 (18.75%)	3 (18.75%)	8* (50%)
3rd	3 (18.75%)	6* (37.5%)	1 (6.25%)	6* (37.5%)
4th	7* (43.75%)	6 (37.5%)	1 (6.25%)	2 (12.5%)

Note: * denotes majority choice for that ranking

Table 18

Speaker Ranking by JMESL Group

	Speaker 1	Speaker 2	Speaker 3	Speaker 4
1st	0	0	21* (100%)	0
2nd	4 (19.5%)	8 (38.1%)	0	9* (42.86%)
3rd	8* (38.1%)	7 (33.33%)	0	6 (28.57%)
4th	9* (42.86%)	6 (28.57%)	0	6 (28.57%)

Note: * denotes majority choice for that ranking

Table 19

Speaker Ranking by NT Group

	Speaker 1	Speaker 2	Speaker 3	Speaker 4
1st	1 (6.25%)	0	13* (81.25%)	2 (12.5%)
2nd	6 (37.5%)	1 (6.25%)	1 (6.25%)	8* (38.1%)
3rd	5 (31.25%)	4 (25%)	1 (6.25%)	6* (37.5%)
4th	4 (25%)	11* (68.75%)	1 (6.25%)	0

Note: * denotes majority choice for that ranking

Table 20

Speaker Ranking by PEER Group

	Speaker 1	Speaker 2	Speaker 3	Speaker 4
1st	0	2 (6.06%)	28* (84.85%)	3 (9.09%)
2nd	7 (21.21%)	10 (30.3%)	2 (6.06%)	14* (42.42%)
3rd	7 (21.21%)	12* (36.36%)	3 (9.09%)	11 (33.33%)
4th	19* (57.58%)	9 (27.27%)	0	5 (15.15%)

Note: * denotes majority choice for that ranking

NMESL Group

More than half of the NMESL group (62.5%) chose Speaker 3 as the most able of the four speakers. Twenty-five percent chose Speaker 1 as the best and 12.5% believed that Speaker 2 belonged in this position. None of the raters in this group selected Speaker 4 as the best. In second place, Speaker 4 received the most ratings, with 50% of raters placing her in this position. In addition, 18.75% of the raters saw Speaker 2 in this place, and the same percentage believed Speaker 3 belonged in second. Speaker 1 was chosen by 12.5% of the raters for this spot. Raters gave both Speaker 2 and Speaker four 37.5% of their votes for third place. Speaker 1 received 18.75%, and Speaker 3, 6.25%. Finally, Speaker 1 was placed by a majority of raters (43.75%) at fourth place. 37.5% of raters selected Speaker 2 for this position, followed by 12.5% for Speaker 4 and 6.25% for Speaker 3.

JMESL Group

The JMESL group was unified in their selection of Speaker 3 as the most able: 100% of the group rated her best. In second place, Speaker 4 received the majority with 42.8%, while 28.1% of raters put Speaker 2 in the second spot, and 19.5% of raters placed Speaker 1 in this position. The ratings were split fairly evenly for third place: Speaker 1 received 38.1% of the ratings, followed by Speaker 2 with 33.3% and Speaker 4 with 28.5%. Speaker 1 received the majority of ratings for the fourth position at 42.86%. The percentages for Speakers 2 and 4 were even at 28.57%.

PEER Group

Again, Speaker 3 was chosen as the most able with 81.25% of the PEER group raters selecting her as the best. Only 12.5 % gave this position to Speaker 4, and 6.2% thought it belonged to Speaker 1. Speaker 3 received none of the ratings for most able. Speaker 1 and Speaker 4 received an almost equal percentage of ratings for second and third place: 38.1% gave second place to Speaker 4, followed closely with 37.25% for Speaker 1. The remaining ratings (approximately 12%) were split evenly between Speaker 2 and Speaker 4. A similar case occurred with the third place results: 37.5% of raters found Speaker 4 to be the third most able, while 31.25% gave this place to Speaker 1; 25% rated Speaker 2 at the third place spot and Speaker 1 received 6.25% of the ratings. Speaker 2 was a clear favorite for the least able position, with 68.755 of the ratings. Speaker 2 was also placed at this spot by 25% of the raters, and Speaker 3 by 6.25%. None of the raters placed Speaker 4 in fourth place.

NT Group

Like the previous three groups, the NT group found that Speaker 3, with 84.85% of the ratings, was the most able. Speakers 2 and 4 received a small percentage of the ratings, with 6.06% and 9.09%, respectively. None of the raters thought Speaker 1 belonged in this spot. The second and third most able positions were again split evenly; Speaker 4 received 42.42% of the ratings for second place, followed by Speaker 2 with 30.3%, and Speaker 1 with 21.21%. Only 6.06% placed Speaker 3 in this position. In the third place spot, ratings were as follows: Speaker 2, 36.36%; Speaker 4, 32.32%; Speaker 1, 21.21%; and Speaker 3, 9.09%. In last position, Speaker 1 received a clear majority of the votes with 57.58%. At the same time, 27.27% of raters thought Speaker 2 belonged in this position, and 15.15% gave this position to Speaker 4. None of the raters placed Speaker 3 in this spot.

Overall Rankings and Group Bias

The results of the FACETS analysis are shown visually in Figure 5. Looking at the figure from left to right across the point of average score (0 on the Measure logit scale), scored by an average Speaker (a speaker whose ability is 0) on an average Score Point (with a logit score of 0) by an rater (a rater whose severity estimate is 0 on the logit scale), an average speaker from an average background is likely to get raw score 8 points.

In more detail, Speaker 3 was clearly most able, though ranges in ability were relatively small, from +.49 to -.25 on the logit scale. The performances of Speakers 1, 2, and 4 were judged quite similarly. With regard to the difficulty of the seven rating categories, fluency was judged most severely, at +.23 logits, compensation techniques, grammar, and pronunciation are clustered around the average with language appropriateness slightly higher, and content of utterance and overall intelligibility rated the most leniently. In the Raters column, one finds the individual raters. Though this paper will not be looking at these differences, one can see that there was greater variation in individual ratings. Finally, one finds the rater group differences, with the NT group giving the harshest raters on average and the NMESL and JMESL groups being the most lenient. However, differences between all four groups were minor (from +.12 to -.06 logits).

```
|Measr|+Speaker|-Scoring point                                                                 |-Raters  |-Backgrounds|S.1  |

+   2 +        +                                                                               +         +            +(15) +
|     |        |                                                                               |         |            |     |
|     |        |                                                                               |         |            |     |
|     |        |                                                                               |         |            |     |
|     |        |                                                                               |         |            | --- |
|     |        |                                                                               |         |            |     |
|     |        |                                                                               | .       |            |     |
|     |        |                                                                               |         |            | 13  |
|     |        |                                                                               |         |            |     |
|     |        |                                                                               |         |            | --- |
+   1 +        +                                                                               +         +            +     +
|     |        |                                                                               | *       |            | 12  |
|     |        |                                                                               | .       |            |     |
|     |        |                                                                               | *       |            | --- |
|     |        |                                                                               | .       |            | 11  |
|     | 3      |                                                                               | **.     |            | --- |
|     |        |                                                                               | *.      |            | 10  |
|     |        |                                                                               | ***     |            | --- |
|     |        | fluency                                                                       | **      |            | 9   |
|     |        |                                                                               | ******* | NS         | --- |
*   0 *        * compensation techniques   grammar                      pronunciation         * **      * PEER       * 8   *
|     | 4      | language appropriateness                                                      | ******  | JESL  NESL | 7   |
|     | 1  2   |                                                                               | ****    |            | --- |
|     |        | content of utternace      overall intelligibility                             | *****   |            | 6   |
|     |        |                                                                               | **      |            | --- |
|     |        |                                                                               | .       |            |     |
|     |        |                                                                               | .       |            | 5   |
|     |        |                                                                               | *       |            | --- |
|     |        |                                                                               | **      |            |     |
|     |        |                                                                               | .       |            | 4   |
+  -1 +        +                                                                               +         +            +(1)  +

|Measr|+Speaker|-Scoring point                                                                 | * = 2   |-Backgrounds|S.1  |
```

Figure 5 FACETS Results

Overall severities and consistency of the four rating groups. Table 21 provides a detailed measurement report, with group severity, error, and fit statistics. The severity span between the most lenient group (NMESL) and the most severe group (NT) as shown in the second column was only .18 logits. The reliability of separation index was high at .93, and the chi-square of 46.6 with 3 *df* was significant at $p<.00$. These indicators suggest that there was significant variation in harshness among rater groups. The third column shows that the level of error was small and equal among rater groups (.02). The last column in the table identifies the infit of each category. If this number exceeds a value of 2.00, a rating group is said to be misfitting. In this case, though the infit of the NT group was a bit high, all four groups were found to be self-consistent.

Table 21

Group Measurement Report for the Four Rating Groups

	Severity (logits)	Error	Infit (mean square)
NMESL	-.06	.02	0.9
JMESL	-.05	.02	0.9
NT	.12	.02	1.3
PEER	-.01	.02	0.9

Notes: Reliability of separation index = .93; fixed (all same) chi-square = 46.6, *df* = 3, significance*:* $p<.00$

The difficulty and consistency of the seven rating categories. Table 22 shows a detailed difficulty measurement report for the seven scoring categories. As seen in the second column, fluency was scored most severely and content of utterance and overall intelligibility were rated most leniently. The difficulty span between the seven categories was small (.56 logits), however the reliability of separation index was very high (.98) and the chi-square of 398.9 with 6 *df* was significant at $p<.00$. These indicators show that significant variation in difficulty among the seven scoring categories did exist. In the third column, it is clear that the level of error was small, but differed slightly among categories (from .02 to .03). Fit values in the last column for all categories were all less that 2.00, indicating that no category was misfitting.

Table 22

Difficulty Measurement Report for the Seven Categories

	Severity (logits)	Error	Infit (mean square)
Grammar	.03	.02	0.8
Fluency	.23	.03	1.1
Pronunciation	.00	.02	1.1
Content	-.33	.03	1.2
Compensation tech.	.04	.02	1.1
Appropriateness	-.12	.02	0.8
Intelligibility	-.33	.03	1.0

Notes: Reliability of separation index = .98; fixed (all same) chi-square = 398.9, $df = 6$, significance: $p<.00$

Rater bias in terms of candidate ability. Table 23 shows the results of the bias analysis in terms of interaction between rating group and candidates. The table lists all 16 possible interactions (four rating groups x four speakers) including ones without significant bias. The first column shows the rater group. The second column shows the speaker number. The next two columns show a total observed score across the seven categories from the rater on that candidate and a total expected score across the seven categories from the rater on that candidate. Each category has a range of 1-15, and groups vary from 16 raters to 33 raters; therefore, the total observed or expected score for the seven categories combined falls within a range from 112 to 3465. Column 5 shows the difference between the average observed score and the expected raw score for the speaker by the rater across the seven rating categories. Columns 6 and 7 show the bias reflected in column 6 in logits and the likely error of this estimate. The *z*-scores for column 7 are given next. A *z*-score greater than + 2 logits suggests that the rater group scored that speaker consistently more harshly than other speakers, and a score under –2 logits indicates that the rater group scored that speaker consistently more leniently than other raters. So, looking at Table 23, one can see the following:

1. The NT group was consistently more lenient with Speaker 4 than other groups
2. The NT group was consistently harsher on Speaker 2 than other groups
3. The PEER group was consistently more lenient with Speaker 2 than other groups
4. The PEER group was consistently harsher with Speaker 1 than other groups
5. The JMESL group was consistently more lenient with Speaker 3 than other groups
6. The JMESL group was consistently more severe on Speaker 4 than other groups

The last column indicates misfit. Since none of the values exceed 3 logits, none of the interactions were identified as misfitting.

Table 23
Rater Group-Candidate Bias Interactions

Rater group	Speaker Number	Observed Score	Expected Score	Observed-Expected Average	Bias (logit)	Error	*z*-score	Infit Mean Square
NT	4	872	754.4	1.05	-0.22	0.04	-5.16	1.1
PEER	2	1780	1639.0	0.61	-0.13	0.03	-4.32	0.8
JMESL	3	1671	1618.8	0.36	-0.10	0.04	-2.19	0.7
NMESL	1	917	881.4	0.32	-0.06	0.04	-1.50	1.2
NT	1	747	717.4	0.26	-0.06	0.04	-1.33	1.5
JMESL	1	1194	1174.1	0.14	-0.03	0.04	-0.73	1.3
PEER	3	2412	2407.4	0.02	-0.01	0.03	-0.15	1.0
NMESL	4	922	922.5	0.00	0.00	0.04	0.02	0.8
NMESL	2	828	832.0	-0.04	0.01	0.04	0.17	0.8
JMESL	2	1093	1108.4	-0.11	0.02	0.04	0.57	0.8
NT	3	1038	1063.7	-0.23	0.05	0.04	1.13	1.1
PEER	4	1760	1820.4	-0.26	0.06	0.03	1.83	0.8
NMESL	3	1189	1220.0	-0.28	0.07	0.05	1.48	0.9
JMESL	4	1172	1228.7	-0.39	0.08	0.04	2.07	0.8
PEER	1	1656	1741.2	-0.37	0.08	0.03	2.59	1.0
NT	2	553	674.5	-1.08	0.28	0.05	5.50	

Rater group bias in terms of the seven categories. Table 24 shows the results of the bias analysis in terms of the interaction between rater groups and categories. It lists all 28 possible interactions (four groups x seven categories) including ones without significant bias. The first two columns show rater group and scoring category. The next two columns give the overall observed score from that rater for all speakers on that category and the expected score from the rater for all speakers on that category. The range for these scores falls between 86-1290 (four rater groups x seven categories x 1-15 possible points). Column 5 shows the average difference between the total observed and expected scores (in columns 3 and 4). The next two columns show the degree of difference indicated in the previous column in the form of a bias logit and the likely error of this estimate. The 7th column provides the *z*-score converted from column 7. A *z*-score greater than +2 indicates that the rater scored consistently more harshly in that category than others, and a score below –2 shows that the rater scored consistently more leniently in that category than other raters. Finally, the last column provides the infit mean square values. Both the mean and standard deviation of the infit mean square value was 1.0 logits.

Table 24 shows that there were nine interactions with a significant bias out of the entire 28 interactions for rater group and category. It shows the following bias patterns within groups:

1. The NMESL group was consistently more lenient in rating pronunciation than other groups.
2. The JMESL group was consistently harsher in scoring two categories: pronunciation and grammar.
3. The JMESL group was consistently more lenient in rating three categories: compensation techniques, language appropriateness and overall intelligibility.
4. The PEER group was consistently more lenient in rating two categories: fluency and grammar.
5. The NT group was consistently more lenient in scoring pronunciation.

The infit mean square values reported in Table 24 indicate that none of these interactions were misfitting; all were within the range within two standard deviations of the mean. Rater groups, therefore, were consistent in category bias patterns across candidates.

Table 24
Rater Group-Category Bias Interactions

Rater group	Category	Observed Score	Expected Score	Observed-Expected Average	Bias (logits)	Error	*z*-score	Infit Mean Square
JMESL	OI	898	837.5	0.72	-0.17	0.05	-3.30	0.9
NT	PR	483	438.1	0.70	-0.15	0.06	-2.65	1.5
JMESL	CN	890	839.1	0.61	-0.14	0.05	-2.62	1.3
JMESL	LA	811	754.9	0.67	-0.14	0.05	-2.77	0.7
PEER	FL	991	908.5	0.62	-0.14	0.04	-3.42	1.0
NMESL	PR	567	531.1	0.56	-0.12	0.06	-2.03	1.0
NT	GR	460	428.3	0.50	-0.11	0.06	-1.89	0.8
PEER	GR	1083	1028.0	0.42	-0.09	0.04	-2.24	0.8
NMESL	CT	540	517.5	0.35	-0.07	0.06	-1.27	0.8
NMESL	CN	636	631.4	0.07	-0.02	0.06	-0.27	1.4
NT	LA	479	472.2	0.11	-0.02	0.06	-0.40	0.9
JMESL	CT	693	688.9	0.05	-0.01	0.05	-0.20	0.9
NMESL	GR	516	520.2	-0.07	0.01	0.06	0.24	0.6
NT	CT	422	425.8	-0.06	0.01	0.06	0.23	1.8
PEER	PR	1028	1048.9	-0.16	0.03	0.04	0.85	0.9
PEER	OI	1217	1235.9	-0.14	0.03	0.04	0.76	0.8
PEER	CT	1000	1022.7	-0.17	0.04	0.04	0.93	0.9
NMESL	OI	616	630.2	-0.22	0.05	0.06	0.83	0.9
PEER	CN	1212	1245.1	-0.25	0.06	0.04	1.38	0.9
NMESL	LA	547	567.4	-0.32	0.07	0.06	1.16	0.6
JMESL	FL	581	610.2	-0.35	0.07	0.05	1.46	0.9
NT	CN	514	536.4	-0.35	0.07	0.06	1.29	1.5
PEER	LA	1077	1119.5	-0.32	0.07	0.04	1.73	0.9
NMESL	FL	434	458.3	-0.38	0.08	0.06	1.40	1.1
NT	OI	507	535.1	-0.44	0.09	0.06	1.61	1.4
NT	FL	345	374.0	-0.45	0.12	0.06	1.81	1.3
JMESL	PR	647	706.9	-0.71	0.15	0.05	2.95	0.8
JMESL	GR	610	692.5	-0.98	0.20	0.05	4.04	0.7

DISCUSSION

The findings presented here show us that there are differences between the four rater groups in this study, but, in most cases, these differences do not appear to be a direct result of their L1 background, or their academic training. The rating patterns as determined by group means appear to be fairly similar across groups. For all four speakers, all rater groups give Fluency the lowest score in relation to other categories. Grammar, Pronunciation, Compensation Techniques and Language Appropriateness are categories of average difficulty for all speakers as assigned by all rater groups. The last two categories, Content and Overall Intelligibility, are given relatively high scores across the four speakers by all rater groups. Speaker rankings are also very similar. Only the PEER group differed from the order suggested by the other groups.

The FACETS analysis provided a more detailed picture of biases that may exist within groups. For the most part, when it came to rater speaker preferences, harsh ratings from one group balanced themselves out with lenient ratings from another group, but no clear patterns seemed to exist based on the two factors considered in this study: L1 and academic training. Interestingly, the NMESL group exhibited no Speaker bias. Looking at bias due to rater group by scoring category, one can see that both native speaker groups were consistently lenient on pronunciation, perhaps because they felt this category was of less importance than the two Japanese rater groups. The Japanese master's degree candidates were biased in one way or another in five of the seven categories. As expected, they scored pronunciation and grammar harshly, two categories focused on in EFL settings in Japan. The three categories that they scored leniently, compensation techniques, language appropriateness, and overall intelligibility are less clearly defined and are normally not given much emphasis in defining language performance in Japan. Finally, the PEER group rated fluency and grammar leniently. It is difficult to guess at the reason for this, however, since this group consists of speakers that are of the same ability or in some cases lower than the speakers, they perhaps felt less comfortable in commenting on these areas, which they themselves might be weak in.

Limitations of this Study

In order to maximize the reliability of rater judgments in performance-based assessments, rater training is required. Unfortunately, due to the large size of the rater groups and the diverse locations, training was impossible in this study. However, since, as Brown (1995) points out, "students will, in most cases, eventually be judged by naïve native speakers of the language", such an investigation is not without value.

Also, the researcher would have preferred the raters to assess a greater number of speakers at a greater number of ability levels performing multiple speaking tasks; however, asking raters to listen to anything over the present 50 minutes of recording seemed unreasonable. Unfortunately, having only four speakers at relatively non-varying ability levels resulted in less-illuminating findings.

Finally, the rating scale, in retrospect, seems flawed. The researcher investigated a variety of rating scales and point systems, which although preferred on a theoretical basis, were impossible to employ due to the use of terminology that was unknown to those without a language training background. Again, if it had been possible to train raters, a more sophisticated scale could have been employed.

CONCLUSION

Because the purpose of this study was to investigate differences in the perceptions of the four rater groups rather than differences in reliability, some general conclusions can be made. Overall, if speaking assessments are to be included in settings which were traditionally grammar-translation oriented, raters (both classroom instructors and entrance exam raters) will have to re-examine their current foci and develop clear guidelines as to what separates an able speaker from a non-able one. They may want to consider areas such as pragmatics and compensation techniques, which appear important to native speakers but have previously been ignored in such exam-based contexts. Teachers will have to be trained in using communicative approaches to instruction and may want to use less traditional forms of assessment as regular components of their lessons. They also should consider having their students take greater responsibility for their own learning; this means instructing students in all components of what makes a competent speaker and having them perhaps evaluate themselves.

This study also points to potential problems in using performance assessments to make high stakes decisions. Three of the four speakers in this study were rated as being almost equivalent levels. Since the judges in this study demonstrated clear rating bias, to use the scores as final placement or evaluation means would seem irresponsible. Therefore, when using performance assessments to make decisions, such as those in university entrance exams, the ethical path would be to have multiple methods of testing or multiple trained raters.

Suggestions for Further Research

More research needs to be conducted into the area of oral performance assessments. Features of successful and non-successful speakers need to be clearly defined and weighted and assessment scales must be devised in order to clearly reflect these features. Raters could be interviewed and a qualitative analysis might be performed to determine categories of importance. Also, it would be useful to conduct conversation analyses of successful and unsuccessful examinees to compare whether or not evaluators reactions are subjective or are truly reflected in the discourse. Finally, it would be useful to look at the examinees' understanding of evaluations and to investigate how the degree of rater feedback given to the examinees influences future performances and beliefs about their speaking abilities.

ACKNOWLEDGEMENTS

Originally this study was designed in collaboration with Naoko Sugiyama, who unexpectedly had to withdraw form the Master's program. I would like to thank her for her help in creating the initial Japanese questionnaires.

I am also grateful for the support and assistance of my SP advisor, Dr. JD Brown. I became interested in language testing after the first course I took with him, and he has amazed me with his knowledge in this area ever since. JD always expected that I learn through my own process of trial and error, and had the confidence that I could succeed even when I lacked self-assurance. This is a more valuable lesson than perhaps any other learned in this program. Thanks to him, regardless of the results of the study, I can be proud of this work because it is truly my own.

I am also extremely appreciative of Dr. Richard Schmidt, whose lessons I have enjoyed during my time in this program, and who agreed, despite his busy schedule, to read an SP on a topic unrelated to his areas of interest. I admire his vast knowledge in this field and was grateful to have him as a reader.

Finally, I would like to give special thanks to Tomomi Hasegawa, Shin Minagawa, Akira Endo, Satomi Uchida, Hiroaki Uchida, and Motohiko Sugihara for their translation of the numerous Japanese questionnaire responses. When I believed that it would no longer be possible to continue this study, they gave me the support I needed to succeed. Without them this study could not have been completed.

REFERENCES

Bachman, L. (1988). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1990). The construct of the FSI Oral Interview. *Language Learning, 31*(1), 67-86.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*, 1-15.

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal, 10*, 15-42.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 1-27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 3,* 29-59.

Carroll, J. B. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.), *Teaching English as a second language: A book of readings.* New York: McGraw-Hill.

Kobayashi, T. (1982). Native and Nonnative Reactions to ESL Compositions. *TESOL Quarterly, 28*(1), 81-121.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias measuring Japanese second language writing performance. *Language Testing, 19(*1), 3-31.

Lado, R. (1961). *Language testing.* London: Longman: 1-37.

Linacre, J. M. (1996). *Facets, version no. 3.0* Chicago: MESA.

Lumley, T., & McNamara, T. F., (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*, 4-71.

Lynch, B. K., & McNamara, T. F., (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*, 158-80.

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17*(3), 289-310.

Savignon, S. (1983). *Communicative competence theory and classroom practice.* Reading, MA: Addison-Wesley.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model For Assessing Foreign Language Learning. *The Modern Language Journal 76(*4)*, 513-521.*

Spolsky, B. (1985). What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing, 2*(2), 180-191.

Vollmer, H. J., & Sang, F. (1983). Competing hypotheses about second language ability: A plea for caution. In J. W. Oller Jr. (Ed.), *Issues in language testing* (pp. 29-79). Boston, MA: Newbury House.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334-354.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Wigglesworth, G. (1993). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4*, 28-47.

APPENDIX A
ORAL INTERVIEW SCHEDULE

Questions

1. I know that you are studying medicine. Could you please tell me why you chose to study that?
2. What are you doing now in Hawaii?
3. How long will you be here?
4. What do you think you need to be a good doctor?
5. Do you know what kind of doctor you want to be?
   a. "Yes" answer – What kind and why?
   b. "No" answer – When do you have to decide?
6. Is there any difference between hospitals in America and hospitals in Japan?
7. Do you have any problems working with American doctors?
8. If you were not going to be a doctor, what other job would you choose? Why?

Role-play Task Description

You met an American girl at a party and she is planning to move to Tokyo for work. She needs advice on what to bring, where to live, and how to find an apartment. Give her some advice and, if possible, prepare her for some of the cultural differences. End the conversation by inviting her to your home in Tokyo.

APPENDIX B

RATER QUESTIONNAIRE (ENGLISH VERSION)

1) What is your native language? Please circle one. English Japanese

2) How old are you? ( ) years old

3) What is your gender? (Please circle one.) Female Male

4) How many years of teaching experience do you have?
None ( ) years ( ) months

5) Where did you teach or are you teaching? Please circle all that apply.

Elementary	Middle	High
College	English conversation school	TOEIC program
Eiken program	TOEFL program	GRE/SAT program
Juku/Yobiko	Business English program	Language institute for academic purposes
Tutoring	Other (	)

6) How long have you lived or did you live in an English speaking country?
Never ( ) years ( ) months

7) Hoe long have you lived or did you live in Japan?
Never ( ) years ( ) months

8) In what country did you do your undergraduate study?
Please name all countries
( )

9) In what country have you been doing or did you do your postgraduate study?
Please name all countries
( )

10) What kind of postgraduate program did you take for your English teaching?
Please circle one, or please specify the type of program

TESOL certificate Master's program Doctoral program Other ( )

11) What do you think are problems of assessing speaking ability?

12) Have you ever had your students self-assess their own speaking ability?
1 2 3 4 5
Never Very often

13) If you chose 2, 3, 4 or 5 for question 12, did you use the self-assessment as part of the students' grade?

14) If you chose 1 for question 12, would you use the self-assessment as a part of the students' grade?

15) Do you think that self-assessment is an effective way of assessing students' speaking ability?
1 2 3 4 5
Not effective Very effective

16) Have you ever had your students assess their classmates' speaking ability?

1 2 3 4 5

Never Very often

17) If you chose 2, 3, 4 or 5 for question 16, did you use the peer assessment as part of the students' grade?

18) If you chose 1 for question 16, would you use the peer assessment as part of the students' grade?

19) Do you think that peer assessment is an effective way of assessing students' speaking ability?

1 2 3 4 5

Not effective Very effective

20) Please write about your teaching and assessing experience in speaking.

21) Have you ever taken a speaking test as a language learner?

APPENDIX C

RATER DIRECTIONS

CD Directions

On the CD, interviews with four speakers are recorded. After listening to each track, please **PAUSE** your CD player (please **DO NOT STOP**), evaluate the speaker according to the 15-point scale and answer the two questions. You may go back to the previous page(s) for rating, but please do not listen to the same track again. When you are finished you will be asked to rank the four speakers from most able to least able. Please do not discuss your evaluations with others.

Content of the CD. The CD contains four tracks:

Track 1	Speaker 1
Track 2	Speaker 2
Track 3	Speaker 3
Track 4	Speaker 4

If you stop the CD by accident, please take a look at the track number written at the top of each page and make sure to return it to the corresponding position on the CD.

Rating Sheet

Speaker 1

Please listen to the 1st track and evaluate the speaker for each category. Circle only ONE number.

Grammar
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Fluency (naturalness of speed, pauses, and flow)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Pronunciation
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Content of utterance (the degree to which the answer satisfies the question)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Compensation techniques (techniques, such as rephrasing, asking for clarification or repetition, fillers, and self-repair)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Language appropriateness (word choice, tone of voice and pragmatics)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Overall Intelligibility (the degree to which the speaker is understood)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Very poor Excellent

Questions

What do you think are good points of this speaker?

What do you think the speaker needs to improve on?

Ranking Sheet

Please rank the four speakers from most able to least able.

Ranking	**Speaker number**
First (most able)	____________
Second	____________
Third	____________
Fourth (least able)	____________

Heather Caban
Department of Second Language Studies
1890 East-West Road
Honolulu, HI 96822

Yearin@@hotmail.com