EVALUATING AN INSTRUMENT FOR ASSESSING CONNECTED SPEECH PERFORMANCE USING FACETS ANALYSIS

YOON AH SEONG
University of Hawai'i at Mānoa

ABSTRACT

In the area of English pronunciation teaching, *connected speech* is increasingly being introduced and covered in pronunciation textbooks (e.g., Hagen, 2000; Weinstein, 2001). Connected speech is a phenomenon in spoken language that collectively includes phonological processes such as reduction, elision, intrusion, assimilation, and contraction. Several research studies have shown that connected speech instruction can help learners to more easily comprehend rapid speech used by native speakers (e.g., Brown & Hilferty, 2006; Celce-Murcia, Brinton, & Goodwin, 1996; Matsuzawa, 2006). Moreover, use of connected speech features can make learners sound more comprehensible and natural with less marked foreign accent (Brown & Kondo-Brown, 2006a; Dauer & Browne, 1992). However, compared to the growing connected speech literature regarding what forms to teach and how, there seems to be very little information on how to assess connected speech especially in terms of production.

Therefore, the purpose of this study was to develop and evaluate a new test of connected speech performance within the context of an English study abroad program. The multi-faceted Rasch software FACETS was used to examine the effectiveness of the test instrument. The analyses used data from two administrations, a pretest and a posttest, and examined the relationships between examinee scores and various aspects of the testing situation (i.e., facets). The four facets investigated in this study were: (a) the examinees, (b) items, (c) raters, and (d) the rater L1 background. The results indicated that assessing the production of certain connected speech forms using this type of test instrument has potential. Detailed inspection of several items, as well as unpredictable examinees' performances, and inconsistent ratings from the raters lead to suggestions for revision and improvement in the item selection (elimination of a single item), rating scales (inclusion of concrete descriptors), and assessment procedures (detailed rater guidelines and training).

The emphasis on communicative competence in English language teaching has placed considerable weight on the speaking and listening abilities of learners. According to Celce-Murcia, Brinton, and Goodwin (1996), this tendency has also brought pronunciation into the spotlight as a crucial factor in oral proficiency. Pronunciation ability is important since it is not only needed for intelligible communication but can also influence individuals psychologically in that accent is a "central component of face-to-face interactions and is consequently part of the

Second Language Studies, 26(2), Spring 2008, pp. 45-101.

process by which speakers present an image of themselves to others" (Pennington & Richards, 1986, p. 215). What is more, pronunciation can establish an individual's identity or sense of affiliation to a certain group (Dalton & Seidlhofer, 1994; Gatbonton, Trofimovich, & Magid, 2005) and even bring motivation and confidence into language learning (Bamgbose, 1998; Pennington, 1994).

In the practice of English pronunciation teaching, the ability to reproduce suprasegmental features such as intonation, rhythm, and sentence stress has long been recognized to be important for achieving overall intelligibility (Anderson-Hsieh, 1990; Dauer & Browne, 1992; Hahn, 2004; Pennington & Richards, 1986). Correspondingly, connected speech has been introduced and covered in many pronunciation textbooks (Brown, in progress; Celce-Murcia et al., 1996; Gilbert, 2005; Hagen, 2000; Weinstein, 2001). Connected speech is a phenomenon in spoken language that collectively includes phonological processes such as reduction, elision, intrusion, assimilation, and contraction, and a number of research studies have shown that learning connected speech can help learners comprehend authentic natural speech used by native speakers (Brown & Hilferty, 1986a, 1986b, 2006; Henrichsen, 1984; Ito, 2006a; Matsuzawa, 2006). Moreover, knowing how to produce connected speech could also help make the learners' speech more comprehensible and natural (Brown, in progress; Brown & Kondo-Brown, 2006a; Dauer & Browne, 1992).

However, just as pronunciation testing is not receiving sufficient attention in research or practice (Koren, 1995; Yoshida, 2004), there seems to be very little available information on how to assess connected speech, compared to the growing connected speech literature. In particular, little has been published regarding how to assess production of connected speech, although a few books (Brown & Kondo-Brown, 2006a; Celce-Murcia et al., 1996) suggest ideas for developing connected speech tests for both perception and production.

Therefore, this study will examine a connected speech performance test developed by the researcher and used in an English study abroad program for Korean children studying in Honolulu, Hawai‘i. The test was developed and utilized for the purpose of assessing the students' production of certain connected speech features learned in class and providing them feedback. The primary focus of this study is to evaluate the effectiveness of this testing instrument using FACETS analysis with the facets of interest being examinees, items, raters, and the rater L1 background. Based on the analysis, suggestions for revising and improving the current test design will be discussed, along with the potential usability of this type of test instrument for assessing connected speech performance.

In order to clarify why a connected speech performance test was developed for an English study abroad program for Korean children, four issues will be discussed in this paper: (a) the role of connected speech in pronunciation teaching, (b) the teaching of pronunciation in Korea, (c) the assessment of connected speech, and (d) the development of the Connected Speech Test (CST).

ROLE OF CONNECTED SPEECH IN PRONUNCIATION TEACHING

Why Connected Speech?

Brown and Kondo-Brown (2006a) define connected speech as an "analysis of the continuous chains in normal spoken language and conversation as compared with the typical linguistic analysis of individual phonemes analyzed in isolation" (p. 284). In other words, connected speech involves the phenomena in spoken language that collectively include phonological processes such as reduction, elision, intrusion, assimilation, contraction and so forth. Brown and Kondo-Brown (2006a) mention that connected speech makes up "a very real part" (p. 5) of the spoken language and occurs in "all levels of speech" (p. 5) from casual to even very formal levels. The naturally occurring speech of native speakers is mostly rapid and continuous with frequent linking, sound alteration, or reduction at word boundaries, which may cause comprehension difficulty when non-native speakers listen to it.

Ito (2006b) describes how non-native speakers would find connected speech very different from what they would have normally heard before in language classrooms, where the speech from teachers and audio materials are typically carefully or slowly articulated. Thus understanding how connected speech functions in English could assist the learners in listening to English more easily.

But what about production? Dauer and Browne (1992) argue that producing connected speech can be beneficial in many ways because it enables the speaker to not only improve his or her intelligibility by developing overall speech rhythm, but also brings psychological relief and confidence as it causes speech to sound more natural. Not using connected speech might even cause a non-native speaker's speech to sound unnatural and choppy, and could bring about frustration to the listener (Brown, 2001; Celce-Murcia et al., 1996).

Connected Speech and Perception

The influence of connected speech on listening has been investigated in several studies

(Brown & Hilferty, 1986a, 1986b, 2006; Henrichsen, 1984; Ito, 2006a). These studies also show how reduced forms in connected speech can interfere with listening comprehension. Henrichsen (1984) hypothesized that reduced forms in listening input would decrease the saliency of the words and therefore make comprehension more difficult for ESL learners. This hypothesis was supported by results showing that both high and low level ESL learners scored significantly lower on a test where the examinees had to write down the citation form of the words in a sentence being said in reduced forms. Comprehending the input with reduced forms, compared to when the sentences were fully enunciated, was more difficult for both levels of students meaning that connected speech was not easy to understand regardless of the level the students were in.

Ito (2006a) further examined this issue using a dictation test by examining the comprehension difficulty difference caused by two types of reduced form, the lexical and the phonological forms. Her assumption was that *lexical reduced forms* such as in the example *won't* exhibit more saliency and thus would be more comprehensible compared to *phonological forms* such as in *he's* where there is no drastic phonological change after the two words, *he* and *is,* form a contraction. The results were similar to Henrichsen (1984) and showed that reduced forms do interfere with listening comprehension. Just as she predicted, non-native speakers scored significantly lower on the dictation test regarding the phonological forms than the lexical forms indicating that different types of reduced forms did distinctively affect comprehension.

Based on the findings that reduced forms in connected speech cause difficulties in listening comprehension, several studies attempted to investigate the teachability and effectiveness of explicit instruction in connected speech on listening. Brown and Hilferty (1986a, 1986b, & 2006) examined the effectiveness of teaching reduced forms to 32 Chinese EFL graduate students. After 30 ten-minute mini-lessons on reduced forms, the group of 16 students who received the instruction as opposed to the other 16 students who did general pronunciation drills was found to have scored higher on two of the three measures used (Integrative Grammar Test from Bowen, 1976, and a reduced form dictation test) suggesting that teaching connected speech does facilitate listening comprehension.

Matsuzawa (2006) did a similar experiment using a pretest-posttest design with 20 Japanese business people to see if they would benefit from connected speech instruction. The student's listening comprehension ability was measured by a dictation test similar to the ones mentioned above, and subsequent to the treatment, the posttest scores indicated that the students had made statistically significant improvement.

Connected Speech and Production

In comparison to the studies on connected speech and perception, few studies have investigated the production of connected speech to see whether it could make speech more intelligible and natural, or whether it could be taught and improved through instruction. Although Anderson-Hsieh, Riney, and Koehler's (1994) research did examine the production of connected speech forms, the study was more about examining how native speakers and non-native speakers differ in the amount of connected speech produced, rather than investigating the effectiveness of instruction on overall pronunciation ability. The results revealed that the higher level students produced more connected speech modifications that were closer to those of English native speakers than the lower level students, which showed that the ability to produce connected speech forms was related to proficiency level. More studies investigating the use of connected speech and its influence on intelligibility and overall pronunciation are needed in order to understand the benefits of learning how to produce connected speech, not to mention more studies examining the teachability of connected speech forms.

PRONUNCIATION TEACHING IN KOREA

Pronunciation is gaining more attention in English classrooms in Korea, as communicative competence becomes a primary goal of English education. In the 7th National Education Curriculum for elementary school English from the Ministry of Education (1998), specific pronunciation learning goals for Grade 3 English stated that students will be able to distinguish different sounds, stress, rhythm, and intonation in English and also speak with appropriate stress, rhythm, and intonation.

Connected speech is increasingly regarded as an important matter in English classes as well (Lee & Jung, 2003; Yoo, 2005). In the case of Lee and Jung's study (2003), they examined five types of textbooks used in junior high schools to investigate the types of connected speech forms covered and how often they were introduced. Further, they conducted an intervention study to examine how explicit instruction on certain connected speech features could enhance listening ability. Yoo (2005) used a dictation test to investigate the types of connected speech features that were causing the most difficulty in listening among Korean high school students. Similarly, the connected speech studies in Korea mainly focus on how teaching connected speech is important and effective for improving 'listening' comprehension skills. The underlying cause for this could

be the compulsory listening component included in school English exams, and especially the Korean College Entrance Exam, wherein speaking ability is currently not measured.

Additionally, the number of children being sent to English kindergartens and private schools is increasing because of the general belief that age is the major factor in phonological acquisition as supported by many studies (e.g., Flege, Yeni-Komshian & Liu, 1999; Munro, Flege, Mackay, 1996; Oyama, 1976; Tsukada, Birdsong, Bialystok, Mack, Sung, & Flege, 2006). This belief is leading many parents and teachers to reckon that starting English earlier would enable attainment of native-like proficiency especially in terms of pronunciation. Moreover, English as an official subject in school starting from grade three or younger has intensified this tendency, which has now become not only a trend but also a significant concern. Parents are spending great amounts of money to send their children to private English schools where they can interact with native speaker teachers. Besides private schools and tutoring, various English camps and study abroad programs are also cropping up one after another so that students can leave for English speaking countries or spend more time in English immersion environments to maximize their opportunity of obtaining native-like English speaking skills (Joe, 2005; Yeo & Park, 2006).

ASSESSING CONNECTED SPEECH

Pronunciation Assessment

Pronunciation is often included as one component of more holistic oral proficiency assessments such as the ACTFL OPI, the SPEAK test, the IELTS examinations, and the like. However, little literature exists on tests for assessing pronunciation *per se* (Celce-Murcia et al., 1996; Koren, 1995; Yoshida, 2004), not to mention connected speech in particular.

For pronunciation tests that are used to assess the ability to distinguish different sounds and patterns, the most commonly used forms are multiple-choice tests where you select a word with a different or same phoneme sound among a set of minimal pairs and dictation or cloze tests where you fill in the blanks or write down what is being said. For testing pronunciation performance, some common forms that are less authentic and more controlled would be listen-and-repeat tests, reading a word or sentence list, or reading a paragraph such as the well-known ‘accent inventory’ of Prator and Robinett (1985). Tests that have students read dialogues are considered to be slightly closer to authentic spoken speech rather than reading sentences and paragraphs, but there are also minimally-controlled styles of pronunciation tests where the examinee produces spontaneous speech by talking freely about a certain topic, describing pictures or stories, and

even performing interviews (Dauer, 1993; Miller, 2006). Another type of test that could be used for assessing not only pronunciation but also general listening and speaking abilities would be to use communicative tasks where the student must role-play or complete a task by listening to certain prompts or instructions carefully designed in order to elicit specific types of pronunciation features (Celce-Murcia et al., 1996).

However, many of these suggested test methods and formats are not very different from ordinary pronunciation activities practiced in classrooms, and they are mainly used for diagnosis and/or feedback purposes. Not many tests provide specific tools such as a scale that could give students some kind of objective rating or a score. However, Kim and Margolis (1999) and Yoshida (2004) developed analytic scales that could not only be used in EFL classrooms in Korea and Japan, respectively, as a tool for assessing pronunciation, but also as a reference point that teachers could use to give feedback to their students.

The English Pronunciation Test (Kim & Margolis, 1999) contains two tasks, a read-aloud passage task for rating overall naturalness and a 30-sentence read-aloud task that is used for rating the other eight categories: first language interference, consonant articulation, vowel articulation, word endings, past plural morphology articulation, word stress, intonation, and rhythm. Each category has multiple numbers of indicators (e.g., *Question* and *Exclamation* indicators for the intonation category) that are each rated on a five-point scale from ‘very poor’ to ‘very good’. The nine categories, each having multiple sub-indicators, resulted in a detailed analytic rubric which could be very useful for teachers to rate specific features more objectively and also give explicit feedback to students regarding each category.

Yoshida’s instrument (2004) is similar to that of Kim and Margolis in terms of having the students perform two different tasks and be rated on an analytic scale. However, her instrument consists of two different types of texts (a prose and a dialogue) that were taken and adapted from Dauer’s (1993) pronunciation textbook, *Accurate English*. Her test has three categories in the rating scale according to three different aspects of pronunciation: segmentals, suprasegmentals, and paralinguistic features. These were again each made up of five indicators (e.g., *Loudness, Rate, Smoothness, Energy, Clarity* indicators for the Paralinguistic features category). The item specifications with the rating guidelines were deliberately structured and prepared for the raters, and the raters had to go through practice sessions to make sure they were familiar with the instrument.

Connected Speech Assessment

Although very little literature deals with the issue of assessing connected speech specifically,

Brown and Kondo-Brown's (2006b) chapter and Celce-Murcia et al. (1996) book on pronunciation teaching provide excellent lists of ideas for testing connected speech in terms of both perception and production. For testing listening ability, the most dominant forms that were proposed and used in research were different variations of dictation or cloze style tests that require students to listen to sentences articulated with connected speech forms and fill in the blanks with their citation forms (Bowen, 1976; Brown & Hilferty, 1986a, 1986b, 2006; Henrichsen, 1984; Ito, 2006a; Matsuzawa, 2006). Another form useful for assessing listening comprehension is a test where the examinees would have to answer comprehension questions after listening to a passage or dialogue filled with connected speech features (Brown & Kondo-Brown, 2006b).

For assessing production, Anderson-Hsieh et al. (1994) used a test including sentence reading tasks and a spontaneous speech task where the students were asked to talk about their most exciting or dangerous experience. However, this test was not intentionally designed to test specific features of connected speech. It contained 15 sentences that were selected from Prater and Robinett's (1985) book that had many consonant clusters at word boundaries, so that connected speech features such as linking and consonant cluster simplification could be induced when they were read out loud. Since the research focused on examining the amount of connected speech produced by people in different groups and not on scoring the performance of these connected speech features, the number of connected speech forms produced by examinees was counted instead of rating the performance on a scale. Other interesting suggestions have been made by Brown and Kondo-Brown (2006b) regarding alternative types of reduced form assessments which could be especially useful for raising self-awareness of the prevalence and importance of connected speech. Some examples include keeping portfolios of recordings and observations of naturally occurring English and peer or self assessment where the students would have to rate their own or each others' performances.

Although ideas and methods for assessing connected speech production do exist as in the case of assessing pronunciation in general, most of them are more similar to class activity ideas and do not include materials such as a passage or a dialogue that is specifically designed for probing features of connected speech nor do they include an analytic scale.

THE DEVELOPMENT OF THE CONNECTED SPEECH TEST

The Program

The Hawai‘i Study Abroad Program is a one-year study abroad program where 10 to 20

Korean elementary and secondary students from grade 1 to grade 9 come to Hawai‘i to learn English. They attend an ordinary school during the daytime and have additional ESL classes in the evening. One of the evening classes is a listening and speaking class focused mainly on pronunciation. Although the class covers everything from phonemes to suprasegmentals, commonly occurring connected speech features such as linking, palatalization, and reduced forms were the predominant topics covered. Indeed, the children enjoyed it, because not only did the parents want their children to attain native-like accent through these classes but also the children themselves desired it. This could be more or less explained by a motivation to sound like their other peer members at school, since they did not want to feel left out and be stigmatized as the ‘ESLers’ because of their foreign accent. Pennington (1994) with Dalton and Seidlhofer (1994) explain this as having integrative motivation that makes ESL learners living in English speaking countries desire native-like accent as one of the means of blending into the target language community.

The Connected Speech Test

In order to measure the students’ achievement in learning, to provide them with feedback, and to guide instruction, a testing instrument that could assess the children’s ability to produce the connected speech forms covered in class was needed. Inspired by the analytical pronunciation instruments of previous researchers, the CST was developed based on a read-aloud dialogue and rating scale.

The test was based on a written dialogue because, according to Ur (1996), a dialogue can be useful for beginning level students who cannot yet talk freely in their L2; it would allow them to produce the target language and may even encourage them to learn formulaic expressions that could contribute to their learning processes. Brown and Kondo-Brown (2006b) also suggest that creating a dialogue containing target forms the teacher has taught could be a useful way to test connected speech production as long as the issue of unnaturalness is carefully avoided by making the dialogue close to oral language. Compared to reading a passage, reading a dialogue serves as a controlled but effective method because such performances more closely resemble authentic speech, and in fact, “mimics the spoken language” (p. 252).

The dialogue is a conversation of two children (See Appendix A), Kris and Jay, who get into a minor argument when Kris feels that he or she is being interrupted in the middle of a hide-and-seek game. Names that could refer to both boys and girls were intentionally selected so that the test takers could take either role interchangeably and not feel any discomfort from gender-

specific names. The setting of a hide-and-seek game and a storyline that involves many emotional statements were purposely selected so that the dialogue could be more interesting, fun, and relevant to children. Three types of target connected speech features, (a) the weak form of *you* using the vowel schwa [ə], (b) palatalization, and (c) commonly used reduced forms such as *wanna* and *gonna*, were used to create the dialogue since they were selected initially to be taught in class. Although it was not an easy task to keep the dialogue as authentic as possible while considering the incorporation of all the target features, care was taken to include multiple indicators for each feature in the dialogue.

When performing the dialogue, the students were asked to read the dialogue aloud in pairs and then switch roles so that each student could perform both roles. Their performances were recorded and rated on a five point scale from 'very poor' to 'very good' for four categories per role. The four categories included the three features mentioned above and an additional fourth category, *naturalness*, so that the overall impression of the examinee's pronunciation could also be rated. Specific descriptions of each category, the items involved in the rating scale, and the item specification list presenting the indicators for each item will be elaborated in the method section.

The Intended Test Use

Since all assessments are designed and utilized in distinctly different situations and settings, and for different purposes, the first step to take prior to evaluating the effectiveness of an assessment procedure is to establish a clear specification of the *intended test use* (Norris, 2008). This specification could be accomplished by determining *who* uses the test information, *what* information is provided by the test, *why* the test exists (for what purposes), and finally the *impact* that will result from the consequences of the test (Norris, 2000, 2008). Accordingly, a description of the intended uses of the test is given in Table 1. Note that the test was created for low-stakes classroom use to primarily measure the student's achievement in acquiring what has been taught in class so that further feedback and supplemental instruction could be provided in whatever aspect needed.

Table 1

The Intended Test Use of the CST

Components	Specifications
Who	Teacher and Program administrator Students Students' parents
What	Ability to produce the connected speech forms learned in class (Evidence of mastery or achievement)
Why	To provide the teacher with the information about - How much the students are able to perform what they have learned - Whether the instruction and/or materials were effective - Which target form(s) need additional focus on instruction - Each student's mastery regarding the performance of target forms so that individual feedback and further guidance could be provided if needed. To provide the students with - A chance to perform what they have learned - Recordings of their own performances so that self awareness could be raised - Feedback so that strengths and further areas of improvement could be identified. To provide the parents with the information of how much their children have achieved in class
Impact	Change in lesson planning and pedagogical method Positive and/or negative wash-back effect on the materials and instruction in class Evaluation purpose (Teacher, instruction, material, student, program, etc.)

RESEARCH QUESTIONS

This study follows the approach used by Yoshida (2004) and examines how facets of the CST such as items, raters, and rater L1 background contribute to the scores of the examinees. This new form of connected speech performance test will be explored in depth for the purpose of making any necessary revisions for improvement, and to evaluate its usefulness as a measure of connected speech. Therefore the research questions for this study will be as follows:

1. Does the test produce reliable test scores?
2. Is the instructional intervention effective?
3. Facet Effects
   A. Examinees:
      How do the 13 examinees differ in terms of ability and how well do their performances 'fit' the model?
   B. Items:
      How do the eight items differ in terms of difficulty and how well do they 'fit' the model; in other words, do they measure a single construct?
   C. Raters:
      a. How do the 44 raters differ in terms of severity in rating and how well do they 'fit' the model; in other words, are their ratings consistent?
      b. How is the rating scale being used by the raters?
   D. Rater L1 Background:
      a. How do English teachers with Korean L1 background (KET group) and English teachers with English L1 background (EET group) differ in terms of rating severity and consistency?
      b. Does the KET group or EET group display bias towards any certain examinee(s) or item(s)?

METHOD

Participants

Examinees. The total of 13 examinees who took the CST were elementary and junior high school students from Korea. The examinees included six girls and seven boys with ages ranging from eight to 13 and a mean age of 10. All 13 participants were studying in the Hawai'i study

abroad program and were in their third month of the program when taking the test. Due to the small number of students in the program, they were all placed into the same after-school ESL class despite large differences in age and English proficiency levels. The class, which was a listening and speaking class focused especially on pronunciation, met during the weekdays from Monday to Friday, 90 minutes each day.

Raters. Since one of the goals of this study was to see how the rating scale is used by teachers, 46 English teachers were initially recruited to be raters, but only 44 were used in this study because two teachers did not return their ratings. Among the 44 raters who did return their ratings, 22 were English teachers that spoke Korean as their L1 (KET) and the remaining 22 were English teachers who spoke English as their LI (EET). Raters from two different L1 backgrounds were recruited so that the two groups could be compared in terms of their rating patterns.

In order to better understand the raters' backgrounds in teaching English and their experience with evaluating English phonology and pronunciation, all of the raters were required to fill out a simple survey form before they turned in the ratings (See Appendix B). All teachers from both the KET and EET groups had experience in teaching English in a variety of contexts; the length of teaching experience in the two groups combined ranged from six months to 16 years. The KET group's teaching experience ranged from 0.5 years to 16 years with a mean ength of 4.43 years, and most of the teachers in the KET group had experience teaching junior high or high school students. Only a few had taught K-6 level children in Korea, and five of them also had experience teaching college level students in an ESL setting in the United States. The EET group's teaching experience ranged from 1.5 years to 13 years with a mean of 5.9 years, and their teaching backgrounds were found to be much more diverse than the KET group teachers. In comparison to the KET group, many of these teachers had teaching experience in both ESL and EFL contexts which included a variety of countries (i.e., China, Hong Kong, Japan, Korea, Singapore, Thailand, and the USA). There were a few more teachers in this group who had experience teaching young children (K-6) compared with the KET group. Only two teachers were familiar with Korean L2 speakers of English from past experience teaching in Korea. Teachers from both groups had background in phonology from courses, teaching experience, or from teacher training; a total of 20 from the KET group and 19 from the EET group were familiar with phonology. However, few teachers had experience in rating pronunciation; only four from the KET group and 10 from the EET group had pronunciation rating experience mostly from evaluating students in the classes they taught or from course projects (see Table 2 for

summary).

Table 2
Characteristics of the Raters

	n	Teaching Length			Phonology Background		Pronunciation Rating	
		M	*SD*	Low-High	Yes	No	Yes	No
KET	22	4.43	4.01	0.5 - 16	20	2	4	18
EET	22	5.9	2.99	1.5-13	19	3	10	12

Materials

A script was developed including the target forms that were taught in class to assess the students' performances on producing the connected speech features they have learned. This section will discuss the following elements of the testing instrument: the three different categories of target connected speech features and the items for the rating scale.

Three categories of target connected speech features. Pronouns and auxiliary verbs are often reduced (Gilbert, 2005) and difficult to hear in spoken English because they are unstressed and said quickly in weak forms using the vowel *schwa* [ə]. Cahill (2006) suggests that many possible combinations of *yes/no* question-phrases consisting of an auxiliary verb with a pronoun (such as in *did you*) are common types of connected speech that are useful to learn. Another phenomenon that frequently occurs in connected speech is palatalization, which is "one sort of a reciprocal assimilation that occurs in NAE connected speech when the [t], [d], [s], or [z] phonemes are followed by a [j] phoneme and combine to become [ʧ], [ʤ], [ʃ], or [ʒ] respectively" (Brown & Kondo-Brown, 2006a, p. 286). For instance, in phrases such as *did you* or *didn't you*, palatalization often occurs because of the [t] and [d] meeting with the glide [j].

Accordingly, reduced forms of *yes/no* question-phrases containing an auxiliary verb and the pronoun *you* were chosen for instruction so that the two connected speech features, weak form of the pronoun *you* and palatalization, could both be addressed. To create the dialogue, most of the *yes/no* question-phrases were selected from Cahill's list (2006), and additional phrases including auxiliary verbs *could*, *would*, and *should* were included to increase the number of occurrences of palatalization (See Table 3).

Table 3

Yes/No Phrases Including Features of the Weak Form of 'You' (WF) and Palatalization (PT)

Full form	Reduced form	Full form	Reduced form
can you	[kænjə]	can't you	[kæntʃə]
will you	[wɪljə]	won't you	[wontʃə]
do you	[dujə]	don't you	[dontʃə]
are you	[arjə]	aren't you	[arntʃə]
were you	[wərjə]	weren't you	[wərntʃə]
have you	[hævjə]	haven't you	[hævntʃə]
did you	[dɪʤə]	didn't you	[dɪdntʃə]
could you	[kʊʤə]	couldn't you	[kʊdntʃə]
would you	[wʊʤə]	wouldn't you	[wʊdntʃə]
should you	[ʃʊʤə]	shouldn't you	[ʃʊdntʃə]

The third type of connected speech taught was reduction of high frequency phrases. Phrases such as *want to* or *going to* are high frequency phrases that often undergo modification and become reduced phrases such as *wanna* and *gonna* (Avery, Ehrlich, Mendelson-Burns, & Jull, 1987). The target high frequency phrases were selected from Matsuzawa's list (2006) with a few more added (See Table 4). In sum, three categories of connected speech features were chosen to be included in the test dialogue, the weak form of *you*, [jə] (WF), palatalization (PT), and reduced form of high frequency phrases (RF).

Table 4

Reduced Forms (RF) of High Frequency Phrases

Full Form	Reduced Form
going to	gonna
want to	wanna
don't know	dunno
got to	gotta
give me	gimme
let me	lemme
leave me	leamme
have to	hafta
has to	hasta

Items and Rating Scale

Although the three features discussed above were the major categories of concern for rating, an additional *naturalness* (NT) category was added to the rating scale for the purpose of making a holistic judgment of the examinee's overall naturalness in pronunciation (See Appendix D for the example evaluation form raters used). Therefore the examinee would not only be evaluated analytically on their ability to pronounce the weak forms of you, palatalization, and reduced forms of high frequency phrases, but they would also be receiving a rating of their pronunciation as a whole. In addition, each examinee was given two ratings for each category, one for Jay's role and one for Kris's, since they were required to take turns with their partners and read both roles. Therefore, the total number of items a single examinee could score was the sum of all eight ratings: NT, WF, PT, and RF for Kris's role and NT, WF, PT, and RF for Jay's role. The specific indicators in the dialogue for each item are summarized in Table 5. This table was also presented to the raters in their rating guidelines so that it could assist them in paying attention to certain phrases they would be basing their judgments on. Note that the indicators for the WF and PT categories for Kris's role (in bold-faced letters) overlap because those phrases involve both features.

Table 5

Items Specifications

Categories	The indicators		Rating criteria
	Jay	Kris	
Naturalness			How natural is their speech overall?
Weak form of you	Are you – Are /jə/ Were you – Were /jə/ Have you – Have /jə/ Do you – Do /jə/	**Shouldn't you – shouldn/ʧə/** **Did you – Di/ʤə/** **Could you – Cou/ʤə/** **Can't you – Can/ʧə/** **Would you – Wou/ʤə/** **Aren't you – Arn/ʧə/**	How well is the pronoun "you" pronounced in the weak form with the *vowel schwa*?
Palatalization	Won't you – Won/ʧə/	**Shouldn't you – shouldn/ʧə/** **Did you – Di/ʤə/** **Could you – Cou/ʤə/** **Can't you – Can/ʧə/** **Would you – Wou/ʤə/** **Aren't you – Arn/ʧə/**	How well do they use palatalization?
Reduced Forms	Give me – Gimme Want to – Wanna Let me – Lemme Going to – Gonna	Got to - Gotta Don't know - Dunno Leave me – Leamme	How well do they produce reduced forms?

Procedures

In order to examine the students' existing ability to produce the connected speech features that were going to be taught, the testing instrument was used for a pretest before instruction took place. Following the pretest, the students received two weeks of instruction in the target features and took the same test once again at the end.

Pretest. The dialogue was distributed to the students before the test, and they were given five minutes of practice time so they could become familiar with the content and therefore produce fewer reading mistakes (Celce-Murcia et al., 1996). The children were well acquainted with the storyline of playing hide-and-seek, since it was a topic they had learned about in a previous lesson. Accordingly, the dialogue was comprehensible for the children, and the possibility of the content's meaning interfering with their performances was eliminated. The children were paired and assigned randomly for taking the test. However, since the number of students was not even (seven boy and six girls), one of the girls was asked to read the dialogue two times, but only her first reading was considered her official performance for the test. Each pair was called to a separate room where they read the script, and their performances were recorded on a digital voice recorder. After reading the dialog the first time, they switched parts and read the dialog again so each examinee could read both Jay and Kris's part. Each pair was thus recorded twice resulting in 14 recordings of seven pairs.

Instructional intervention. Beginning the day following the pretest, eight lessons on the target connected speech features took place. On the first day, to introduce the issue on connected speech, the children were asked if they had any trouble understanding what their native speaker friends say at school either because it was too fast or sounded different from what they knew. When presented with examples such as '*cause* for *because*, *wanna* for *want to*, or '*ssup* for *what's up*, they strongly acknowledged the fact that they would often encounter those forms being said and that they were not easy to understand. The remaining days were spent on teaching the features, and every lesson started with a review of the previous lesson. Various listening activities using dictation and cloze were used along with reading-aloud exercises; these activities were mostly created, selected, or adapted from pronunciation books of Rost and Stratton (1978), Weinstein (2001), and from Cahill's (2006) study.

Posttest. The posttest was administered using the same process as the pretest. The pairings were the same as in the pretest in order to avoid partner effect on the rater's judgment. In other words, the partners were left unchanged so that the ratings would not be influenced by having a different partner and thus the scores of the examinees in the pretest and posttest could be comparable. Likewise, five minutes were given for practice before the test, and the performances were all recorded on a digital recorder. Table 6 summarizes the overall testing and instruction from the pretest to the posttest.

Table 6

Procedure Summary

Day	Topic	Content
1	Pretest	dialogue read-aloud test
2	Introduction to reduced forms	discussion on spoken English
3	Weak form of you 1	can/will/do/are/were/have + you
	Weak form of you 2	can/will/do/are/were/have + you
4	Palatalization 1	did/could/would/should/ + you
5	Palatalization 2	all negative forms (e.g., won't/didn't)
6	High frequency phrases 1	going to & want to
7	High frequency phrases 2	got to, have to, & has to
8	High frequency phrases 3	give me, let me, leave me, & don't know
9	Review	listening fill-in-the-blank exercises
10	Posttest	dialogue read-aloud test

Rater guidelines and scoring. All 28 recordings, 14 from the pretest and 14 from the posttest, were collected and then randomly ordered to be burnt on CDs and sent to the raters. Three different CD versions were produced (type A, type B, and type C) in which the order of the recordings varied. In this way, the order of the recordings would not affect the overall scores, should fatigue influence the raters' judgments towards the end of the rating process. Each voice file ran about 1.5 minutes, and so the time needed for listening to the whole CD one time was approximately 45 minutes. Subsequently, the different CD versions were randomly delivered to 46 raters by the researcher. When handing over the recordings, the researcher met with the raters individually and went over the guidelines of the rating scale and target items for approximately five to ten minutes. Other than that, there was no other rater guidance or training.

The materials given to the raters included the dialogue script (Appendix A) with highlighted phrases identifying the indicators, the rater survey (Appendix B), the instructions for rating (See Appendix C), and the item specification table (Table 5) presented above. They were asked to listen to each track at least twice in order to give separate ratings for Kris and Jay and were also told that reading errors were not to be counted as markers of pronunciation ability. Finally, the four categories, NT, WF, PT, and RF for each role were to be rated on a five point likert scale from 1 (*very poor*) to 5 (*very good*) based on how well the features were performed. An

additional comment section was added on the evaluation form (Appendix D) so that the raters could make comments for explanations of (a) particularly low or high ratings, (b) any other particular features of the performance, or (c) difficulty or easiness they had in making decisions, and so forth. This comment section was included for research purposes. It was hoped that the comments would assist the researcher in understanding how the raters were using the rating scale in more depth.

Analytical Method and Procedure

FACETS analysis. FACETS (or Multi-faceted Rasch) analysis was used in this study to investigate important factors in the test design. Unlike the test formats where examinees simply choose an answer which is either correct or incorrect, a performance assessment requires involvement of *raters* who judge the examinees' performances on a *rating scale*. Therefore, the rater and the scale that might influence test scores variably were included as other additional factors in the test. In particular, raters can be a source of variability in the scores when raters differ in severity or use of the scale (McNamara, 1996). For instance, a student being assessed by two harsh raters would not receive a comparable score to that of a student who is assessed by two lenient raters. That is to say, even if they receive the same score, it is obvious that the former student is more 'able' than the latter one. Therefore, the consequential score of an examinee (from which their ability is inferred) is not only affected by the difficulty of the item but also by the characteristics of the raters. FACETS analysis makes the investigation of the rater and other multiple factors, also called *facets,* possible. Based on overall response patterns, measures of each facet such as the ability of examinee, severity of rater, and difficulty of item are estimated and presented in *logits.* These facets can also be displayed all at once for comparison on a single *logit scale* which is a 'true interval scale' (Henning, 1984, p. 129) that has consistent interval value between units (Bond & Fox, 2007) enabling the visual representation of the relationships among the facets.

Beyond this, FACETS analysis can provide *fit statistics* for all the individual elements within each facet to see if they 'fit' the model expectation. Furthermore, *rating scale diagnostics* to see how the rating scale is used by the raters and *bias analysis* concerning "the identification of systematic patterns related to different interactions of the various facets" (Yoshida, 2004, p. 40) may also be conducted. In the present study, FACETS 3.62.0 version (Linacre, 2007b) was used to examine the four facets of the CST setting: examinee ability, item difficulty, rater severity, and the L1 background of raters.

The sum of the average ratings of the eight items was used as the total score for each examinee, and using SPSS 11.5, descriptive statistics and the inter-rater reliability coefficients for both pretest and posttest were computed separately.

Then, to investigate test-score reliability, the *person reliability indices* from the pretest and posttest's FACETS analyses were used. There are two reasons why the person reliability index from FACETS was used instead of the Cronbach alpha or Kuder-Richardson 20 which are typically reported for language test-score reliability. First of all, the person reliability index indicates the extent to which the test is able to separate examinee abilities from each other and thus is closely identifiable with test-score reliability coefficients (Linacre, 2008). Secondly, since the CST was used as a criterion-referenced test, and so the scores of the students were negatively skewed and not normally distributed especially in the posttest, those reliability coefficients that are "very sensitive to the magnitude of standard deviation" (Brown, 2005, p. 199) could not be used for the present study. Therefore, person ability reliability index, that is analogous to the K-R20 or Cronbach Alpha (Linacre, 2007a), was chosen. This reliability index, like other reliability coefficients, ranges from 0 to 1.

The effectiveness of the instructional intervention was examined by making three comparisons of how the students performed differently in posttest and pretest. First, a paired *t*-test was conducted using SPSS to examine whether the students' raw total scores from the pretest and posttest differed. Secondly, a pretest to posttest item-by-item comparison was performed by using eight paired *t*-tests with examinees' scores on each item in both tests to further verify the extent to which examinees' raw test scores actually changed for those features that were taught. Finally, the *vertical rulers* and measurement reports of the examinees' abilities from the FACETS outputs of both pretest and posttest were also juxtaposed to see how the ability (measured in logits) of the students changed. The *vertical ruler* is a graphical description of all the measures of the facets under investigation where the elements of each facet and their standings can be compared in one graph. Such comparisons show the degree to which the assessment is sensitive to instructional intervention and so is effective for measuring what the test was initially designed to measure.

For examining the detailed descriptions of each facet and its elements, the measurement reports for only the 'posttest' were used in this study. The reasons are not only due to limited space but also because the examinees took the posttest after they had learned about connected speech, and thus it would be closest to the actual classroom achievement setting for which the test was originally designed. Four facets were investigated (the examinee, item, rater, and L1

background of the raters) with only the examinee facet allowed to float and the item and rater facets centered on '0' logits. The L1 background of the raters was included as a 'dummy facet' which is a classification facet excluded from estimation but used for the purpose of examining interaction alone (Linacre, 2007a). Thus, the elements of this facet (KET group and EET group) were all anchored at '0,' since this facet is merely a classification of the raters and does not actually affect the measurement.

The rater L1 background facet was specially added to examine the question of whether native and non-native teachers rate similarly when assessing connected speech. While several studies have used FACETS to examine the relationship of L1 background differences to rater performance within oral performance tests (Brown, 1995; Caban, 2003), Yoshida's study (2004) specifically examined this facet with an analytic pronunciation test. The results revealed that that the differences found in severity among the raters were not associated with rater L1 background confirming the fact that non-native speakers (in this case Japanese speakers) were equally capable of rating pronunciation as were the native speakers of English. In the same vein, the difference in severity for the two groups (KET and EET) was examined using a paired *t*-test[1] and the fit statistics for these groups were compared to verify whether they differed in rating severity and consistency.

Finally, bias analyses were conducted in order to see whether these two different rater groups showed bias in their ratings for certain examinees or certain items. In the studies mentioned above (Brown, 1995; Caban, 2003; Yoshida, 2004), the results of bias analyses showed that bias *did* exist towards certain examinees and items across different raters or rater groups with distinct L1 backgrounds. In the case of Yoshida's study (2004), L1 Japanese raters had a larger number of significant bias interactions than the L1 English raters when rating items related to segmentals and suprasegmentals. For rater and examinee interactions, all the raters had bias interactions with certain examinees; generally the high and low level students were found to be more harshly rated than the middle level students.

[1] Ten *t*-tests were conducted within the same data set for this study in total: one for pretest-posttest total score comparison, eight for the pretest-posttest item-by-item comparison, and one for comparison between severity logits of two rater groups. Hence, it is important to note that the Bonferroni adjustment was applied (.05/10 = .005) and the more conservative alpha level ($p < .005$) was used for making statistical decisions based on these 10 *t*-tests (Brown, 2001).

RESULTS

Descriptive Statistics and Test Score Reliability

For each examinee, all the scores from the raters for each item were averaged and added up to calculate the total score. Therefore, each examinee could get a maximum score of 40, since there were eight items in total, and the score for each item ranged from 1 to 5. The Cronbach alpha *inter-rater reliability* was found to be high with a value of .98. Table 7 shows the descriptive statistics for both the pretest and posttest. The increase in the means and the change in the standard deviation immediately tell us that the examinees got higher scores on average on the posttest and their scores became less variable. A detailed comparison of the pretest and posttest performances and each examinee's standing relative to the other examinees will be discussed shortly.

The person reliability indices from the FACETS analyses were .98 for the pretest and .96 for the posttest meaning that the examinees who took the test vary reliably in ability. In other words, these high reliability indices indicate that the test is reliably differentiating examinees from each other. Furthermore the chi-square values, which examine the null hypothesis that all persons are equal, were 769.6 (df = 12) for the pretest and 300.7 (df = 12) for the posttest and were both significant at $p < .00$ level. This also confirms the fact that the examinees are significantly different from each other.

Table 7

Descriptive Statistics of Pretest and Posttest Total Scores

	N	*M*	*SD*	High	Low	Person Reliability index
Pretest	13	24.21	2.82	28.64	19.50	*.98
Posttest	13	29.90	1.64	32.95	27.73	*.96

Notes: *Pretest - fixed (all same) chi-square: 769.6, *df*: 12; significance: $p < .00$
*Posttest -fixed (all same) chi-square: 300.7, *df*: 12; significance: $p < .00$

Pretest and Posttest Gain

Initially, a paired *t*-test based on the raw total scores of the students was conducted in order to see if there was a significant difference between the pretest and posttest scores of the examinees. The result showed that the students received significantly higher scores on the posttest $t(12) = -$

9.97, $p < .005$. The examinees' scores for pretest and posttest were also compared for each item item using eight *t*-tests for each pair, and the results are summarized in Table 8. As can be seen, examinees received significantly higher scores on all eight items on the posttest (marked with asterisks) confirming once again that the examinees performed better on the posttest for all three features (WF, PT, and RF) that were taught.

Table 8

Descriptive Statistics of Pretest and Posttest Items Scores and Their Mean Comparisons

	K-NT	K-WF	K-PT	K-RF	J-NT	J-WF	J-PT	J-RF
Pretest								
M	3.20	2.79	3.52	2.72	2.90	2.55	3.20	2.60
SD	0.72	0.40	0.48	0.52	0.59	0.37	0.71	0.53
High	4.02	3.30	4.20	3.43	3.75	3.18	4.02	3.57
Low	2.07	2.16	2.55	1.84	1.82	2.11	2.18	1.89
Posttest								
M	3.56	3.48	3.96	3.69	3.42	3.67	3.85	3.63
SD	0.65	0.26	0.33	0.63	0.37	0.21	0.36	0.52
High	4.34	4.00	4.32	4.36	4.11	4.05	4.30	4.48
Low	2.07	3.07	3.20	2.43	2.89	3.32	2.91	2.84
df	12	12	12	12	12	12	12	12
t	***-4.39**	***-5.62**	***-3.89**	***-6.48**	***-5.39**	***-12.16**	***-4.17**	***-6.20**

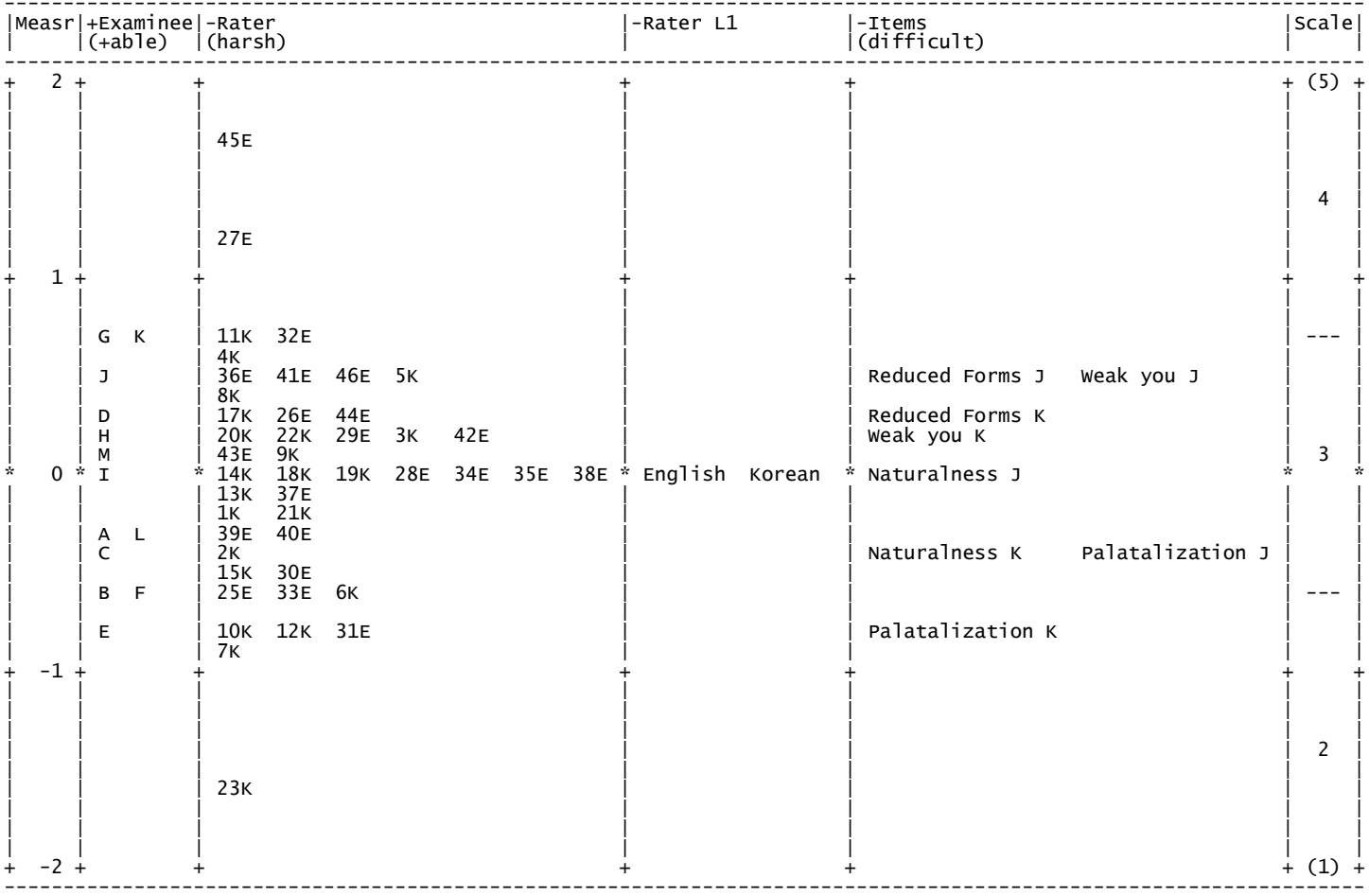
Notes: * $p < .005$

Figures 1a and 1b are the Rasch analysis vertical rulers or visual summaries for the pretest and posttest, respectively. In the case of the pretest vertical ruler (Figure 1a), the first column shows the logit scale, and notice that all four facets are displayed along this scale and can be compared to each other all at the same time. The second column shows the standings of the 13 examinees, ranging from students G and K with the highest scores (most able) to student E with the lowest score (least able). The third column shows rater severity where rater 45E is the harshest and rater 23K is the most lenient (E and K stands for English and Korean L1 raters, respectively). The fourth column shows the rater L1 background facet (English or Korean) and

notice that the two groups are anchored at ‘0’ logits, since they were set as dummy facets. Finally the items are shown in the fifth column with the most difficult item on the top (J-RF, Reduced Forms for Jay’s role) and the easiest on the bottom (K-PT, Palatalization for Kay’s role).

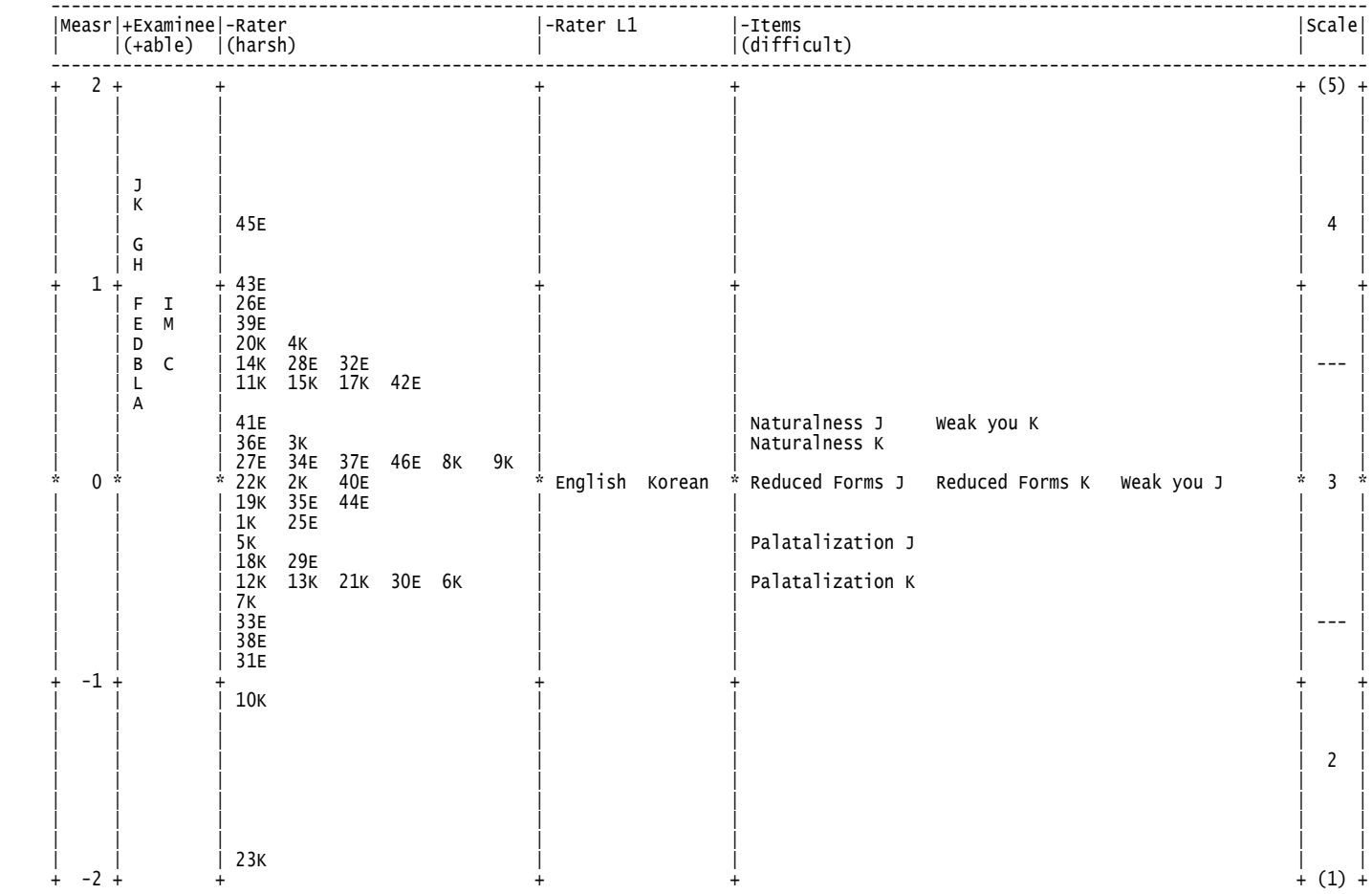
Compared to the pretest, the performances of the examinees, raters, and items differed in the posttest. It is interesting to see that the examinees’ abilities shifted upwards in the posttest ruler, with everyone above 0 on the logit scale. Besides improvement in performances, there was less variation compared to the pretest. Table 9 demonstrates how the ability logit of each examinee changed from the pretest to the posttest. All 13 examinees’ ability logits increased, ranging from a minimum increase of 0.38 logits to a maximum of 1.63 logits.

Another noticeable change is the alteration of the items’ difficulty measures. The naturalness items, K-NT and J-NT, actually switched positions with the reduced form items, K-RF and J-RF, and became the most difficult items in the posttest. This indicates that although the examinees performed better in the posttest overall, the naturalness items became more difficult in the posttest. In contrast, the palatalization items, K-PT and J-PT, remained as the easiest items in the posttest. For the weak form items, K-WF and J-WF, their positions switched and K-WF became more difficult than J-WF, which is opposite to the pattern seen in the pretest.



Notes: Rater labels are made up of number and E for EETs and K for KETs

Figure 1a. Vertical Ruler for Pretest Four Facet Analysis (Examinee, Rater, Item, & Rater L1 Background)



Notes: Rater labels are made up of number and E for EETs and K for KETs

Figure 1b. Vertical Ruler for Posttest Four Facet Analysis (Examinee, Rater, Item, & Rater L1 Background)

Table 9

Examinees' Ability Measurement Comparison Report for Pretest and Posttest

Examinee	Pretest		Posttest		Change in
	Ability logit	Model error	Ability logit	Model error	Ability
A	-0.30	0.06	0.42	0.06	0.72
B	-0.65	0.07	0.60	0.06	1.25
C	-0.37	0.06	0.63	0.06	1.00
D	0.31	0.06	0.69	0.06	**0.38**
E	-0.80	0.07	0.83	0.06	**1.63**
F	-0.64	0.07	0.85	0.06	1.49
G	0.74	0.06	1.20	0.07	0.46
H	0.22	0.06	1.11	0.07	0.89
I	0.04	0.06	0.92	0.07	0.88
J	0.49	0.06	1.50	0.07	1.01
K	0.71	0.06	1.38	0.07	0.67
L	-0.29	0.06	0.49	0.06	0.78
M	0.06	0.06	0.79	0.06	0.73
M	-0.04	0.52	0.88	0.07	
SD	0.52	0.00	0.34	0.00	

Facet Effects

Examinees. Table 10 shows the report for each of the 13 examinees that took the posttest in ascending order of ability from the least able examinee (A) to the most able examinee (J). Each column from left to right presents the examinee label, examinee's ability (measured in logits), error, the mean square value and *t* value for model fit. As seen in the vertical ruler (Figure 1b), the students' ability measures are all above 0 on the logit scale where '0' means having a 50% chance of getting the raw score of 3 on the rating scale when being rated on a average difficulty item by a rater of average severity. Although there seems to be a narrow spread among the examinees, the reliability index was high (.98) and the chi-square of 300.7 ($df = 12$) was significant at $p < .00$ level, indicating that the examinees differ consistently from each other in ability.

The fourth and fifth columns show the fit statistics, which provide information on how the observed empirical data fit the model by calculating the discrepancies between the expected

estimate and the actual observed values. Fit is then reported in terms of mean square values or converted into standardized *t* values. The mean square values have an expected value of 1 and, depending on the variation of the observed value, the mean square could be less than 1 when there is less variation than expected and more than 1 when there is greater variation than expected. Mean square values between 0.75 and 1.3 are seen as acceptable, and as for the *t*-value, values inside the range of -2 to +2 are considered to be acceptable. Values larger than +2 would indicate 'misfit' (not meeting the model in an unpredictable way) and values less than -2 would describe 'overfit' (not meeting the model in a 'too' predictable way, Bond & Fox, 2007).

According to the fit statistics in Table 10, we can see examinees C, D, E, and K are showing overfit while examinees L and H are misfitting. Overfit of examinees means that their performances on the test have less variation than expected by the model and thus are too predictable. For example, if a person gets all the difficult items wrong and all the easy items correct with no exceptions, this person's performance is seen as "too good to be true" (Bond & Fox, 2007, p. 240). However, this is not considered to be as problematic as misfitting examinees, whose performances are unpredictable and show too much deviation from the model. For instance, an examinee who would get easy items wrong and then difficult items right would be identified as showing unpredictable or misfitting performance. According to McNamara (1996), the cause for person misfit could be of various reasons such as guessing, different levels of mastery for certain skills, or even the condition of the test taker. He suggests that it is useful to review the actual score records of those certain misfitting students to examine what the problem might be, and accordingly, that is what was done.

Table 10

Ability Measurement Report for Examinees in Posttest

Examinee	Ability logit	Model error	Infit	
			Mean Square	*t*
A	0.42	0.06	1.01	0.00
L	0.49	0.06	1.54	***6.60**
B	0.60	0.06	1.07	0.90
C	0.63	0.06	0.85	**-2.10**
D	0.69	0.06	0.84	**-2.30**
M	0.79	0.06	1.12	1.60
E	0.83	0.06	0.77	**-3.40**
F	0.85	0.06	1.00	0.00
I	0.92	0.07	0.99	-0.10
H	1.11	0.07	1.21	***2.70**
G	1.20	0.07	0.90	-1.30
K	1.38	0.07	0.84	**-2.20**
J	1.50	0.07	0.88	-1.60
M	0.88	0.07	1.00	-0.10
SD	0.34	0.00	0.21	2.70

Notes: Reliability of separation index = .96; fixed (all same) chi-square: 300.7, *df*: 12; significance: $p < .00$

* Misfitting examinees

Table 11 shows the overall average raw scores for each item for those examinees who were found to be overfitting (C, D, E, and K) or misfitting (H and L). The items are ordered horizontally from the most difficult (left) to the easiest (right), and the examinees are ordered vertically from the least able (top) to the most able (bottom). The pattern that would be expected would essentially be the lowest scores in the upper left corner, highest scores in the lower right corner and a smooth transition of scores in between (i.e., scores should steadily increase as you move down or to the right). Although erratic performances on certain items cannot easily be spotted for misfitting examinee H ($t = 2.70$) except for the sudden drop in the scores of items J-WF and K-PT, examinee L's performance shows a remarkably inconsistent pattern. Not only does he perform better than more able examinees on a number of the items (g., J-NT, K-WF, J-RF, J-WF, and J-PT), he performs quite poorly all of a sudden for items K-NT, K-RF, and especially for K-PT which is actually the easiest item of all.

Table 11

Response Patterns of Overfitting and Misfitting Examinees for Eight Items in Posttest

Examinee	Item								Total
	Most difficult							*Easiest*	
	J-NT	K-WF	K-NT	J-RF	J-WF	K-RF	J-PT	K-PT	
***L**	3.66	3.30	**2.07**	3.82	4.05	**2.43**	4.16	**3.66**	27.15
C	3.14	3.07	3.27	2.98	3.59	4.16	3.77	3.95	27.93
D	3.23	3.57	3.82	2.84	3.43	3.55	3.84	4.02	28.30
E	3.14	3.59	3.16	4.07	3.64	3.73	3.84	3.89	29.06
***H**	3.95	3.16	3.77	4.48	**3.64**	4.23	4.11	**3.20**	30.54
K	3.64	3.41	4.25	3.91	3.89	4.18	4.25	4.32	31.85

Notes: * Misfitting examinees

Items. The estimated difficulty measures and the fit statistics for all eight items are shown in Table 12, organized in ascending order. Unlike the ability estimates for the examinees, the item difficulty logits range from negative (-0.47) to positive value (0.34) with the logit value '0' being average difficulty. The items are not spread out very widely on the logit scale, although there are reliable differences in terms of difficulty as shown by the reliability index (.96) and the chi-square value (192.8, $df = 7$) significant at the $p < .00$ level. These indicate that the items in the test are consistently different in difficulty, indicating that the test would reveal similar results if the same items were given to another similar group of people with the same behavior in performance (Bond & Fox, 2007, p. 41).

The palatalization items, K-PT and J-PT, were the easiest items, with the reduced form items (K-RF and J-RF) and weak form items (K-WF and J-WF) being approximately in the middle, and finally the naturalness items were the most difficult. One noticeable point is that two items of the same feature, as in the case of J-WF and K-WF, may vary in terms of difficulty (with a 0.28 logit difference), where K-WF is more difficult. This could also be seen with the reduced form items (J-RF and K-RF).

Table 12

Difficulty Measurement Report for Items in Posttest

Item	Difficulty logit	Model error	Infit	
			Mean Square	*t*
K-PT	-0.47	0.05	0.99	-0.20
J-PT	-0.29	0.05	1.03	0.50
K-RF	-0.03	0.05	1.03	0.50
J-WF	-0.01	0.05	1.00	0.00
J-RF	0.04	0.05	1.09	1.60
K-NT	0.16	0.05	0.97	-0.50
K-WF	0.27	0.05	1.16	***2.80**
J-NT	0.34	0.05	0.79	**-4.00**
M	0.00	0.05	1.01	0.10
SD	0.27	0.00	0.11	2.00

Notes: Reliability of separation index = .96; fixed (all same) chi-square: 192.8, *df*: 7; significance: $p < .00$
* Misfitting item

Fit statistics show that the K-WF item is misfitting with a *t*-value of 2.80 and J-NT item is overfitting with a *t*-value of -4.00. Fit indices for items allow us to determine whether the items satisfy the condition of *unidimensionality* referring to "the focus of one attribute or dimension at a time" (Bond & Fox, 2007, p. 32). In other words, the fit index of each item tells us whether the item is contributing well to the test measuring a single dimension or construct. The item that is found to be misfitting means that the item is departing unacceptably from the predicted difficulty pattern of the other items and thus might indicate that the item is not working as in the same way as the other items. For instance, if there is a minimal pair pronunciation item included in a reading test, doing well on this item would not necessarily predict good performance on the other items of the test because it measures a different ability.

An overfitting item, on the other hand, means that the responses to this item are too predictable. McNamara (1996) says that overfit items may be 'redundant' because they are "just doing what all the other items are doing in combination" (p. 218), as the items are heavily dependant on the scores of the other items. For instance, the rating given to an item that is asking for an overall wholistic rating of a performance (e.g., overall speaking ability) could be strongly influenced by the ratings of other items (e.g., pronunciation ability, fluency, grammar, and so forth). This could be the reason why the J-NT item, requiring a more holistic rating, was found to be overfitting. In fact, according to Bond and Fox (2007), items asking for overall

impressionistic ratings are ‘typical’ overfitting items, and omitting such an item that does not provide any additional information, might be preferable in the sense that it could cause the model frame to shift and enable misfitting item(s) to fit the model. Therefore, another analysis was conducted after excluding the overfitting J-NT item, and the result showed no overfitting or misfitting items (See Table 13). With the *t*-value of 1.70, item K-WF was no longer misfitting.

Table 13

Difficulty Measurement Report for Items in Posttest without J-NT Item

Item	Difficulty logit	Model error	Infit	
			Mean Square	*t*
K-PT	-0.41	0.05	0.95	-0.80
J-PT	-0.24	0.05	1.01	0.20
K-RF	0.01	0.05	.98	-0.20
J-WF	0.04	0.05	.97	-.50
J-RF	0.09	0.05	1.09	1.50
K-NT	0.20	0.05	0.94	-1.10
K-WF	0.31	0.05	1.10	**1.70**
M	0.00	0.05	1.01	0.10
SD	0.25	0.00	0.07	1.20

Notes: Reliability of separation index = .96; fixed (all same) chi-square: 137.6, *df:* 6; significance: $p < .00$

Raters. Table 14 is the severity measurement report for the raters in the posttest. The first to fifth columns show the measures of severity, error, and fit statistics of the KET raters, and the sixth to the last columns show measures and statistics for the EET raters arranged in ascending order of severity. The raters are the most widely spread out with severity measures ranging across more than 3 logits, from -1.87 (23K) to 1.30 (45E), which was much greater than the range of measures seen for examinees or items.

The reliability index (.96) and chi-square value (971.2, $df = 43$, $p < .00$) show that the raters are significantly different in terms of severity, which means they show consistent disagreement in their ratings. Note that this is different from ‘inter-rater’ reliability which is actually the correlation of the raters’ ratings. Being highly correlated does not necessarily mean that they are rating identically. It tells that there is consistency among the raters’ ‘rank order’ of the examinees, but it does not provide information regarding how the raters differ in severity (Bond & Fox,

2007). Therefore, although the inter-rater reliability is high (.98) showing that the raters' ratings strongly correlate, the high Rasch reliability index (.96) tells us that they are consistently different in terms of severity.

While the range of severity measures and the reliability index indicate how raters are different from each other in severity, the fit statistics tell us the intra-rater consistency of each rater. As could be seen in Table 14, a number of raters were found to be misfitting or overfitting, which is not a good sign. There are 10 misfitting raters (marked with an asterisk) with *t*-values over +2 (8K, 27E, 29E, 45E, 21K, 38E, 18K, 1K, 33E, 44E: from largest to smallest misfit) indicating that these raters were rating in unpredictable ways. Moreover, 14 overfitting raters with *t*-values under -2 (22K, 40E, 34E, 39E, 32E, 4K, 35E, 43E, 37E, 14K, 36E, 41E, 17K, 15K: from largest to smallest overfit) report that these raters were using the rating scale with very little variation.

The possible explanation for the overfitting raters could be because the raters were not using the whole scale but only a part of it, such as in the case where the rater avoids using extreme scores (*central tendency*) or because of *halo effect*, a common rater error (Engelhard, 1994), where all or most of the items are rated similarly or identically indicating the rater's wholistic approach to the items (e.g., 4444, 3333). To see if this was the reason for the overfitting raters in the present study, the rating pattern of the eight items for each examinee was reviewed. This was exactly the case, and Table 15 shows an example of each rater error type using the rating observations of the two highly overfitting raters (22K: -5.90; 39E: -5.50)

Table 14

Severity Measurement Report for Raters in Posttest

KET	Severity	Model	Infit		EET	Severity	Model	Infit	
Raters	Logit	error	Mean Square	*t*	raters	Logit	error	Mean Square	*t*
23K	-1.87	0.18	1.32	1.90	31E	-0.88	0.14	1.08	0.60
10K	-1.11	0.15	1.03	0.20	38E	-0.85	0.14	1.72	***4.20**
7K	-0.61	0.13	1.05	0.30	33E	-0.70	0.13	1.41	***2.60**
12K	-0.54	0.13	0.97	-0.20	30E	-0.46	0.13	1.23	1.50
13K	-0.49	0.13	1.09	0.60	29E	-0.39	0.13	1.96	***5.40**
21K	-0.48	0.13	1.75	***4.40**	25E	-0.24	0.12	0.80	-1.40
6K	-0.46	0.13	1.25	1.70	35E	-0.13	0.12	0.55	**-3.90**
18K	-0.38	0.13	1.62	***3.80**	44E	-0.09	0.12	1.33	***2.20**
5K	-0.25	0.12	1.17	1.20	40E	-0.05	0.12	0.39	**-5.90**
1K	-0.19	0.12	1.47	***3.00**	27E	0.05	0.12	1.98	***5.70**
19K	-0.12	0.12	0.94	-0.40	34E	0.07	0.12	0.41	**-5.70**
2K	-0.05	0.12	1.06	0.40	46E	0.11	0.12	0.77	-1.80
22K	-0.03	0.12	0.39	**-5.90**	37E	0.15	0.12	0.59	**-3.50**
8K	0.07	0.12	2.35	***7.40**	36E	0.19	0.12	0.69	**-2.60**
9K	0.08	0.12	0.94	-0.40	41E	0.31	0.11	0.71	**-2.40**
3K	0.22	0.12	1.00	0.00	42E	0.48	0.11	0.91	-0.60
17K	0.49	0.11	0.72	**-2.30**	28E	0.57	0.11	0.89	-0.80
11K	0.50	0.11	1.12	0.90	32E	0.59	0.11	0.48	**-5.00**
15K	0.50	0.11	0.73	**-2.20**	39E	0.81	0.11	0.45	**-5.50**
14K	0.62	0.11	0.63	**-3.30**	26E	0.90	0.11	0.86	-1.10
20K	0.65	0.11	0.88	-0.90	43E	1.03	0.11	0.61	**-3.60**
4K	0.70	0.11	0.52	**-4.60**	45E	1.30	0.11	1.69	***4.60**
M all	0	0.12							
SD	0.61	0.01							
M KET	**-0.13**	0.12							
SD	0.62	0.01							
M EET	**0.13**	0.12							
SD	0.59	0.01							

Notes: *All raters*, Reliability of separation index = .96; fixed (all same) chi-square: 971.2, *df:* 43; significance: $p < .00$
KET raters, Reliability of separation index = .96; fixed (all same) chi-square: 433.0, *df:* 21; significance: $p < .00$
EET raters, Reliability of separation index = .96; fixed (all same) chi-square: 502.1, *df:* 21; significance: $p < .00$
* Misfitting raters

Table 15

Example of Overfitting Rater Pattern in Posttest

22K (halo effect)								**39E (central tendency)**							
KNT	KWF	KPT	KRF	JNT	JWK	JPT	JRF	KNT	KWF	KPT	KRF	JNT	JWK	JPT	JRF
3	4	4	3	3	4	4	4	3	3	3	2	3	3	3	3
4	4	4	3	2	4	4	4	3	2	2	3	3	3	2	2
4	4	4	4	3	4	4	3	3	3	3	3	3	3	2	2
4	4	3	3	3	4	4	4	4	3	3	3	3	3	3	3
3	4	4	4	3	4	4	4	3	3	3	3	3	3	3	3
2	4	4	4	2	4	4	4	**3**	**3**	**2**	**2**	**3**	**3**	**3**	**3**
4	4	4	4	3	4	4	4	4	4	4	4	3	3	3	3
4	**4**	**4**	**4**	**4**	**3**	**4**	**4**	4	3	4	3	3	3	3	4
4	3	4	4	3	4	3	4	4	4	4	4	3	3	3	3
4	4	4	4	4	4	4	4	4	3	3	3	4	4	4	4
4	4	4	4	2	4	4	4	5	4	4	4	4	3	3	3
2	3	4	3	4	4	4	4	2	2	2	2	2	2	2	2
4	4	4	4	3	4	4	3	4	4	4	3	3	3	3	3

Although overfitting and misfitting raters are both problematic, having many misfitting raters is considered a more serious problem (Bachman, Lynch & Mason, 1995). The reason for the large number of inconsistent raters was further investigated by reviewing the comments that were made on the evaluation forms, and certain comments were found to appear repeatedly. Interestingly, most of them expressed the raters' confusion when having to rate examinees that did produce the target forms in their speech but did not necessarily sound natural. Many raters mentioned that it was difficult to decide whether an examinee was to be rated as 'good' since the target forms existed in the speech, or 'poor' because they did not sound natural. The following examples show a few of their comments.

"Quantity or quality?"

"Forcing the target form caused unnaturalness: over-generalized and overstressed ya"

"Maybe overstressing was due to the emphasis of this point in their instruction."

Especially for examinee L who exhibited extremely erratic performance, many of the comments from the raters aligned with this same issue stating that he did not sound natural despite the fact he was producing the target forms. A few raters, therefore, added that they

decided to give a neutral rating of ‘3’ for those items, which were ambiguous. This kind of uncertainty in the rating process could have been one of the major causes of the many misfitting raters.

Rating scale diagnostics. Rating scale diagnostics provide information on how the rating scale is functioning by giving us frequency measurement reports for each point of the scale and the *step difficulty threshold,* which is essentially the cut-point for each point on the scale. Table 16 is the summary of the scale diagnostics for the posttest. The first column shows each point on the scale from ‘1’ to ‘5’. Then the second and third columns present the frequency counts and their percentage values so that we could see how often each point is being used. Notice how ‘3’, ‘4’, and ‘5’ are used the most. The fourth column is the average measure for each point. It reports the average ability (in logits) of all of the examinees who received that point on any of the items in the test. So, the ability measure of 1.55 for ‘5’ on the scale shows that 1.55 is the average ability for all the examinees that received a score of ‘5’ on any of the items on the test. The average measures increase from 0.16 to 1.55 along the scale which indicates that the points of the scale actually represent steps of increasing difficulty. The fifth column contains the fit statistics, and according to Bond and Fox (2007), outfit mean squares that are greater than ‘2’ indicate that the particular point on the scale is causing ‘noise’ in the measurement process. However, in the current diagnostics, no rating point was found to be troublesome in terms of fit. Finally the sixth column shows step difficulty thresholds. Fair distance among the thresholds demonstrates that each point defines a distinct position in the measure of the construct. Thresholds are ideal when they are at least 1.4 logits apart but less than 5 logits apart (Bond & Fox, 2007). Yet, the result for the posttest shows that the distance between the thresholds of points ‘3’ and ‘4’ is not desirable (0.89) compared to the distances between other threshold points which are ideal (between ‘2’ and ‘3’: 1.66; between ‘4’ and ‘5’: 2.40).

Table 16

Rating Scale Diagnostics (Frequency Measurement Report for Rating Scale in Posttest)

Rating Scale	Count	%	Average Measure	Outfit Mnsq	Step difficulty
1	66	1%	0.16	1.2	
2	483	11%	0.37	1.2	-1.88
3	1254	27%	0.57	0.9	-0.52
4	1925	42%	0.94	1.0	0.37
5	848	19%	1.55	0.9	2.03

These threshold estimates can also be visually represented by the intersection of probability curves (See Figure 2). Probability curves show the degree to which each point on the scale is distinct and overlaps with each other. If there is too much overlap among the curves and the curves are relatively flat, this would mean that those points on the scale are not distant from each other and thus are not functioning ideally. Considering Figure 2, notice that there is considerable overlap especially for the curve of point ‘3’ which suggests that ‘3’ is not serving as a distinctive point on the scale.

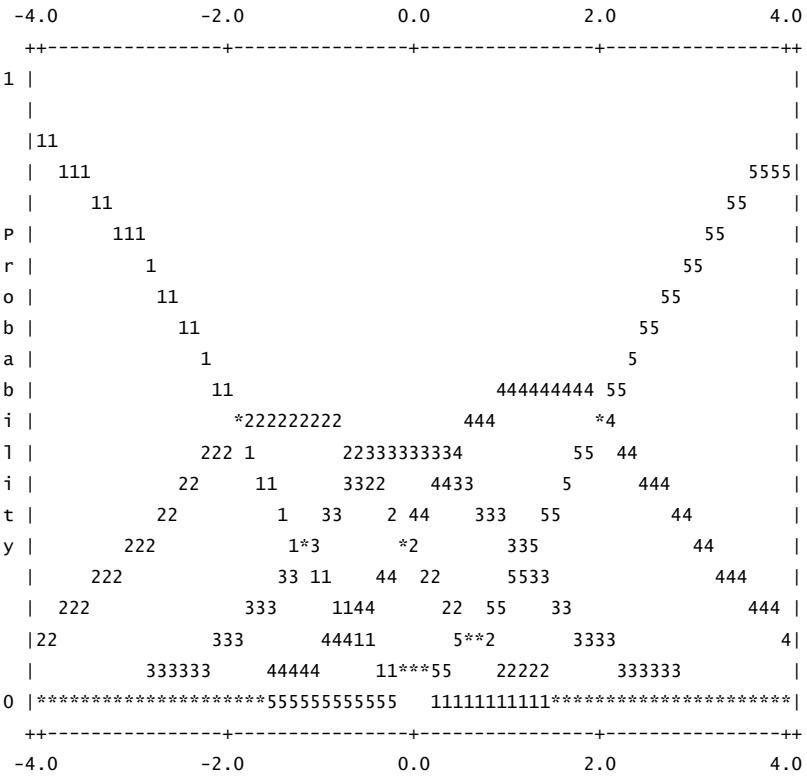


Figure 2. Rating Scale Probability Curves (Posttest)

Rater L1 background. The differences in the severity and consistency between the two rater groups (KET and EET) are summarized in Table 17. The range of the severity measures differs in two groups; KET raters' measures range from -1.87 to 0.70 and the EET raters' range from -0.88 to 1.30, with the most severe rater being from the EET group and the most lenient rater being from the KET group. The mean severity logit of the KET group was -0.13 and 0.13 for the EET group showing 0.26 logit difference. A paired *t*-test was conducted to determine whether the difference in severity is statistically significant, and the result revealed that the EET group rated significantly more harshly than the KET group ($t(21) = -5.51$, $p < .005$). In terms of consistency, the EET group had a larger number of raters that were both misfitting ($n = 6$) and overfitting ($n = 9$) than the KET group (misfit: $n = 4$; overfit: $n = 5$).

The severity among the raters *within* each group was found to differ for both KET and EET groups; the reliability indices for the groups were the same (.96) with significant chi-square values (KET: 433.0, $df = 21$, $p < .00$; EET: 502.1, $df = 21$, $p < .00$). However, the inter-rater reliability coefficients revealed that both groups had high inter-rater reliability (KET = .98; EET: .97).

Table 17

Comparison Report of KET and EET Rater Groups

	n	Severity logits						Misfits			Overfits		
		low	high	*M*	*SD*	*t*	*df*	*n*	low	high	*n*	low	high
KET	22	-1.87	0.70	**-0.13**	0.62	***-5.5**	21	4	3.00	7.40	5	-5.90	-2.20
EET	22	-0.88	1.30	**0.13**	0.59			**6**	2.20	5.70	**9**	-5.90	-2.40

Notes: KET raters, Reliability of separation index = .96; fixed (all same) chi-square: 433.0, *df:* 21; significance: $p < .00$

EET raters, Reliability of separation index = .96; fixed (all same) chi-square: 502.1, *df:* 21; significance: $p < .00$

* $p < .005$

Rater L1 interaction with examinees and items. Finally, bias analyses were conducted to see if there were any rater group (KET or EET) interactions with certain examinees or items. A single rater group might show a certain pattern of harshness or leniency towards particular examinee(s) or specific item(s) and not others. Table 18 shows the measures for all possible interactions between the two rater groups and the 13 examinees. The second and third columns report the observed score in the data and the expected score that the model predicts, and then the discrepancy between the observed and the expected scores are calculated into bias measures (column 4) with the error value for each measure in the following column (column 5). The sixth

column displays the converted standardized t scores of the bias measures, and as in the case of fit *t*-values, bias *t*-values within -2 to +2 are considered acceptable. In other words, t-values above +2 or below -2 indicate the presence of statistically significant bias. If the rater-examinee interaction *t*-value is below -2, we can say that the rater is systematically rating the examinee more leniently compared to others. For a *t*-value over 2, it would mean the opposite; the rater is rating that examinee more harshly. It is intriguing to see that no significant bias interaction was found between both rater groups and the examinees. Although large bias sizes were spotted for interactions between both groups and examinees E (KET: -0.16; EET: 0.15) and K (KET: -0.11; EET: 0.11), the *t*-values indicate that these are not significant for they are within the range of -2 to 2. This tells us that none of the rater groups rated any examinee more harshly or leniently than the others.

Table 18

Rater L1 Group x Examinee Bias Interaction Report for Posttest

	L1 - Korean (KET)					L1 - English (EET)				
Ex	Observed Score	Expected Score	Bias Size	Error	*t*	Obsvd Score	Exp. Score	Bias Size	Error	*t*
A	610	604.1	0.05	0.09	0.52	567	572.8	-0.04	0.09	-0.51
B	626	626.1	0.00	0.09	-0.01	596	595.7	0.00	0.09	0.02
C	644	629.6	0.12	0.09	1.31	585	599.3	-0.11	0.09	-1.27
D	635	637.4	-0.02	0.09	-0.22	610	607.5	0.02	0.09	0.22
E	635	653.4	**-0.16**	0.09	**-1.71**	643	624.4	**0.15**	0.09	**1.68**
F	660	656.3	0.03	0.09	0.34	624	627.5	-0.03	0.09	-0.32
G	694	695	-0.01	0.10	-0.10	670	668.8	0.01	0.09	0.11
H	690	685.4	0.04	0.10	0.45	654	658.4	-0.04	0.09	-0.42
I	671	664.6	0.06	0.09	0.60	630	636.2	-0.05	0.09	-0.57
J	725	724.3	0.01	0.10	0.08	700	700.5	-0.01	0.10	-0.05
K	702	712.8	**-0.11**	0.10	**-1.09**	699	688	**0.11**	0.10	**1.07**
L	613	612.4	0.00	0.09	0.05	581	581.5	0.00	0.09	-0.04
M	647	649.5	-0.02	0.09	-0.24	623	620.3	0.02	0.09	0.24

Table 19 is the bias interaction report for rater group and items. Unlike the case with the examinees, two significant biases existed. Both rater groups were found to be biased towards the

same item (J-NT) but in opposite directions. The KET group consistently rated this item leniently than other items (*t* = -2.31), and the EET group rated the item more harshly *(t* = 2.28). Besides this, no other item(s) were found to have significant bias interaction with the two rater groups.

Figure 13 shows this bias pattern visually with the horizontal axis representing the items and the vertical axis showing the *t*-value of the bias interaction. The significant bias interactions (values in excess of ± 2.0) between both rater groups and the J-NT item are immediately noticeable (circled points), and although not significant, notice that the EET group raters tended to be harsher regarding most of the items except the ones concerning naturalness (K-NT and J-NT).

Table 19

Rater L1 Group x Item Bias Interaction Report for Posttest

Item	L1 - Korean (KET)					L1 - English (EET)				
	Obsvd Score	Exp. Score	Bias Size	Error	*t*	Obsvd Score	Exp. Score	Bias Size	Error	*t*
K-NT	1028	1040.9	-0.07	0.07	-0.93	1006	992.9	0.07	0.07	0.92
K-WF	1024	1019	0.03	0.07	0.36	965	969.9	-0.02	0.07	-0.34
K-PT	1169	1152.2	0.11	0.08	1.33	1095	1111.5	-0.09	0.07	-1.26
K-RF	1087	1076.9	0.06	0.07	0.74	1021	1030.8	-0.05	0.07	-0.70
J-NT	971	1003.8	-0.16	0.07	**-2.31**	987	954.0	0.16	0.07	**2.28**
J-WF	1081	1073	0.04	0.07	0.59	1019	1026.7	-0.04	0.07	-0.55
J-PT	1128	1122.8	0.03	0.08	0.40	1075	1079.9	-0.03	0.07	-0.36
J-WF	1064	1062.3	0.01	0.07	0.12	1014	1015.4	-0.01	0.07	-0.10

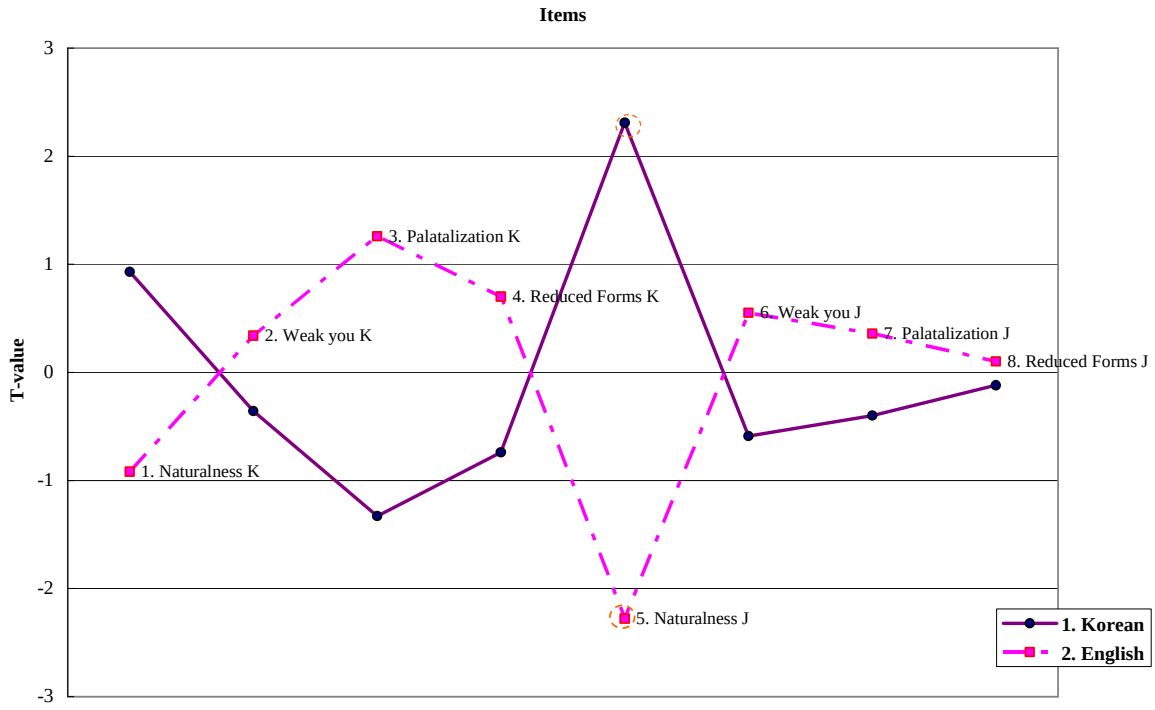


Figure 3. Bias Interaction: Rater L1 with Items (Posttest)

DISCUSSION

The results of the analyses will be reviewed and further discussed in the order of the research questions.

Does the Test Produce Reliable Test Scores?

The calculated inter-rater reliability alpha value for the test was high at .98, and the person reliability indices from the FACETS analyses for both pretest (.98) and posttest (.96) show that the test is differentiating examinees' abilities from each other well. Thus, this indicates "the replicability of person ordering we could expect if this sample of persons were given another parallel set of items measuring the same construct" (Bond & Fox, 2007, p. 40), and therefore we can conclude that the test scores are to a high degree reliable.

Is the Instructional Intervention Effective?

Validating a test is important for the test developer to ensure that the test is appropriately designed for measuring "what it claims, or purports, to be measuring" (Brown, 2005, p. 220). Although it is impossible to demonstrate criterion-related validity due to the absence of a pre-

existing criterion scale for connected speech, content and construct validity, the other two ways for investigating validity (Brown, 2005), are more pertinent for this CST. One of them, construct validity, is supported by the results of the present study. The fact that the students were able to perform better in the posttest and become more capable in producing all three target features covered in class shows the test is sensitive to instructional intervention and that the construct (ability to produce certain connected speech features) is being measured by the test. Specifically, all of the examinees' ability measures increased in value for the posttest, which was visually apparent in the vertical rulers as well (Figure 1a and 1b). These results demonstrate not only the construct validity of the test, but also the teachability of producing connected speech. In terms of 'unidimensionality,' or the issue of whether the items in the test are indeed measuring a single construct, the Rasch analysis fit statistics in the item difficulty report indicated that, by deleting one redundant item, all the other items 'fit' the unidimensional model.

According to McNamara (2000), content validity involves thinking of the content, for instance, how relevant it is to the construct being measured and what the test takers are asked to do. The CST, then, can be said to be valid in terms of content since the test provides an assessment of the target connected speech forms. The test content includes commonly used connected speech forms identified in the connected speech literature, and reflects the material covered during instruction.

How do the 13 Examinees Differ in Terms of Ability and How Well Do Their Performances 'Fit' the Model?

The examinees' ability logits ranged from 0.42 to 1.50 with reliable differences among them. Examinee ability measures in the posttest were also skewed somewhat above the range of the item difficulty due to the instruction, which is preferable for a criterion-referenced test since it indicates 'mastery'. The results of the fit statistics and in-depth examination of the misfitting and overfitting examinees' performances show that four examinees had a rather too predictable pattern in their performances while two problematic misfitting examinees received unexpected scores for certain items. As discussed in the results section, the comments from the raters, especially regarding examinee L's producing the target forms too forcefully and unnaturally, offered some explanation for inconsistent scores. In addition to those comments, several raters also pointed out that L was mumbling a lot and had frequent unnecessary pauses in his speech that often made his performance hard to rate.

How Do the Eight Items Differ in Terms of Difficulty and How Well Do They 'Fit' the Model; In Other Words, Do They Measure a Single Construct?

The item difficulty differed significantly (reliability index = .96) ranging from -0.41 to 0.34 with palatalization items (K-PT and J-PT) being the easiest of all, weak form of 'you' and reduced forms generally coming next, and finally naturalness being the most difficult. However, as mentioned previously, the J-WF item was more difficult than K-WF. Referring back to the item specification table (Table 5), all of the indicators for item K-WF involve two connected speech features simultaneously: palatalization and the weak form of 'you' (e.g., the question phrase 'did you (Di/ʤə/)'). In contrast, item J-WF had indicators involving only the latter. This could explain why producing the weak form of 'you' might have been more difficult when performing Kris's role. The fit statistics for the items indicated item J-NT to be overfitting (t = -4.00) and K-WF to be misfitting (t = 2.80). However repeating the analysis without the redundant overfitting J-NT item resulted in all the other items perfectly fitting the model. This shows that eliminating the J-NT item is preferable, in the sense that it is one less item to rate for the rater without any change in the reliability (.96). Moreover, this makes sense since the K-NT item is also for rating the examinee's overall naturalness in pronunciation, and so there is no need to have another naturalness rating for Jay's role.

How Do the 44 Raters Differ in Terms of Severity in Rating and How Well Do They 'Fit' the Model; In Other Words, Are Their Ratings Consistent?

The 44 raters differed significantly from each other in terms of severity with a reliability index of .96, and the range was noticeably wide going from -1.87 to 1.30. Fit t-values revealed the raters that may need more training. Almost 23% (10 out of 44) of the raters showed misfit which means that their ratings were inconsistent, and nearly 32% (14 out of 44) of the raters were found as overfitting which indicates that significantly less variance was found in their ratings than what the model expected. The overfitting raters were explained by common rater errors such as central tendency and halo effect, and the probable causes of rater misfit were investigated using qualitative data obtained through the raters' comments. The most common concern raised among the raters was the issue of 'quality or quantity'; that is, whether the examinee's score should be based upon the 'doing' of the target forms or the 'excellence' of the doing.

How is the Rating Scale Being Used by the Raters?

Results of the rating scale diagnostics confirm that the points on the scale in fact represent increasing levels of ability with '1' indicating the lowest ability to '5' indicating the highest

ability of the measured construct. However, the large overlapping areas of the probability curves and the small distance between threshold estimates of points ‘3’ and ‘4’ (0.89) indicate that point ‘3’ is not functioning in a way distinct from the other points on the scale.

How Do the KET Group and EET Group Differ in Terms of Rating Severity and Consistency?

The two groups’ severity measures (KET = -0.13 logits; EET = 0.13 logits) differed from each other where the EET raters rating more harshly than the KET raters. In terms of consistency, more misfits and overfits were found in the EET group than in the KET group. This contradicts the findings in Yoshida’s study (2004) where in a general pronunciation test the difference in rating severity was not related to the L1 background of the raters. However, each group showed high inter-rater reliability (KET = .98; EET = .97), and the reliability index for both groups had a value of .96, indicating the degree to which the raters showed consistent differences in severity from each other.

Does the KET Group or EET Group Display Bias Towards Any Certain Examinee(s) or Item(s)?

The bias interaction report for rater L1 group interaction with examinees showed that KET and EET raters did not rate any certain examinees more harshly or leniently than the others. However, the report for rater L1 group and item interaction showed significant bias interactions between the two rater groups and a single item (J-NT). It was found that the Korean L1 raters were rating this item more leniently than the other items, and in contrast, the English L1 raters were rating it more harshly. An interesting pattern was also discovered where all items except the two naturalness items (K-NT and J-NT) were being rated more harshly by the Korean L1 raters and more leniently by the English L1 raters than predicted by the Rasch model. Finally, it is worth noting that the only item that showed significant bias interaction with the rater groups was also the only overfitting item. Once again, this suggests that getting rid of this item would not only result in other items fitting better to the model, but also eliminate biased interaction.

CONCLUSION

Limitations

The limitations of this study lie in the small number of examinees and the relatively small differences in their abilities. Although there were a large number of raters in this study, performances from only 13 examinees had to be used because of the limited number of students in the study abroad program. The students also did not vary much in ability, with their ability

measures ranging within two logits, and this caused limited variance in the data. Finally, the length of period they received explicit instruction on the target forms was two weeks followed immediately by a posttest. This might have been the reason why many of the students were forcefully and unnaturally producing the target forms. If more time were spent on the instruction so that the students were able to have enough time to gradually improve in both articulation and naturalness and further internalize what they had learned, there might have been more variation in the performances of the examinees. Finally, though there were nearly 45 raters, over half of them were found to be overfitting or misfitting. If minimal rater training were possible before having them do the ratings, the results may have indicated less noise in their ratings.

Suggestions for Revision

Despite the limitations of this study, the findings from the analyses provide a great amount of information needed for finding the shortcomings of the test and suggesting ideas for improvement. Accordingly, three major issues were raised for revision in the test design and administration procedure: (a) elimination of item J-NT, (b) rater training, (c) and addition of detailed descriptors for each rating scale point. These will be discussed in order.

First of all, the J-NT item can be eliminated in the revised version of the test. This would make the test more reliable and valid by making all the other items fit the model perfectly and would save time for the raters who would have one less item to rate, as well as removing an item rater bias interaction.

Second, the great number of raters who were found to be misfitting or overfitting could benefit from rater training (McNamara, 1996). If they could practice and become more familiarized with the rating system while under guidance, they may be more self-consistent when it comes to rating on their own. To prevent overfitting raters showing central tendency and halo effect, the rater training session could be equally useful since raters could be guided into using the scale more thoroughly through the session.

Third, the rating scale could be improved by adding detailed descriptors to each point of the scale. Many of the raters thought that the students were forcing the target forms even though they could not use them naturally, and this was identified as one of the potential sources that might have caused rating confusion. Although the examinees scored higher in the posttest for all the items, the FACETS output showed alteration in relative difficulty among the items from the pretest to the posttest as was illustrated in the vertical rulers (Figure 1a and 1b). Recall that the naturalness items, K-NT and J-NT (both of which were relatively easier items in the pretest administration), became two of the toughest items in the posttest. In other words, the students' increased use of the

connected speech features might have negatively influenced their overall naturalness.

If students who are in the process of learning how to produce connected speech are facing the issues of quantity ‘and’ quality, these aspects should both be added in judging the student’s ability to produce connected speech features. Knowing how to do it would determine quantity, but using it naturally would determine the quality. Therefore, the descriptors explaining expected performance in terms of the two criteria for each point on the scale can be added to the rating scales so that they could assist the raters to make clear and unambiguous scoring decisions (Bond & Fox, 2007). See Table 20 for an example of the descriptors that could be used to cover these two criteria. Not only would using these descriptors reduce the confusion when rating and assist the raters to use the scale more thoroughly, it might ultimately resolve the problem of the less distinguishable point ‘3’ on the scale as well. Furthermore, this scale could be more useful for the teachers when determining where a student stands: whether the student needs more practice to produce the target form more naturally or whether the student does not yet know the mechanisms of producing it.

Table 20
Example of Descriptors for Each Scale Point

Scale	Descriptors
1	Not able to produce the forms at all
2	Can produce very few and does not sound natural
3	Can produce approximately half of them with some sounding natural and some not
4	Can produce many of them mostly sounding natural with few unnatural ones
5	Can produce it most of the time and sounds natural

Future Research Suggestions

As the goal of this study was to find the potential and underlying problems of the test and seek ways to revise it, the first suggestion for a future study would be to redo a FACETS analysis after collecting ratings from raters who have used the new version of the rating rubric and have gone through training. Although rater training might not close the gaps between the different severity logits of the raters, the self-consistency has been shown to improve (Lumley & McNamara, 1995; Weigle, 1998; Wigglesworth, 1993).

Another interesting study would be to examine relative importance of the variance components of various facet of the test by conducting a generalizability study (G study) which

enables the examination of multiple sources of error in the measurement procedure (Shavelson & Webb, 1991). Then using the variance components estimated in the G study, a decision study (D study) that allows the tester to figure out the most ideal conditions of the test, such as deciding the number of raters or the number of items, would be useful for creating a more feasible yet reliable design for both CRT and NRT purposes. A D study would be particularly useful to do after ensuring that all the facets are functioning well using FACETS analysis.

Finally, in addition to the conventional way of investigating language test validity in terms of construct, content, and criterion-referenced validity, looking at validity in a broader view as in Messick's approach (1988), where validity is also examined in terms of the adequacy of the inferences and decisions made upon the test scores and its actual use, would be worthwhile.

Summary and Implications

A new instrument for assessing connected speech performances was developed and evaluated. The reliability and validity of this test was examined using FACETS analysis, where a number of pieces of information suggested that the test is reliable and valid. In particular, the reliability index of the items showed that the items were spread out in terms of difficulty, and the gains of the examinees' performances as well as the fit statistics of the items supported the construct validity of the test. Although a single item was found to be misfitting, this problem was solved by eliminating another overfitting item. Also, examining unexpected performances of the examinees and raters through examinee and rater fit statistics and reviewing the rating scale diagnostics permitted deeper understanding of the drawbacks in the rating scale design and provided answers to how the scale descriptors might be worded. Finally, analyses of the raters' too-predictable or unpredictable performances suggested that there is a need for a rater training session where the potential raters could learn how to use the whole scale and use it consistently.

However, one concern remains that rater groups with different L1 backgrounds were found to have different rating patterns. This issue needs further investigation in future studies, since it raises a crucial question of whether non-native and native speakers are equally suitable when rating English pronunciation. This question could also be extended to the issue of using the test in different contexts, such as in EFL or ESL, as the students' scores might be dependant upon the L1 of the teacher using it.

Nevertheless, the first attempt to create and evaluate a testing instrument with an analytic scale for assessing connected speech production turned out to be quite successful in the sense that it suggests a new, reliable, and valid way of assessing connected speech and to some extent provides a specific model of a rating scale. This type of test could potentially serve as a feasible way of assessing connected speech performance in classrooms for diagnostic, achievement, or

feedback purposes. Although the current test and its dialogue only cover a small portion of all existing connected speech features in English, teachers who wish to teach their students how to use connected speech could always develop other dialogues involving different connected speech features (e.g., vowel to vowel linking, consonant deletion, and flapping), and use them as test prompts in their classrooms.

ACKNOWLEDGEMENTS

First and foremost, I am grateful for the support and guidance of my advisor, Dr. JD Brown, who has inspired my interest in the field of language testing, and who has made it so fascinating that all the “numbers” could not frighten me away. Specifically, he has continuously reminded me that ‘anomalies’ are not something to be concerned about, and that I should try to make the most of them by trying to understand what is going on. This word of encouragement has been the main driving force that has pushed me to complete this paper and gain confidence in what I had found. I would also like to thank my second reader, Dr. John Norris, who agreed to read this paper on short notice. Despite his very busy schedule, he read it with the greatest care and provided me with extremely insightful and stimulating feedback.

Also, I am very appreciative to my “mini-advisors,” Dennis Koyama, Lawrence Davis, and Munehiko Miyata, for their patience in answering all my questions through the process of completing this paper by sharing their amazing knowledge and expertise in writing, statistics, and academic experiences overall. Moreover, I cannot express enough thanks to my friends here in Hawai‘i who have always given me tremendous support mentally and have made my experiences here in the program and Hawai‘i exciting, extraordinary, and unforgettable.

Finally, and most of all, I would like to thank my family with all my heart for their unconditional support and love. Without their encouragement and consolations, my studies here in this program could not have been completed with such great satisfaction.

REFERENCE

Anderson-Hsieh, J. (1990). Teaching suprasegmentals to international teaching assistants using field-specific materials. *English for Specific Purposes, 9*, 195-214.

Anderson-Hsieh, J., Riney, T., & Koehler, K. (1994). Connected speech modifications in the English of Japanese ESL learners. *Issues and Developments in English and Applied Linguistics, 7*, 31-52.

Avery, P., Ehrlich, S., Mendelson-Burns, I. & Jull, D. (1987). Part 1: The sound systems of English. *TESL Talk, 17*, 14-72.

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

Bamgbose, A. (1998). Torn between the norms: Innovations in world Englishes. *World Englishes, 17*, 1-14.

Bond, T., & Fox, C. (2007). *Applying the Rasch Model*. New Jersey: Lawrence Erlbaum Associates.

Bowen, J. (1976). Current research on an integrative test of English Grammar. *RELC Journal, 7*, 30-37.

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

Brown, H. (2001). *Teaching by Principles: An interactive approach to language pedagogy* (2nd ed.), New York: Longman.

Brown, J. D. (2001). *Using Surveys in Language Programs.* New York: Cambridge University Press.

Brown, J. D. (2005). *Testing in Language Programs: A comprehensive guide to English language assessments.* New York: McGraw-Hill. Brown, J. D. (In progress). *Shaping students' pronunciation: Teaching the connected speech of North American English*. Honolulu, HI: University of Hawai'i.

Brown, J. D., & Hilferty, A. (1986a). Listening for reduced forms. *TESOL Quarterly, 20*, 759-763.

Brown, J. D., & Hilferty, A. (1986b). The effectiveness of teaching reduced forms for listening comprehension. *RELC Journal, 17*, 59-70.

Brown, J. D., & Hilferty, A. (2006). The effectiveness of teaching reduced forms for listening comprehension. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on Teaching Connected Speech to Second Language Speakers* (pp. 51-58). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Brown, J. D., & Kondo-Brown, K. (Eds.). (2006a). *Perspectives on teaching connected speech to second language speakers*. Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center.

Brown, J. D., & Kondo-Brown, K. (2006b). Testing reduced forms. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 247-264). Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center.

Caban H. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21,* 1-44. www.hawaii.edu/sls/uhwpesl/21(2)/Caban.pdf

Cahill, R. (2006) Teaching reduced interrogative forms to low-level EFL students in Japan. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 99-125). Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. New York: Cambridge University Press.

Dalton, C., & Seidlhofer, B. (1994). *Pronunciation.* Oxford: Oxford University Press.

Dauer, R. (1993). *Accurate English: A complete course in pronunciation*. Englewood Cliffs, NJ: Prentice Hall.

Dauer, R., & Browne, S. (1992). *Teaching the pronunciation of connected speech.* Paper presented at the 26th annual meeting of TESOL, Vancouver, BC (ERIC Document Reproduction Service No. ED354777).

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.

Flege, J., Yeni-Komshian, G., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41,* 78-104.

Gatbonton, E., Trofimovich, P., & Magid, M. (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly, 39,* 489-511.

Gilbert, J. (2005). *Clear speech: pronunciation and listening comprehension in North American English (3rd Ed.).* Cambridge: Cambridge University Press.

Hagen, S. (2000). *Sound advice: A basis for listening*. New York: Pearson Education.

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38,* 201-223.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing, 1,* 123-133.

Henrichsen, L. (1984). Sandhi-variation: A filter of input for learners of ESL. *Language Learning, 3,* 103-126.

Ito, Y. (2006a). The comprehension of English reduced forms by second language learners and its effect on input-intake process. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 67-81). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Ito, Y. (2006b). The significance of reduced forms in L2 pedagogy. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 17-25). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Joe, J. (2005). Korean college students' experiences in English camps in English speaking countries. *STEM Journal, 6,* 153-174.

Kim, D., & Margolis, D. (1999). Teaching English pronunciation to Koreans: Development of the English pronunciation test - EPT. *KOTESOL proceedings: The Second Pan Asian Conference*, pp. 89-112.

Koren, S. (1995). Foreign language pronunciation testing: A new approach. *System, 23,* 387-400.

Lee, B., & Jung, K. (2003). A study of English connected speech through middle school textbooks. *Foreign Language Education, 10,* 175-196.

Linacre, J. (2007a). *A User's Guide to FACETS.* [Computer software manual] Chicago: Winsteps.com.

Linacre, J. (2007b). Facets (Version 3.62.0). [Computer software] Chicago: Winsteps.com.

Linacre, J. (2008). *A user's guide to WINSTEPS: Rasch model computer programs.* [Computer software manual] Chicago: Winsteps.com.

Lumley, T., & McNamara, T. (1995). Reader characteristics and reader bias: Implications for training. *Language Testing, 12,* 54-71.

Matsuzawa, T. (2006). Comprehension of English reduced forms by Japanese business people and the effectiveness of instruction. In J. D. Brown, & K. Kondo-Brown, (Eds.), *Perspectives on teaching connected speech to second language speakers* (pp. 59-66). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

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

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun, (Eds.), *Test validity* (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Miller, S. (2006). *Targeting Pronunciation: Communicating clearly in English.* New York: Houghton Mifflin Company.

Ministry of Education (1998). *The 7th education curriculum manual.* Seoul: Daehan.

Munro, M., Flege, J., & Mackay, R. (1996).The effects of age of second language learning on the production of English vowels. *Applied Psycholinguistics, 17,* 313-334.

Norris, J. M. (2000). Purposeful language assessment. *English Teaching Forum, 38,* 18-23.

Norris, J. M. (2008). *Validity evaluation in language assessment.* New York: Peter Lang.

Oyama, S. (1976). A sensitive period in the acquisition of a non-native phonological system. *Journal of Psycholinguistic Research, 5,* 261-285.

Pennington, M. (1994). Recent research in L2 phonology: Implications for practice. In J. Morley (Ed.), *Pronunciation pedagogy and theory: New views, new directions* (pp. 92-108*).* Alexandria, VA: TESOL.

Pennington, M., & Richards, J. (1986). Pronunciation revisited. *TESOL Quarterly, 20,* 207-225.

Prator, C., & Robinett, B. (1985). *Manual of American English pronunciation.* Chicago: Holt, Rinehart, & Winston.

Rost, M., & Stratton, R. (1978). *Listening in the real world: Clues to English conversation.* Tucson, AZ: Lingual House.

Shavelson, R., & Webb, N. (1991). *MMSS generalizability theory: A primer*. Newbury Park, CA: SAGE Publications.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34,* 153-175.

Ur, P. (1996). *A course in language teaching: Practice and theory.* Cambridge: Cambridge University Press.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Weinstein, N. (2001). *Whaddaya say? Guided practice in relaxed speech* (2nd Ed.). New York: Longman.

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

Yoo, H. (2005). A study on the comprehension of phonological processes in English connected speech. *Foreign Languages Education, 12,* 273-293.

Yeo, M., & Park, G. (2006). The effective teaching of English pronunciation for raising communicative competence of primary school students. *English Education Research, 33,* 120-139.

Yoshida, H. (2004). An Analytic Instrument for Assessing EFL Pronunciation. *Dissertation Abstracts International, 65*(04), 1337A. (UMI No. 9315947)

APPENDIX A

Dialogue Script

Jay Hi, Kris! What are you doing?

Kris Jay? Shouldn’t you be home right now?

Jay I'm sleeping over at Debbie's house.

Kris Really? Did you get permission to go?

Jay Uh-huh. Of course I did. Well what were you doing?

Kris We were playing hide-and-seek, and I'm IT.

Jay Oh, so have you found anybody?

Kris No, not yet. I’ve got to find them after I finish counting.

Jay Well, do you need any help? I can help you.

Kris No, you can’t. You don’t know how to play this game!

Jay Yes, I do! I’m good at finding people. Really! Just give me a chance!

Kris I don’t need your help. Could you please leave me alone?
I'm in the middle of a game you know.

Jay Then can I just watch? I want to see you find everyone.

Kris No, please just go, Jay! Can’t you see I’m busy?

Jay Why won’t you let me stay?

Kris Jay! Would you please stop bothering me? And besides, aren’t you late?
Debbie will be waiting you know.

Jay Okay, okay! I don't like hide-and-seek anyways. I’m going to go now! Bye!

Kris Okay….98, 99, and 100. Ready or not, here I come!

APPENDIX B

Rater Survey Questions

1. What is your first language?

2. Your **English** teaching experience.
   Where have you taught for how long? And what did you teach?
   (e.g., ESL, University, 2 years, Listening and speaking)

3. Do you have English phonology or phonetics background?
   If yes, from where? (e.g., Course, book, teacher training)

4. Do you have pronunciation evaluation experience? If yes, what kind?

APPENDIX C

Rating Guideline Directions

Here are the directions for your rating:

1. You will listen to the tracks in numerical order on your CD.
2. Write the track number down on each evaluation form.
3. You will be rating separately for Kris and Jay on each track.
4. The following table (equivalent to Table 5 above and so excluded here) shows the list of target items what you will be rating the performance on.
5. Listen to the track twice; one time for rating Jay and one time for rating Kris.
6. You will give each a wholistic rating for four indicators (naturalness, weak form of 'you,' palatalization, and reduced forms) according to a 5-point scale from 1 = Very Poor to 5 = Very Good.
7. For the comments section, you may add comments for 1) explanations of particularly low or high ratings, 2) descriptions of any other particular feature of the performance, or 3) explanations of difficulty or easiness of making decision, and so on.
8. When evaluating students' performance, please focus on <u>pronunciation.</u> Students were directed to keep reading, even if they made mistakes. You will notice some student actually skipping some words or substituting different words, but please do not include these as errors when rating.
9. When listening to the CD, you can use a CD player or a computer, whichever you feel convenient. However, please use only ONE device throughout the evaluation so that using different kinds of equipment does not affect your rating consistency.
10. Since the sound quality is not good, it would be helpful for you to do this in a less distracted and quiet place.

APPENDIX D

Evaluation Form

JAY

Naturalness				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Weak form of "you"				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Palatalization				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Reduced forms				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Comments				

KRIS

Naturalness				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Weak form of "you"				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Palatalization				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Reduced forms				
1	**2**	**3**	**4**	**5**
Very Poor	Poor	Fair	Good	Very Good
Comments				