The Dissertation Committee for Lin Alison McGregor Certifies that this is the approved version of the following dissertation:

An Examination of Comprehensibility in a High Stakes Oral Proficiency Assessment for Prospective International Teaching Assistants

Committee:

Diane Schallert, Supervisor

Randy Diehl

Marilla Svinicki

Wanda Griffith

Keenan Pituch

An Examination of Comprehensibility in a High Stakes Oral Proficiency Assessment for Prospective International Teaching Assistants

by

Lin Alison McGregor, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy
The University of Texas at Austin
August 2007

Dedication

To my past, present, and future students

Acknowledgements

Many people have been instrumental in enabling me to complete this dissertation. The seed was planted in my mind as an undergraduate when Harriet Ottenheimer told me I had the insight to do a Ph.D. I am indebted to Harold Edwards for putting me on the path to become an accent modification trainer and pushing me to extend his methods.

I would like to thank my doctoral committee members for believing in and challenging me to excel academically. I am especially grateful to my esteemed advisor, Diane Schallert, for enduring and facilitating my learning curve. I give kudos to Wanda Griffith and Sherlock Campbell for their statistical assistance. I also thank Rob Donald for his indispensable technology guidance and encouragement.

I want to express my deepest gratitude to all my friends, but especially Ken Ingraham, Michelle Achacoso, Yoonjung Cho, Mary Knill, and Danelle Wright for their consistent support through all stages of this process and more. Finally, I would like to acknowledge my family for giving me roots and wings.

An Examination of Comprehensibility in a High Stakes Oral Proficiency Assessment for Prospective International Teaching Assistants

Publication No._____

Lin Alison McGregor, Ph.D.
The University of Texas at Austin, 2007

Supervisor: Diane Schallert

This study investigated the construct of comprehensible English in the context of oral proficiency assessment for international teaching assistants. I carried out a three-part mixed method design to explore instructor rater judgments, results of a speech analysis, and how specific speech variables might have influenced judgments on the assessment criteria. Each step focused on a failed/passed assessment comparison made possible through archived data from which 10 individuals initially failed the oral proficiency test but within the same year retook the task and received a passing score.

Part A evaluated the perspective of the instructor raters through the rating scale judgments provided on the assessment evaluation forms. In the second part of the study, I coded and scored grammatical, temporal, and phonological variables that occurred on two-minute excerpts of a field-specific summary task from the set of 10 failed and then subsequently passed assessments performed by the same individuals. I inspected the speech analysis results to evaluate differences in the values of specific speech variables on the set of failed performances in comparison to the set of passed performances. In Part C, I conducted 10 case studies to compare each individual's rating scale judgments and rater comments on grammar, fluency, and pronunciation from their failed and their passed assessment with the results from the speech analysis of grammatical, temporal, and phonological variables. The case study approach facilitated a broad inspection of the interrelation among the rater perspectives on the assessment criteria and the speech analysis results.

The study findings showed evidence of an interrelation between temporal and phonological variables on rater judgments of comprehensibility, as well as the role of pronunciation as a criterion for oral proficiency assessments. I concluded with implications for future research on the interrelation among speech variables that influence listener perceptions of comprehensibility and the use of pronunciation as a speaking assessment criterion.

Table of Contents

List of Tables

CHAPTER 1: INTRODUCTION

> “I think there should be a [higher] minimum English equivalency," said Russell Baird, a mechanical engineering senior, "I have nothing against international TAs, but when they can't speak well enough to get the point across it hinders their ability to teach."

In this dissertation, I investigated the construct of comprehensible English in the context of oral proficiency assessment for prospective international teaching assistants seeking to meet a state legislated English proficiency requirement to become eligible to assist with undergraduate courses. The study focused on a comparison between a set of failed and later passed oral proficiency assessments by the same 10 international graduate students at a Research 1 institute in the southwestern U.S. My aim was to explore what in the non-native English speaking performances might have triggered the overall score improvement from a failed to a passed oral proficiency screening result. Specifically, I conducted a detailed speech analysis on grammatical, temporal, and phonological variables produced in the set of failed and passed oral proficiency assessments. I hypothesized that closer approximation to native-like use in the speech variables had positively influenced the rater judgments on specific assessment criteria of pronunciation, grammar, fluency, and comprehensibility resulting in the failed to passed improvement on the overall test scores.

I approached the primary research question by using a three-part mixed method design to investigate (a) the scores indicated by the rating scale judgments on an

institutional oral proficiency test with pronunciation, grammar, fluency, and comprehensibility as assessment criteria, (b) the grammatical, temporal, and phonological characteristics as produced on a 2-minute segment of one of the assessment tasks, and (c) whether the speech variables analyzed showed any relation to the instructor rating scale judgments. In the remainder of this chapter, I provide an introduction to oral proficiency assessments for international teaching assistants. Following this background information, I describe in brief the relevant literature that situates this work in a larger context to provide the rationale for the problem addressed. Finally, I present the research questions that guided this three-part study.

The Development of Oral Proficiency Assessment for International Teaching Assistants

Institutions of higher education commonly require, and are sometimes mandated by state legislatures, to maintain standards on oral proficiency levels for international teaching assistants, often referred to as ITAs. Impetus for this legislation of oral proficiency for non-native English speaking teaching assistants started in the 1980s when enrollment of American students in graduate programs in the natural sciences and engineering fields decreased while the number of international graduate students increased. Due to the shortage of qualified American graduate students, institutions gave a large number of teaching assistant (TA) positions to non-native English speakers. Eventually, undergraduate complaints about the difficulty in understanding non-native teaching assistants' spoken English led to language proficiency requirements for

international teaching assistants (Anderson-Hsieh & Koehler, 1988; M. Bresnahan & Kim, 1993). The burden to prove language competency by passing the oral proficiency requirements ultimately falls on the international graduate students.

Oral proficiency screening in academic settings attempts to insure comprehensible language skills from non-native English speakers (Derwing, Rossiter, Munro, & Thomson, 2004; Hoekje & Linnel, 1994). The stakeholders for such oral proficiency assessments include the international graduate student, who may be in need of the financial support of an assistantship in order to pursue higher education in the United States (Myers, 1995); the program or institutions, who find themselves responsible for administering the oral proficiency assessment as well as ITA training courses; and, finally, the undergraduates, who will interact with ITAs. In fairness to all stakeholders, valid and reliable assessment is essential. The crux of the issue, however, resides in being able to measure the phenomenon of comprehensible, or "clear," English.

Currently, there are a reported 96,981 international scholars teaching and researching on U.S. campuses (ACTFL, 2007). In spite of the high demand for such assessments and over 25 years of research on oral proficiency assessment, no consensus exists on the appropriate instrument to evaluate spoken language proficiency for ITAs (2002). Popular standardized testing instruments in use include: Educational Testing Service's SPEAK (Spoken Proficiency English Assessment Kit), and TSE (Test of Spoken English). The OPI (Oral Proficiency Interview) is another. According to Saif

(2002), two decades of research has shown that these standardized tests "are not adequate for measuring the spoken language ability of ITAs" (p.146).

For decades, second language acquisition researchers and especially those focusing on language testing have investigated the nature of language proficiency components (Bachman, 1991; Bachman & Palmer, 1982). Language ability, at one time conceived as a single, global trait, is currently based on multi-componential models. These models include but are not limited to linguistic, sociolinguistic, discourse, and strategic knowledge (Bachman & Palmer, 1982; Skehan, 1998). Since the 1990s, language test design and development have been based on such multi-componential theoretical models of language proficiency. In addition, testing research has identified the influence of measurement factors on speaking assessments such as the test tasks, raters, scale criteria, and contextual variables, all seen as affecting measurement of oral proficiency (McNamara, Hill, & May, 2002). In brief, predictions of oral proficiency relate to how language users are able to cope with the demands of using actual language in real time as well as the incorporation of the test method and performance conditions of the test (Bachman, 1991; Luoma, 2004; Skehan, 1998). Consequently, speaking test results include construct-related variance as well as method-specific variance (Chalhoub-Deville, 1995).

Assessing oral language proficiency relies on defining communicative competence for the ITA context. Consequently, because of the debated range of language competencies involved in the various types of communicative acts within an

ITA position, construct validity in ITA oral proficiency assessment is controversial. ITAs need speaking skills for various duties including successfully teaching undergraduate classes, conducting lab sessions, and holding tutorial and office hours. Saif (2002) argued that the ITA language use context naturally requires a bilateral type of discourse used during interactions as ITAs interact with undergraduates. He explained that communication in this capacity "involves the speaker in both the production and comprehension of the spoken language, a complex activity which requires the employment of all aspects of communicative competence" (p. 147). Despite the interactive nature of ITA communication with undergraduates, common oral proficiency assessments include only monologic tasks (Wennerstrom, 2000).

In addition to the speaking skills needed to fulfill teaching assistant duties, there is currently still debate about whether or to what extent, pedagogical skills, field-specific knowledge, and cultural awareness should be included in the screening process. Because native speaking teaching assistants are not typically evaluated for teaching skills, some have argued that testing international graduate students for such skills is unfair (Saif, 2002). The issues revolving around oral proficiency screening for ITAs are resolved at the institutional level where administrators choose the testing instruments and possibly additional requirements, like mock teaching demonstrations, as part of the screening evaluation towards oral proficiency certification. The end result is the existence of a wide variety of ITA screening and training programs across the U.S. (Gorsuch, 2003).

Each program, however, is faced with the common goal of insuring oral proficiency in the context of non-native teaching assistants to work in undergraduate education. This real-world demand with state legislation requiring oral proficiency screening presents a practical need for the design and development of valid and reliable speaking tests for non-native English speakers. The practical need also stimulates theoretical questions related to better understanding the nature of oral proficiency components for non-native English speakers. This includes further investigation of competencies represented in oral proficiency models as well as how these components interact. An example is how a non-native English speaker's production of linguistic features impacts a native listener's perception of understanding. This study addressed both practical and theoretical level issues by comparing 10 failed and passed oral proficiency assessments to determine how they differed, including an exploration of how speech variables might explain the improved rater judgments on the assessment criteria of an oral proficiency assessment.

Supporting Literature for the Study Rationale

Oral Proficiency Assessment for ITAs

Oral proficiency assessment for international teaching assistants is a thorny area from all fronts, namely, the definition of the construct, the measurement, and the context. What is oral proficiency? How do we create valid and reliable speaking assessments? Which proficiencies (language, pedagogic, cultural) are needed to perform the duties of a teaching assistant? These fundamental questions are concerns for the fields of second

language acquisition (SLA), language testing, and teaching English to students of other languages (TESOL). The intersection of these fields only heightens the complexity of valid and reliable assessment of communicative competence on a variety of speech acts required in the ITA context. Although the field of ITA evaluation and training has existed for only the past three decades, more historically researched areas, specifically, fluency and pronunciation provide significant contributions to understanding the construct of oral proficiency from their much longer lines of research.

Fluency for Non-Native English Speakers

The premise behind oral proficiency assessment is to insure a level of language competence for ITA duties. The term *proficiency* in a second or foreign language is often synonymous with being fluent. Definitions of fluency include a broad sense of having a high level command of the language as well as a narrower sense referring to speed and smoothness of oral delivery (Lennon, 2000). Fluency might be described as flow, continuity, automaticity, or smoothness of speech (Riggenbach, 2000). In the narrow sense, fluency is commonly used as a criterion on speaking assessments representing a dimension or component of oral proficiency (Kormos & Denes, 2004). As an assessment criterion, fluency is often described in terms of timing. The linguistic predictors of fluency include: speech rate (speed), mean length of runs (amount of speech produced between pauses), disfluencies, and the number of filled pauses (uh, uhm, mmm) and unfilled pauses (silence) (Kormos, 2006). From an assessment raters' perspective, judging fluency involves giving the impression that "the psycholinguistic processes of

speech planning and speech production are functioning easily and efficiently" (Lennon, 1990). Considering American undergraduates as target listeners, fluency needs to reflect the non-native TA's ability to focus the students' attention on the message by presenting a finished product, rather than cumbersome production mechanisms as a speaker struggles to put together a sentence. One limitation of fluency research, however, is the lack of studies that investigate a combination of syntactic (grammar), temporal (timing), and phonological (sounds) variables, and the interaction among them (Kormos & Denes, 2004).

Pronunciation in English as a Second Language

The importance of various phonological aspects of English for second language learners has been the focus for the area of pronunciation. Since the 1960s, the intelligibility principle has guided pronunciation instruction in ESL (Levis, 2005; Munro & Derwing, 1999). The principle holds that learners are striving to be understandable and recognizes that communication can be remarkably successful even when foreign accents are noticeable or even strong. Levis (2005) argued there is no clear correlation between accent and comprehension. Instead, it is possible that certain features of speech seem to have a disproportionate role in impairing comprehensibility. Behind these statements are decades of debate among instructors and researchers concerning what aspects of the phonological system in English impact intelligibility and what features should be prioritized in instruction. Based on research and anecdotal evidence, there has been a call to shift the focus of pronunciation instruction from the segmental (isolated

sounds) level to a suprasegmental level (speech features applied across more than a segment) to have a greater impact on communicative competence (Anderson-Hsieh, Johnson, & Koehler, 1992; Jones, 1997; Pennington & Richards, 1986).

The term *suprasegmental* is often used interchangeably with *prosody* or when someone is referring collectively to the prosodic features of English. In the English as a Second Language literature, prosody is generally defined as intonation, stress, and rhythm of English (Anderson-Hsieh & Koehler, 1988; Derwing et al., 2004). These features, in English, function in combination to create the timing, phrasing, and highlighting of main information-bearing words (Derwing et al., 2004; Snow, 2001). In other words, prosody provides a framework directing the listeners' attention to information the speaker regards as important (Anderson-Hsieh et al., 1992; Derwing et al., 2004; Pickering, 2001; Wennerstrom, 2000). As Levis (2005) claimed, not all of the speech characteristics weigh the same in affecting comprehension. There is some evidence that, in rapid speech, prosodic deviance may more adversely affect comprehension than segmental deviance (Anderson-Hsieh & Koehler, 1988).

Differentiating Accent, Intelligibility, and Comprehensibility

Up to this point, several terms have been used without definitions or differentiation. In brief, the terms *foreign accent*, *intelligibility*, *comprehensibility*, and *fluency* represent related but distinct concepts that contribute to perceptions of oral proficiency. A *foreign accent* is defined as an impression of how much the speech of a non-native speaker deviates from a listener norm. In contrast, *intelligibility* refers to how

much a native listener can actually process from non-native speech or it can be conceived as how much of what is spoken can be written down. Being comprehensible in layman's terms refers to being understandable. *Comprehensibility*, as a dimension of oral proficiency or as an assessment criterion, means the overall impression of understanding a speaker. Munro and Derwing (1999) claimed to have empirically demonstrated that accent, intelligibility, and comprehensibility are related but distinct concepts. Derwing, Rossiter, Munro, and Thomson (2004) have added that listeners' perceptions of comprehensibility had a clearer tie to fluency judgments than their assessments of accentedness. This indicated that when a native listener perceived a non-native English speaker as being more fluent, the speaker was also rated higher on comprehensibility.

Although intelligibility and comprehensibility seem similar, the perception of what a listener understands (comprehensibility) is not always equivalent to what the listener has actually acoustically processed (intelligibility) (Derwing, Munro, & Wiebe, 1998; Munro & Derwing, 1999). In this sense, Munro and Derwing (1999) emphasized the human tendency to make perception-based judgments regarding reality. A native speaker might, for example, walk away from a conversation with a non-native speaker and quip to a friend, "I only understood about half of what that guy said." In a research setting, this speaker could be asked to write down what the non-native speaker actually said in the interaction. The Munro and Derwing definition implies that it was likely that the person would be able to write down much more than the 50% claimed to be perceived as understood, differentiating the degree of intelligibility from comprehensibility. The

distinction between intelligible versus comprehensible speech provides a subtle change in perspective on the nature of oral communication. Adopting a "comprehensibility" principle in assessment or instruction would be a theoretical paradigm shift in that it forces us to grapple not only with the phenomenon of meaning-making but also grounds it in our real-world contexts and adds the element of perception. In the next section, I give a brief description of the study context and of how the study was conducted.

Context of the Study

This study used results from the English Oral Proficiency Test (EOPT), the ITA screening instrument used at a large Research 1 university where the study was conducted. The assessment, modeled on the Educational Testing Service's standardized SPEAK test, is comprised of five tasks evaluated on a total of 10 different rating scales judgments. Two experienced ESL instructors rate each rating scale criteria from 0 to 3, the highest rating level. The rating scales represent the four assessment criteria evaluating pronunciation, grammar, fluency, and comprehensibility. In an attempt to fit the ITA context better, four of the five tasks use field-specific material based on the prospective ITA's major. The five test tasks include: summary of an article, pronunciation of terminology, explanation of terms, interpretation of a graph, and role-play (Appendix A). The summary of an article and pronunciation of terminology tasks are field-specific, that is, the summary topics and list of terms to be pronounced relate to each prospective ITA's major field of study.

Results of the EOPT are converted to a 300-point scale and fall into three categories: pass (above 250), conditional pass (230-245), and fail (below 225) (Appendix C). Students who receive a passing score are eligible to attend the ITA Teaching Workshop. Upon completing the workshop, the participants then qualify to apply for teaching assistantships. When students conditionally pass the EOPT, they have the option of retesting or taking a culture and communication course in order to improve their competence in speaking, pedagogy, and cultural competence. If students pass the course, no further testing is required, and the student qualifies to apply for teaching assistantships. Finally, those who fail to meet the minimum oral proficiency requirement are allowed to re-take the exam up to five times with three months between each assessment.

Research Questions

Part A: Investigation of Rater Judgments

For this study, I had access to the audio taped assessments together with the judge's rating sheets with their impressions of the students' performance for 10 prospective ITAs who had initially failed and subsequently passed the EOPT. Thus, the data I used in this study were part of an archive held by the office that administers the test. In Part A, I identified the change in rater judgments based on the rating scale score differences between the set of failed and passed assessments. To isolate the change in scores between the set of failed and passed assessments, I used a repeated measure multivariate analysis of variance (MANOVA) to analyze score differences on (a) the

assessment's five tasks, (b) the rating scale criteria of pronunciation, fluency, grammar and comprehensibility, and (c) the same four assessment rating scale criteria only on the field-specific summary task. In other words, because the EOPT was comprised of five tasks, each rated by one or more of the four possible assessment criteria, I wanted to determine whether score improvement on the passed assessment occurred on any one specific task. Similarly, I also wanted to analyze the rating scale criteria of pronunciation, fluency, grammar, and comprehensibility to test whether improvement in overall oral proficiency could be attributable to one or more of these competencies. Prior to conducting the speech analysis, I also wanted to investigate the summary of an article task to identify if the fail/pass score difference was due to performance on a specific area of skill competence represented in the rating scale criteria judgments. In Part A, I investigated three research questions using a repeated MANOVA method.

- Do overall scores on the set of failed and passed oral proficiency assessments by the same ten individuals differ statistically?
- Are there statistical differences on any of the assessments' five tasks between the set of failed and passed oral proficiency assessments for the same ten individuals?
- Are there statistical differences on any of the assessments' four rating scale criteria scores between the set of failed and passed oral proficiency assessments for the same ten individuals?
- Is there a statistical difference on the summary task criteria scores between the set of failed and passed oral proficiency assessments for the same ten individuals?

The results of the first investigation were used to establish how the set of failed and passed assessments scores differed by task and by rating scale criteria. Because one task, the summary task seemed particularly promising as the focus of a detailed speech analysis, I wanted to confirm whether the rating scale criteria scores on this task differed significantly between failed and passed assessments. It was 40% of the overall test score and was the only task on which raters provided assessments on all four of the rating scale criteria.

Part B: Speech Analysis

In Part B, I conducted a detailed speech analysis on a 2-minute excerpt from the summary task recordings. I hypothesized that the rating scale judgments were to some extent reactions to the oral production characteristics from the prospective ITAs performance. To this end, I expected the candidate's grammar errors would influence the rater's judgment of grammar, that an examinee's temporal measures of speech would impact the rater's perception of fluency, and that a speaker's phonological accuracy would influence the judgment of pronunciation. To test my hypothesis, I coded the grammatical, temporal, and phonological variables on the 20 speech samples taken from the summary task on both the failed and passed performances. Table 1.1 lists the speech analysis variables that I coded and scored. I also conducted this same speech analysis on speech samples from five native English speaking graduate students performing the field-specific summary task.

Table 1.1 Speech Analysis Variables

Variables	Explanations
Grammatical	*Grammar error counts*
	Categorization of errors
Temporal	*Timing*
Total Time	Total amount of speaking time
Speech Rate	Number of total syllables/total time
Articulation Rate	Number of total syllables/(total time-pause time)
Total Pause Time	Total amount of pauses > .33 seconds
Mean length of runs	Number of syllables between pauses > .33 seconds
Uhs, um, mmm	Counts of filled pauses
Repetitions	Number of repeated syllables
Phonological	*Sounds and prosodic features*
Segmentals	Error counts
Thought-groups/utterance	Number of "chunks" of languages in an utterance
Prominence/thought-group	number of emphasized syllables (stress)/thought-group

To obtain results from the speech analysis in Part B, I compiled descriptive statistics for evaluating whether the mean scores in grammatical, temporal, and phonological accuracy from the failed assessments differed from those on the passed oral proficiency assessment. I used dependent t-tests to evaluate the differences between the

set of failed assessments and the same individuals' results on the set of passed assessments for each variable. In Part B, I investigated the following research question.

- Do mean scores on the grammatical, temporal, and phonological variables increase when overall oral proficiency ratings increased from a failed to a passed level?

Part C: Multi-Case Studies

While conducting the first two investigations, it became apparent that a more detailed individual level analysis was needed to understand the factors influencing the change from a failed to a passed oral proficiency assessment score. Consequently, I turned to qualitative analysis and conducted 10 case studies to investigate how each individual's failed assessment differed from the passed assessment. Within each case, I could explore (a) the difference in the rating scale judgments and rater comments, (b) whether or not the speech analysis results could explain the change in rating scale judgments, and (c) how a holistic view of all the data at hand for each case could explain why each candidate's performance had been judged as failing but later judged as a passing performance. The specific research questions that guided the multi-case studies were the following:

- How do the rating scale judgments and comments differ between the failed and passed performance?
- How do each individual's failed and passed performances differ based on the results of the speech analysis variables, that is, grammatical, temporal, and phonological characteristics?

- Based on all data from each case, how can each individual's change in oral proficiency from a failed to a passed performance be explained?

Summary and Outline of Chapters

The goal of this study was to investigate the performance differences between a set of 10 failed and passed oral proficiency assessments of prospective international teaching assistants. I conducted a context specific investigation to examine (a) the change in failed and passed rater judgments, (b) the relation among speech analysis results and rating scale judgments, and (c) the interrelation among the rater judgments and comments, the assessment criteria, and the speech variables as these related to ratings of comprehensibility.

In Chapter 2, I review literature on oral proficiency assessment, international teaching assistants, and the language constructs of accent, intelligibility, comprehensibility, and fluency in relation to each other and their underlying speech factors. In Chapter 3, I describe the method employed in each phase of this study. I report the results of this three-part study in Chapter 4. Finally, I discuss the findings and limitations of the study as well as directions for future research in Chapter 5.

CHAPTER 2: LITERATURE REVIEW

"The sound of people's speech is meaningful, and that is why it is important for assessing speaking" (Luoma, 2004, p. 10).

A literature review on listener perceptions of comprehensibility within the context of oral proficiency screening for international teaching assistants requires examination in several areas. Chapter 2 has been organized into four main sections: oral proficiency assessment; the language constructs of accent, intelligibility, comprehensibility, and fluency; international teaching assistant studies; and synthesis and summary. The purpose of the first section on oral proficiency assessment is to highlight issues within assessment of speaking and to introduce Bachman's (1990) multi-componential model of language ability. In the beginning of the next section on language constructs, I start by defining the linguistic terminology. I summarize several key research studies delineating differences in the language constructs of accent, intelligibility, comprehensibility, and fluency as well as the underlying speech characteristics related to them. In the third section, I then review several international teaching assistant studies investigating language constructs and speech features relevant to the current study. Finally, in the last section, I summarize how the literature relates to and justifies this study.

Oral Proficiency Assessments

Language Ability and Models of Communicative Competence

Performance assessment within a second language context involves much complexity, largely because of the difficulty in defining what communicative ability in a second language is understood to be (McNamara, 1996). Luoma (2004) pointed out that if speaking tests are to be relevant to something outside the test, they need to be related to theoretical models of language ability. Therefore, understanding the nature of second language proficiency is fundamental to the design and development of valid and reliable language tests.

Historically, definitions of language performance have run a continuum from single, global trait models that have evolved into complex multi-componential models depicting the nature of communicative competence in a second language. According to Skehan (1998), Chomsky claimed that language performance was governed by an underlying rule-based grammatical knowledge system. Hymes added that appropriate language use was also an important aspect of communicative competence because it demonstrated the individual's understanding of social relations and how these relate to language use. Canale and Swain (1980) put forth a framework for communicative competence and extended previous models by including not only linguistic (derived from Chomsky) and sociolinguistic (Hyme's contribution) components but also discourse and strategic dimensions. Discourse was defined as the ability to handle language beyond the level of the sentence and to understand the rules of discourse, including how spoken and

written text are organized and how to make inferences that recover the underlying meaning of what has been said in the connection between utterances. In terms of strategic competence, Canale and Swain explained that when unable to cope with the other competences, compensatory strategies come into play, that is, "a range of devices may be drawn on to achieve the intended meaning or even to abandon the original meaning and resort to a simpler and more easily achieved goal" (Skehan, 1998).

More recently, Bachman (1990) put forth his Communicative Language Ability (CAL) model with three main components: language knowledge or competence, strategic competence, and psychophysiological mechanisms. Bachman's model consisted of both knowledge, or competence, and the capacity for implementing or executing that competence in appropriate, contextualized communicative language use (Bachman, 1990). By language knowledge or competence, Bachman (1991) meant the domain of information that is specific to language ability and that is stored in long-term memory. He used *knowledge* to refer to both conscious and tacit, analyzed and unanalyzed knowledge. The concept of "language knowledge" rejects the idea of isolated skills like reading, writing, listening, and speaking, because they are language use activities which all draw on components of knowledge identified within the communicative language ability model (Luoma, 2004).

In addition, Bachman (1990) explained that his theoretical model is simply a metaphor because "it captures certain features at the expense of others" (p.86). He also explicitly pointed out that these components, listed in Table 2.1, interact with each other

and with features in the language use context. To stay within the scope of the present study, I continue with only further discussion of grammatical knowledge, which relates most directly to the present study.

Table 2.1 Bachman's Areas of Language Competence

Type of Knowledge	Definition
Organizational knowledge	How utterances or sentences and text are organized
Grammatical knowledge	How individual utterances or sentences are organized • Knowledge of vocabulary • Knowledge of syntax • Knowledge of phonology/graphology
Textual knowledge	How utterances or sentences are organized to form texts • Knowledge of cohesion • Knowledge of rhetorical or conversational organization
Pragmatic knowledge	How utterances or sentences and texts are related to the communicative goals of the language users and to the features of the language-use setting
Functional knowledge	How utterances or sentences and texts are related to the communicative goals of language users • Knowledge of ideational functions • Knowledge of manipulative functions • Knowledge of heuristic functions • Knowledge of imaginative functions

Type of Knowledge	Definition
Sociolinguistic knowledge	How utterances or sentences and texts are related to the features of the language-use setting • Knowledge of dialects/varieties • Knowledge of registers • Knowledge of natural idiomatic expressions • Knowledge of cultural references and figures of speech

Bachman (1990) explained that grammatical competence consists of a number of relatively independent competencies such as the knowledge of vocabulary, morphology, syntax, and phonology/graphology. He demonstrated that the choice of words to express signification, their forms, their arrangement in utterances to express propositions, and their physical realizations, either as sounds or as written symbols through an example.

> Suppose, for example, a test taker is shown a picture of two people, a boy and a taller girl, and is asked to describe it. In so doing, the test taker demonstrates her lexical competence by choosing words with appropriate significations (boy, girl, tall) to refer to the contents of the picture. She demonstrates her knowledge of morphology by affix in the inflectional morpheme (-er) to 'tall.' She demonstrates her knowledge of syntactic rules by putting the words in the proper order, to compose the sentence 'The girl is taller than the boy.' When produced using the phonological rules of English, the resulting utterance is a linguistically accurate representation of the information in the picture (Bachman, 1990).

The psychophysiological mechanism or modalities of performance can be applied to the above example to show this component of Bachman's model. For example, the candidate used linguistic competence to form the sentence, "the girl is taller than the boy." She used visual skill to view the picture and auditory skills to obtain the administrator's directions for the task. Finally, she used articulatory skill in pronouncing the words correctly and providing the appropriate stress and intonation (Bachman, 1990). This demonstrates the incorporation of performance modalities into Bachman's model of communicative language ability.

One of the limitations to componential models of communicative competence is the static view of communication and emphasis on language characteristics (Luoma, 2004). Skehan (1998) critiqued Bachman's framework by pointing out that it "lacks a rationale grounded in pscyholinguistic mechanisms and processes (and research findings) which can enable such a model to move beyond 'checklist' status and instead make functional statements about the nature of performance and the way it is grounded in competence" (p. 164). In addition, Chalhoub-Deville (1995) stated that although the multi-componential nature of second language proficiency may be generally agreed upon, the nature of the components continue to be debated. Specifically, she highlighted the need for research to specify the attributes of the second language oral construct and to explain factors such as task and rater that might confound the interpretation of proficiency.

Factors in Oral Proficiency Assessments

Assessing second language oral proficiency is one of the most difficult skills to assess reliably because of the complex array of factors influencing the performance (Luoma, 2004). Basically, a performance-based assessment involves the candidate performing elicitation tasks from the instrument, being observed, and then judged using an agreed upon set of rating scale criteria (McNamara, 1996). Speaking test performances include: the task (the vehicle for performance), the raters (who observe the performance) equipped with scales and criteria to make judgments about it, and contextual factors. The raters and elicitation tasks are principal factors influencing the assessment of learner's second language oral proficiency(Chalhoub-Deville, 1995)

The rater factor in oral proficiency assessment. Second language speaking performances necessarily involve subjective judgments. Performance-based assessments then naturally involve a performance or behavior being judged or rated, by means of a scale or other kind of scoring device. This introduces a new type of interaction between the rater and the scale that mediates the interaction between scoring and performance. Although ratings are an integral part of proficiency, the scores may reflect, in addition to the learner's ability, raters' assessment schemes or interpretations of the assessment criteria (Chalhoub-Deville, 1995).

According to McNamara (1996), raters may differ in the following ways:

1. Raters may differ in their overall leniency.

2. Raters may display specific patterns of harshness or leniency in relation to a particular group or candidates or in relation to tasks or certain aspect(s) of language.
3. Raters may differ in their interpretation of the rating scale being used.
4. Raters may differ in terms of their consistency or inconsistencies; that is the extent of the random error association with their ratings. (p. 123-124)

In addition, rater judgment issues also involve the difference between using experienced versus inexperienced raters. In many testing situations, ESL instructors or teachers administer oral proficiency assessments as part of their regular employment duties. However, because of continued exposure to learners' non-native English speaking, the ESL instructor serving as a rater may adopt certain criteria that differ from non-teacher raters. Chalhoub-Deville (1995) pointed out that second language teachers have been found to focus more on grammatical and pronunciation errors as well as to match their evaluations of second language learners' oral ability to the models with which they had been trained.

The criterion factor in oral proficiency assessment. Assessment criteria make implicit reference to competencies that then emerge as the object of measurement. Rating scale criteria with clearly defined levels of achievement can enhance reliability of judge-mediated ratings by limiting raters' tendencies to subconsciously bias scores. McNamara (1996) admitted, however, that sometimes a rater's orientation to relevant language features can be quite deep-seated, and involve substantial reinterpretation of the

criteria presented to them by test developers. In other words, rating scores constitute a threat to construct validity in that they do not necessarily represent the intended construct because of how raters use the criteria despite training and simple descriptions. (1996b) argued that language performance tests are "relatively strong or weak according to whether the assessment criteria reflect real-world evaluative criteria, or focus more narrowly on dimensions of language performance. In either case, the role of non-linguistic and interactional factors, either acknowledged as explicitly relevant, or assumed (wrongly) to be irrelevant, needs to be investigated in the empirical validation of inferences from test scores" (p. 45). Assessments might depend not only upon which particular features of speech (e.g., pronunciation, accuracy, fluency) the rater pays attention to at any point in time, but also upon a host of other factors such as language level, gender, and status of rater, his or her familiarity to the candidate and other personal characteristics of the rater and candidate (Luoma, 2004).

Language Constructs

Definition of Terms

One line of English as a second language research has investigated language constructs by evaluating the linguistic features that influence accent, intelligibility, comprehensibility, and fluency. Prior to reviewing these research studies, I reviewed the major linguistic features relevant to these investigations. In particular, I explain in the next section the difference between segmental and suprasegmentals as well as describe the features of prosody.

In the area of pronunciation in English as a second language, the terms *segmental* and *suprasegmentals* dominate an ongoing debate about which linguistic aspect of English positively impacts communication and therefore should be prioritized in instruction. A *segment* can be defined as an isolated sound. More specifically, *phonemes* in a language comprise a limited number of classes of sounds of a particular language that are used contrastively to distinguish words of that language (Asher & Simpson, 1994).

Suprasegmental refers generally to the analysis of phonological features above the segmental level (Fox, 2000). Suprasegmental features are elements of the sound system that can be superimposed above the segment level on a string of segments. These phonological features include pitch, loudness, voice quality, and length. Most of these features are expressed on the laryngeal level through the rate of the vocal fold vibration, amplitude of vocal fold vibration, manner of phonation, timing, and presence or absence of velic closure (Snow, 2001).

Suprasegmental is often used interchangeably with *prosody* or when someone is referring collectively to prosodic features. Prosody in the English as a second language literature are generally defined as intonation, stress, and rhythm (Anderson-Hsieh & Koehler, 1988; Derwing et al., 2004). Any earnest attempt at defining prosodic features, however, quickly becomes a terminological minefield. According to Fox (2000), the description and definition of these features have always been something of a problem for linguists and have been relatively neglected especially in the last 30 years. There is no

universal consensus among phonologists about either the nature of prosodic features themselves or the general framework for their description. Fox defined prosodic features as length, accent and stress, tone, intonation, and potentially a few others. Cruttenden (1986) included only pitch, length, and loudness. Below I describe intonation, stress and rhythm.

Intonation is a multifaceted phenomenon (Brazil, 1997) that conveys meaning to a phrase or utterance, thereby setting it apart from other prosodic and non-prosodic features (Fox, 2000; Ladd, 1996). The function of prosodic features is that they combine to create the timing of a language, phrasing, and highlighting of main information-bearing words (Snow, 2001). In other words, prosody provides a framework directing the listener's attention to information the speaker regards as important (Anderson-Hsieh et al., 1992). Falling intonation, for example, could assign the meaning of a statement or completion of utterance while rising might indicate a question or an incomplete message. In addition, "intonation may indicate a discourse meaning like inviting a listener to make a contribution to the conversation, or an attitudinal meaning like being condescending" (Cruttenden, 1986).

Intonation is a pattern of pitch movement across a phrase or utterance. Acoustically, pitch is measured in fundamental frequency variation over one or more successive syllables (Cruttenden, 1986). Intonation in English uses a pitch range of 60-240 hertz for adult men and 180-400 hertz for women (Reed, 2006). Many terms have been put forth to identify the primary unit on which intonation is applied, or the spoken

"chunks" of language. These discrete spoken units have been called tone groups (Ladd, 1996), intonational phrases, tone units (Brazil, 1997), intonation units (Tench, 1996), sense groups, breath groups, or thought-groups. The variation in terminology may be due to the unresolved issue of whether these units are determined semantically, physiologically, psycho-acoustically, or based on other reasons (Brazil, 1997). Despite the disagreement over the terms, it is agreed upon that pauses indicate the boundaries of intonation units (Cruttenden, 1986). Thought-groups in native speaker speech tend to range from one to seven words. In this study, I adopted the use of the term *thought-groups*.

In thought-groups, syllables become prominent by stress created through pitch, length, loudness, and/or intensity. The stress pattern of an utterance is generated by considerations of the information structure (Pierrehumbert & Hirshberg, 1990). The relative stress on each syllable functions at two levels: it highlights the syllables of important words in utterances as well as generating the pattern of stressed and unstressed syllables that defines the rhythm in English. English, like other Germanic languages, is stressed timed. That is, the rhythm is created by the combination of strong or weak stress put on each syllable. In contrast, there are syllable-timed languages like Japanese in which each syllable has equal contribution to the rhythm.

The terms used to refer to the accentual phenomena include but are not limited to: *accent, accentuation, stress, prominence, emphasis, salience, intensity,* and *force*. These terms are used in different kinds of accent, that is, syllable accent, word-accent, sentence

accent, in contrast to different kinds of stress namely word-stress and sentence stress (Fox, 2000). For the purposes of this study, I adopted stress as the term to refer to the prominence of syllables generated through pitch, length, and loudness (Snow, 2001). Perceptual experiments have clearly shown that the features of pitch, length, loudness in English form a scale of importance in bringing syllables into prominence. Pitch has been found to be the most efficacious, and loudness the least so (Cruttenden, 1986). Simply, stress on a syllable is created through pitch, as defined earlier, the fundamental frequency level. Length of a syllable also generates stress through the relative duration of a given syllable or a number of successive syllables. The loudness of a syllable or the relative loudness of a number of successive syllables also creates stress. Finally, rhythm, the third component, describes the combination of the stressed and unstressed syllables in English which generate a specific rhythmic quality.

There have been two traditions in the study of prosody, one more phonetic in contrast to the other more phonological. The former aimed at quantifying the acoustic features of prosody while the other aimed at describing prosody in relation to grammar and phonology. The phonetic approach defines prosody in concrete terms and views a direct correlation between function and form, that is, the specific message with specific acoustic parameters. Experimental psychologist and phoneticians have used acoustic instruments in research to make this specific connection. The more abstract approach investigated any phenomenon that involved phonological organization at the level above the segment (Cutler & Ladd, 1983). Linguist and language teachers have been more

interested in describing how prosodic features function in different languages than in their acoustic parameters.

Foreign Accent, Intelligibility, Comprehensibility, and Fluency

Why do foreign accents exist and what factors influence them? A foreign accent or "accentedness" is defined as the extent to which non-native speech is perceived to differ from a native speaker (NS) norm (Flege, Munro, & Mackay, 1995; Munro & Derwing, 1998). The phenomenon of a foreign accent has been investigated for over 30 years (Piske, Mackay, & Flege, 2001). This research has primarily focused on two major themes: (1) explanations of why accents exist in a second language (L2) including factors that influence the degree of "accentedness," and (2) examinations of foreign accent in relation to other concepts such as intelligibility, comprehensibility, and fluency (Derwing et al., 1998; Derwing et al., 2004; Munro & Derwing, 1998, 1999, 2001).

One of the most noticeable differences between first language acquisition (L1) and second language acquisition (L2) is the degree of foreign accent often incurred. In the late 1960's, the seemingly poor ability of adults to master a second language phonological system was attributed to neurophysiologic maturation or the critical period hypothesis. A critical period, first used by ethologists to explain the origins of animal behavior, was applied to language acquisition by the American psycholinguist, Eric Lenneberg (Crystal, 1997). From this theoretical influence, accented speech was claimed to be due to cerebral lateralization, a change to the central nervous system occurring

around puberty indicating that adult learners simply had missed the acquisition "window" for sounds (Flege, 1981; Gass & Selinker, 2001).

Viewing the critical hypothesis period as an insufficient explanation for adult accents, Flege (1981) proposed the phonological translation hypothesis, where mature speakers had a tendency to interpret sounds in a foreign language in terms of their native language sound system. The cause of a foreign accent was attributed to the language learner mapping of new foreign language sounds onto the first language sound system. Flege recognized additional phonological influence beyond just sounds by suggesting that listeners were "likely to base a judgment of foreign accent on some combination of segmental, subsegmental, and suprasegmental differences which distinguish the speech of a native from that of a non-native speaker" (p. 445). The majority of research on foreign accent, however, continued to focus primarily on age-related constraints and segmental deviation to explain accented speech of non-native language speakers. Similar to the earlier phonological translation hypothesis, Flege (1995) later put forth the more developed second language Speech Learning Model to explain segmental differences in foreign accented speech (Birdsong, 1999).

Piske, Mackay, and Flege (2001) reported a review of 30 years of research on the factors affecting the degree of foreign accents in L2. Age of learning, length of residence, gender, formal instruction, motivation, language learning aptitude, language use, and amount of L1 use were the factors identified as affecting the degree of a foreign accent. Age of learning was claimed to be the strongest predictor of a foreign accent.

The authors pointed out, however, the great variation found in the designs and methodologies of the studies reviewed. Briefly, these studies reviewed differed in subject populations, elicitation techniques, rating techniques, and factors considered to impact the existence of an accent. Next, I turn to the literature that began to compare accent with other language concepts.

Distinguishing between accent, intelligibility, and comprehensibility. Munro and Derwing (1999) investigated the interrelation among accent, intelligibility, and listeners' perceptions of comprehensibility. They used descriptive narrations of humorous cartoons to generate speech samples from 10 native Mandarin speakers with advanced level English. Then, they had 18 native English speaking college students rate the speech samples on comprehensibility and accent, using a 9 point rating scales. Intelligibility was measured by asking the listeners to transcribe orthographically the speech samples. In addition, intonation was rated on a 9 point rating scale from native-like to not at all native-like. Two native speaker participants were included in the study as a baseline for the speech analyses.

The complete set of 36 utterances were phonetically transcribed by the researchers and coded for phonemic, phonetic, and grammatical errors. Phonemic errors were defined as any deletion, insertion, or substitution of a segment different from a clearly interpretable English phoneme. Phonetic errors involved the production of segments that sounded noticeably nonnative. The intonation of each speech sample was rated on a 0 to 9 rating scale where 1 = native-like and 9 = not at all native-like.

Munro and Derwing (1999) found a significant negative correlation between perceived comprehensibility and the transcription intelligibility scores representing accent. The accent ratings showed even weaker reflections of intelligibility represented in the transcriptions than judgments of comprehensibility. Mean distributions between perceived comprehensibility and accent scores were noticeably different. Listeners tended to assign harsher scores when rating accent. It was suggested that the extra processing time required in understanding accented utterances might explain the low comprehensibility scores. However, the accent scores were a much poorer reflection of the listeners' actual understanding of an utterance than the perceived comprehensibility scores.

The most important finding was that some utterances rated as moderately or heavily accented were perfectly transcribed in evaluating intelligibility. Phonemic, phonetic, and grammar errors as well as intonation were significantly related to ratings of accent by 70% of the listener. Yet, a surprising lack of correlation between the phonemic, phonetic, grammatical errors, and intonation ratings suggested that these categories were independent of one another.

To summarize, Munro and Derwing (1999) made several important claims based on their results: (1) foreign accent, intelligibility, and perceived comprehensibility were empirically demonstrated to be related but unique concepts; (2) the use of a 9-point rating scale allowed more appropriate comparison than binary judgments of accent and comprehensibility; and (3) extemporaneous speech samples reflected more naturally

occurring speech than a reading task. The following limitations were also pointed out. The study was based on 10 native Mandarin speakers of English and 18 Canadian college students being paid $10 to complete the experiment. The influence of sample size set statistical limitations on the types of analyses that could be conducted as well as amount of statistical power possible. In spite of the limitations, the study illuminated how accent, intelligibility, and comprehensibility could impact communication differently.

The effect of speech rate on accent and comprehensibility. I turn now to three studies that focused on one primary phonological feature, namely speech rate, and its impact on foreign accent and comprehensibility. In Table 2.2 the participants, judges, type of speech sample, and measurements are summarized for comparison. The purpose of this section is to review findings on speech rate.

In the first study by Anderson-Hsieh and Koehler (1988), speech rates of slow, regular, and fast were compared to scores on listening comprehension, accent, and pronunciation, specifically segmentals, syllable structure, and prosody (stress, rhythm, and intonation). The speech samples came from three non-native and one native speaker on reading passages with 310 to 475 syllables. The three categories of speech rate (slow, regular, and fast) ranged from 2.4-4.5 syllables per second, a somewhat slower range typically found in native speaker studies. After listening to the recordings, 224 college students answered listening comprehension questions and rated the speakers' accent and speech rate. A 7-point scale was later used by one of the authors to impressionistically

rate the speech samples on segmental, syllable structure, and prosody, defined as stress, rhythm, and intonation.

Results indicated that accent ratings generally matched the non-native speakers' Test of Spoken English levels. The accent ratings also remained fairly constant across the three speech rates. Finally, Anderson-Hsieh and Koehler (1998) found that higher ratings on sound segments, syllable structure, and prosody generally correlated with higher listening comprehension scores. However, later Anderson-Hsieh and Koehler (1992) claimed that "…there is some indication that not all of the speech characteristics weigh the same in affecting comprehension. There is some evidence that prosodic deviance may more adversely affect comprehension at the fast rate than segmental deviance does." (p. 585)

Table 2.2 Research Studies Investigating Speech Rate

	Studies		
	Anderson-Hsieh & Koehler (1988)	Munro & Derwing (1998)	Munro & Derwing (2001)
Participants			
Number	n=3 non-native/ 1 native	n=10 non-native/ 10 native	n=48 non-native/ 2 native
First language	Mandarin/ Native English	Mandarin/ Native English	12 different languages/ Native English
Judges	224 American college students	20 Canadian college students	48 Canadian college students

	Studies		
	Anderson-Hsieh & Koehler (1988)	Munro & Derwing (1998)	Munro & Derwing (2001)
Speech Sample	Reading narrative	Reading narrative	Reading sentences
Measures			
Factor	Listening comprehension	Comprehensibility	Comprehensibility
Measurement	6 test questions over passage	9pt Likert scale	9pt Likert
Factor	Accent	Accent	Accent
Measurement	5pt Likert scale	9 pt Likert scale	9pt Likert scale
Factor	Speaking rate	Speaking rate	
Measurement	5pt Likert scale	Number syllables/total time	
Factor	Segments, syllables structure, prosody (stress, rhythm, intonation)		
Measurement	7pt Likert scale		

The aim of the study by Munro and Derwing (1998) was to investigate how speech rate influences accentedness, comprehensibility, and listener preference. The rationale behind the study was to identify accentedness as fundamentally a perceptual phenomenon because listener judgments of accented speech relate to: age of L2 learning,

segmental error frequency, prosodic goodness, and acoustic characteristics. They argued, "it is necessary that perceptual research on accented speech controls sociolinguistic variables as much as possible, in order to focus on phenomenon affecting listeners' processing of L2 production" (p.160). To incorporate these controls Munro and Derwing used a reading passage to keep the sample free of grammatical errors, restricted the number of first languages to eliminate prejudice as a factor, and used a one-way listening task.

In the experiment, 10 native Mandarin speakers of English were asked to read a passage twice, once at a normal rate were a second instructed to read at half as fast as normal. The 40 speech samples created from the readings and were played for 20 Canadian college students. The samples were rated on 1-9 point scales for accent and comprehensibility. Munro and Derwing hypothesized that slow second language speech would be judged by native listeners as less accented and easier to understand. It was assumed that better execution of articulations from the non-native speakers would lead to more accurate production of phonetic targets and more time for listeners to process what was being said. However, the results indicated that these Mandarin speakers were generally judged to be less comprehensible and more accented when they slowed down.

In a later study, Munro and Derwing (2001) hypothesized a curvilinear relationship between speech rate and listeners' perceptions of accent and comprehensibility. Here, 48 adult intermediate ESL students read a list of standard statements for the recorded speech samples and 48 Canadian college students rated accent

and comprehensibility on four sentences taken from each of the 48 non-native speakers' recordings. Interrater reliability for accent ratings was .70 and .72 for comprehensibility. Two regression equations were calculated, one with comprehensibility as the dependent variable and the other with accent as the independent variables. Regression on accent accounted for R-squared = 15% for rate on accent ratings and an even smaller amount of variance 7%, for rate on comprehensibility ratings. They concluded that speech rate was indeed related to accentedness and comprehensibility but to a relatively small degree.

Munro and Derwing (2001) did draw attention to the fact that this study was limited in only being suggestive of correlational relationship. In other words, the independent contribution of rate to listeners' evaluations of accentedness and comprehensibility could not be determined. "It is possible, for instance, that the speakers in the study who spoke slowly also made the largest numbers of phonological errors." Therefore, "it would be useful to examine more closely the importance of rate in comparison with other variables, such as segmental accuracy" (p. 460).

Fluency on different tasks. Although accent, intelligibility, and comprehensibility are useful concepts, it is not uncommon in language pedagogy and language testing to focus on levels of oral proficiency often termed fluency. There seems to be no consensus on the definition of fluency or how is should be measured (Kormos & Denes, 2004). One definition defined native-like fluency as having the ability to produce fluent stretches of discourse like a native speaker. In a broad sense, fluency seems to be indicative of global oral proficiency or having a high command of a foreign language. Lennon (1990)

defined fluency as “an impression on the listeners that the psycholinguistic process of speech planning and speech production are functioning easily and efficiently” (p. 391). In a narrower sense, fluency can be considered a component of oral proficiency, which is often used in assessing candidates' oral language skills in language testing (Kormos & Denes, 2004) Generally, fluency definitions seem to contrast two aspects, those that consider fluency to be a temporal phenomenon and those that regard fluency as competence in oral proficiency. Fluency research, however, lacks investigations that combine linguistic, temporal, phonological, and interactional variables (Kormos, 2006).

Derwing, Rossiter, Munro, and Thomson (2004) conducted a study in which 20 high-beginner native Mandarin speaking ESL students were asked to perform three speaking tasks: a narrative, monologue, and conversation. These tasks were then rated by 28 college students for fluency, comprehensibility, and accent. In addition, the evaluated prosody (intonation and rhythm) and temporal measures including pauses, speech rate, self-repetition, mean length of run, and pruned syllables. This last measure was a count of the total number of syllables and subtracted self-corrections, self-repetitions, false starts, non-lexical filled pauses, and asides, similar to “disfluency” category in Kormos and Denes (2004). Also, relevant for the current study, comprehensibility and fluency had between .20 and .30 higher correlations than fluency and accent on the three tasks. Through multiple-regression analysis, pruned syllables were reported to have accounted for 65% of the fluency.

Most studies on fluency have concluded that "the best predictors are speech rate, that is, the number of syllables articulated per minute, mean length of runs, that is, the average number of syllables produced in utterances between pauses of .25 seconds and above" (Kormos & Denes, 2004). Phonation time, the ratio of time spent speaking divided by time taken to produce the speech sample, has also been found to be a good predictor of fluency.

Investigations based on instruction. With a hypothesis on the value of teaching stress, intonation, and rhythm, Derwing, Munro, and Wiebe (1998) compared three types of pronunciation instruction to evaluate the influence on accent, comprehensibility, and fluency. Intermediate level ESL learners in Canada were divided into three instructional groups: no instruction, segmental, and global. Group one had no specific pronunciation instruction and served as an experimental control. The segmental group practiced the elicitation of individual sounds and syllables. Finally, in the global approach group, stress, intonation, and rhythm were the focus of instruction. In study 1, 48 native speakers evaluated read sentences on comprehension and accent using 9-point rating scales. In a second study, six ESL instructors rated 45-second excerpts of spontaneous speech from picture narrations on comprehension, accent, and fluency. From the results, the researchers recommended both global and segmental benefits for ESL students. Derwing et al. (1998) explained that

> in the case of a communication breakdown caused by mispronunciation, a student who has received segmental training might be able to focus on the mispronounced

> form in a self-repetition. On the other hand, global instruction seems to provide the learner with skills that can be applied in extemporaneous speech production despite the need to allocate attention to several speech components (p. 407).

Interestingly, when Derwing and Rossiter (2002) compared ESL learners' perceptions of their pronunciation needs and their strategies in communication breakdown, a disconnect emerged. While one third of the 100 adult immigrants responded they often or very often had trouble being understood because of their accent, paraphrasing and self-repetition were the most common strategies used to resolve the communicative breakdowns. The researchers pointed out that if prosodic features are indeed primary to being understood, learners need to be taught communicative strategies involving those elements of speech.

Gorsuch (2001) reported that after 32 hours of production-focused instruction, students' perception of suprasegmentals seemed to improve although their production did not. Reasons for these results included the limitation of a once a week class resulting in insufficient time to develop real changes. Limited room for improvement on the tasks as well as unfamiliar tasks in the assessment as opposed to practice activities were also brought into question as possible confounds on the results. However, Gorsuch pointed out the need for clarification in the relationship between speech perception and production. In other words, do learners need to be able to perceive the target phonological feature prior to producing it or will production practice improve the learner's ability to perceive target segmental or suprasegmental? The cause and effect relationship between speech perception and speech production is debated but the

evidence that speech perception does play a significant role in second language phonological acquisition is clear.

In summary, the terms *foreign accent, intelligibility, comprehensibility*, and *fluency* represent related but distinct concepts. A foreign accent describes speech that deviates from a norm, but it is not necessarily equivalent to intelligibility, how much of an utterance a listener can successfully process. Likewise, intelligibility and comprehensibility have been shown to be distinct because the perception of what a listener understands is not always equivalent to what has actually been acoustically processed (Derwing et al., 1998; Munro & Derwing, 1999). It should also be pointed out that fluency is based on temporal phenomena and language competence (Kormos & Denes, 2004). In a recent study, Derwing, Munro, and Thomson (2004) indicated that listeners' perceptions of comprehensibility had a clearer tie to fluency judgments than did their assessments of accentedness.

Research on International Teaching Assistants

In the next section, I review research on speaking assessments and whether SPEAK is a good predictor of oral proficiency. I introduce several factors found to impacting undergraduates' receptivity of international teaching assistants. Finally, I return to language constructs research of intelligibility and fluency conducted within the international teaching assistant context.

Educational Testing Services' SPEAK Test

One of the most well known screening instruments for prospective international teaching assistants is the Speaking Proficiency English Assessment Kit (SPEAK) designed by the Education Testing Service (ETS), the nonprofit testing agency (who is also responsible for the GRE, SAT, TOEFL, and TOEIC). SPEAK is a standardize assessment of oral communication skills designed by ETS but administered locally by universities and other agencies. The ETS administered equivalent to SPEAK is called the Test of Spoken English (TSE). Academic institutions, corporations, government agencies, health care systems, and other organizations use TSE scores to guide their decisions regarding graduate assistantships in teaching and research, employment, and licensing and certification (ETS, 2005)

In both the TSE and SPEAK, the test lasts about 20 minutes, the questions are given by a recorded interviewer, the candidate is asked to describe a 6-panel picture story as well as a graph. On the SPEAK, typical tasks are to demonstrate the ability to give directions from a map or verbally present to a group changes to a schedule. On the OEPA, the international graduate student is asked to role-play a teaching assistant making an announcement regarding course related information.

Is SPEAK a good predictor of acceptable classroom performance for non-native English speaking teaching assistants? This was the focus of a study done at Iowa State University in 1986. Mathematics, Physics, and Chemistry departments were surveyed to determine if the university screening process for international TAs was being successful.

The survey results indicated that they had very few complaints about the speaking ability of non-native teaching assistants who had passed SPEAK as well as an additional locally designed test for teaching assistants. It was pointed out, however, that undergraduates do not always voice their complaints to administrators. Survey results from undergraduates indicated that non-native teaching assistants were consistently rated below native speaker teaching assistants. However, mean responses for all non-native speaker TAs who had passed the university screening process were within an acceptable level of instructor competence or within a quarter of a point below it. Interestingly, length of time spent in an English speaking country did not relate to scores on SPEAK except when the number of years was above five with those students scoring about a 200 on the SPEAK. TOEFL scores also did not relate to SPEAK scores except when above 600, when most scored 230 on the SPEAK. The best indicator of speaking ability appeared to be self-ratings by the examinee (Abraham & Plakans, 1988).

Factors Affecting Undergraduate Receptivity of ITAs

Undergraduate complaints about the barriers in communicating with international teaching assistants are typically blamed solely on language. ESL specialist and others recognize, however, the interaction of three major factors impacting the ITA communication context, namely, language, teaching, and culture (Gorsuch, 2003).

Bresnahan and Kim (1993) found authoritarianism, dogmatism, and individualism as strong predictors contributing to a lack of receptivity of international teaching assistants by US undergraduate college students. These authors hypothesized that (a) an

unintelligible foreign accent will cause negative affective response (b) preference will be shown to a friendship identity compared to a teaching assistant identity, and (c) students with a strong ethnic identity, more so than those with weaker ethnic identity, will exhibit more negative attitudes and emotional responses to people with accents. In the study, 311 college students were presented speech samples of native and non-native speakers reading a text about friendship as well as a text simulating a college of communication class lecture. After listening to the speech samples, participants completed a questionnaire measuring.

Results confirmed that positive attitudes and affect were associated with accents that were rated as more intelligible. The participants also showed more positive attitudes when listening to the friend speech sample than to the teaching assistant communication lecture. Finally, students with stronger ethnic identity tended to rate the preference for native speaker English higher than those with weaker ethnic identity. According to the authors, intelligibility was supported as "a key issue in receptivity to foreign teaching assistants" (p. 181-182). In fact, the non-native speech samples with higher intelligibility were judged to be more attractive and dynamic. This study was based on social identity theory a perspective that suggests people exhibit a preference for the variety of language associated with their in-group (M. J. Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002).

ITA Intelligibility

In a study by Anderson-Hsieh, Johnson, and Koehler (1992), the main research question focused on what impacted ratings of intelligibility and acceptability on the SPEAK Test, an oral proficiency assessment given to prospective international teaching assistants at many American colleges and universities to screen for language ability prior to assigning international graduate students to teaching assistant positions. Three experienced ESL teachers impressionistically rated the oral reading samples of 60 SPEAK recordings produced by speakers of 11 language groups. The criteria for judging pronunciation were based on ratings of intelligibility and acceptability. In the study, pronunciation deviance was divided into three major areas: segmentals, syllable structure, and prosody. Segmental errors included deviation on consonants and vowels. Syllable structure errors were based on addition or deletion of syllables. Prosodic features, that is, stress (word and tone group), rhythm, intonation, and phrasing were rated impressionistically rated on a 3-point rating scale with 0 as least native-like to 3 as most native-like. Inter-rater reliability ranged from .80 to .89. Overall prosody rating was also measured and eventually solely used in the analysis because there was such a high correlation among the prosodic features and the overall prosody score. Intelligibility and acceptability were judged on a 7-point Likert scale. Pearson correlation and multiple-regression were used to predict the variance in segmentals, syllable structures, and prosody from ratings of intelligibility and acceptability.

The results of the adjust R-squared indicated 89% of the variation in the pronunciation rating was accounted for by the independent variables of segmentals, syllable structures, and prosody. The standardized multiple regression coefficients results were as follows: segmental errors = -.27, syllable structure error = -.25, and prosody = .58. Prosody and syllable structure showed the highest correlation at r = .69. Based on these results, the researchers claimed that prosody had a greater influence on raters' judgments of pronunciation than the influence of segmental errors or syllable structure errors (Anderson-Hsieh et al., 1992).

SPEAK Fluency Ratings

Wennerstrom (2000) extended research on the construct of second language fluency by conducting a conversation analysis exploring the role of intonation as an important variable in the perception of fluency. She pointed out that based on conversational analysis studies perceived fluency is influenced by frequency and type of hesitation phenomena, frequency of repair, amount and rate of speech, and other interactive features of conversation management. The assessment criterion of fluency included definitions describing less fluent speech as halting, fragmented, and with numerous pauses; in contrast to fluent speech described as smooth and effortless. She also, however, raised the issue of rater-scale interaction in claiming that simple assessment criterion description of fluency fail to define specifically what aspect of speech influences the rater's judgments. The problem is that "while such descriptions may be appropriate to instruct test raters to think holistically in the judgments of fluency,

we are left without a clear notion of what linguistic features actually correspond to these general characteristics" (p. 103). She also rationalized that because fluency and comprehensibility scores from the SPEAK Test had a correlation of .85, there was strong evidence that fluency strongly contributed to overall language ability in the perception of raters.

In the study, speech samples were both rated by native judges and then analyzed on a Computerized Speech Lab machine to measure the pitch patterns of the non-native speakers. Wennerstrom used two SPEAK trained raters to judge informal conversational dialogues between native and nonnative speakers of English, including four Korean speakers, three Japanese speakers, one Mandarin speaker, one Thai speaker, and one Swiss speaker of Italian. The raters listened to tape recorded segments of the conversations and then individual utterances from the non-native speakers that were judged 0 to 3 on the areas of pronunciation, fluency, and comprehensibility. Averages of the fluency rating scores were taken across both tasks and raters to assign a single mean score on fluency. The speech analysis "focused on (1) the relationship between the pitch levels of individual words and the contribution each word made to the discourse and (2) the direction of the boundary tones in the context of the turn-taking sequence" (Wennerstrom, 2000). The three lexical categories were words that contributed new or contrasting information, words that conveyed given information, and words of low information value such as articles and prepositions.

Results of the study indicated a difference in intonation patterns depending on the level of fluency. Speakers who were rated more fluent were better able to signal relationships among words and parse their speech for turn-taking in the conversations. Wennerstrom (2000) described that "…it is not longer utterances or shorter pauses per se that lead to a perception of fluent speech; instead, it is the ability to speak phrasally rather than word-by-word, focusing the main idea of each utterance in a coherent manner and collaborating in the turn-taking process" (p.125). She concluded that the use of pitch on lexical items and use of pitch at boundaries signaling the hold or relinquishing of turns were aspects of intonation that need to be included in interactional models of fluency in English.

Synthesis and Summary

This literature review covered oral proficiency assessment, language constructs, and international teaching assistant studies. In the first section, I used Bachman's CAL model to depict the multi-componential nature of oral proficiency and introduced *grammatical knowledge* the component of oral proficiency on which this study focused. Next, I reviewed linguistic terminology to familiarize the reader with definitions of phonological features that appear in the research investigating the language constructs of accent, intelligibility, comprehensibility, and fluency. Distinctions and relationships between accent, intelligibility, comprehensibility and fluency were demonstrated as well as investigations specific to ITAs, the context of the current study.

Table 2.3 provides a summary of the terms and how they were operationalized within the specific studies already reviewed. The purpose of the table as shown by the column categories is to identify (a) the research studies, (b) the constructs under investigation, (c) the phonological features of each respective study, and (d) how the features were measured.

Table 2.3 Review of Empirical Studies Including Linguistic Variables

Study	Phonological Features	Definition/Measurement
(Anderson-Hsieh & Koehler, 1988)	Segmentals	1-7 pt Likert from least native-like to near native-like
Accent	Syllable structure	1-7pt Likert from least native-like to near native-like
	Speaking rate	Syllables per second
	Prosody (stress, rhythm, & intonation)	1-7pt Likert from least native-like to near native-like
(Anderson-Hsieh, et. al., 1992)	Syllable structure	Number of syllables added & deleted/total number of syllables in target sample
Intelligibility/ acceptability	Word and phrase stress	Correct syllables stressed in words; appropriate accent in tone groups
	Phrasing	Appropriate, pauses occur at syntactic boundaries
(Munro & Derwing, 1998)	Speech rate	Number of syllables/total speaking time
Accent		

Study	Phonological Features	Definition/Measurement
(Munro &Derwing, 1999)	Segments	Number of errors in production
Accent	Phonemic	Deletion, insertion, or substitution of segment
Intelligibility	Phonetic	Errors in production of segment
Comprehensibility	Intonation	9 pt Likert from native-like to not at all native-like
	Grammar	
(Munro & Derwing, 2001)	Speech rate	Syllables per second
Accent/comprehensibility		
(Derwing, et. al., 2004)	Pruned syllables	Total syllables minus disfluencies/ total number of seconds
Fluency	Mean length of runs	Number of syllables btw unfilled pauses of 400 ms or longer
Accent	Number and duration of pauses	Counted syllables
Comprehensibility	Goodness-of-prosody (intonation & rhythm)	1-9 Likert native-like to extremely not native-like
(Kormos & Denes, 2004)	Speech rate	Total number of syllables/total time
Fluency	Articulation rate	Total number of syllables/total time minus pause time

Study	Phonological Features	Definition/Measurement
	Phonation-time ratio	Time speaking/total time of speech sample
	Mean length of runs	Avg. number of syllables btw pauses above .25 sec.
	Number silent pauses/ minute	Pauses over .2 sec. considered
	Mean length of pauses	Total length of pause/total number of pauses
	Number of filled pauses	Total number of uhm, er, mmm/total time
	Number of disfluencies/ minute	Repetitions, restarts, and repairs/total time
	Pace	Number of stressed words/minute
	Space	Proportion of stressed words to total number of words

There are several interesting points upon a review of these studies. First, phonological features influence listener perceptions of accent, intelligibility, comprehensibility, and fluency. The variation in this literature contributes to broadening our knowledge of the interrelationships between phonological features and oral proficiency concepts. It is evident that these researchers are in search of defining the linguistic variables that influence listener perceptions of oral proficiency concepts.

However, a review of the existing studies also points to limitations of this literature. The shortcoming in identification of primary contributing prosodic features on

comprehensibility may be due to methodology. First, measuring prosodic features through ratings does not identify specific variability. In other words, the dichotomy or even range of sounding native-like or not offers no explanation as to what in the phonological features deviated and how it deviated. Second, although a one time rating of accent, intelligibility, comprehensibility, and fluency in relation to rated or counted error phonological features offers correlational evidence, lacks the element of change over time or how phonological features change when perceptions of comprehensibility change. Finally, sample size and the time consuming nature of dealing with phonological features either impressionistically or instrumentally has in the past imposed certain limitations on this area of research. Partly because of these reasons, the my own study was different in that (1) linguistic variables were quantified through coding and scoring, not simply through ratings, and (2) investigated grammatical, temporal, and phonological variables based on change in failed and passed comprehensibility rating scores. From the studies listed in Table 2.3, I adopted the following linguistic variable definitions and measurements for the study:

a. Segments (Anderson-Hsieh & Koehler, 1988; Munro & Derwing, 1999)
b. Speech rate (Anderson-Hsieh et. al., 1992; Munro & Derwing, 1998, 2001)
c. Syllable structure (Anderson-Hsieh & Koehler, 1988)
d. Phrasing/thought-groups (Anderson-Hsieh et. al., 1992)
e. Mean length of runs (Derwing et. al., 2004)
f. Disfluencies (Kormos & Denes, 2004)

CHAPTER 3: METHOD

In this study, I used a three-part mixed method design to explore the difference between a set of failed and passed oral proficiency assessments. First, I investigated the change in rater judgments between the failed and passed performance evaluations. Next, I conducted a grammatical, temporal, and phonological speech analysis to explore whether specific variables from these categories differentiated the two differently rated performances. Lastly, I carried out 10 case studies to evaluate how the unacceptable versus acceptable oral proficiency performances differed at the individual level.

Three-Part Study Overview

In Part A, I used multivariate analysis of variance (MANOVA) to analyze how the rater judgments on the assessments' five tasks and four rating scale criteria changed between the failed and passed assessments. The dependent variables in the first MANOVA were the total scores from each of the five speaking tasks; that is, from the summary of an article, the pronunciation of terms, the explanation of terms, the interpretation of a graph, and the classroom role-play tasks. The dependent variables in a second MANOVA were the criteria rating scale scores on pronunciation, grammar, fluency, and comprehensibility averaged across task. Finally, I ran a third MANOVA on the rating scale scores from only the summary task. Results of this investigation were expected to be indicative of which tasks and criteria had changed in order to promote this set of ten initially failed performances to a passing level. My main purpose for this first

look at the data was to identify an appropriate task on which to conduct the speech analysis for Part B.

Based on the results from Part A, I selected the task that required an explanation and summary of an article, from here on referred to as summary task, for a detailed speech analysis. The process involved coding and counting three categories of speech production variables on a 2-minute speech sample from each failed and passed assessment. Recall from Table 1.1 that I was interested in the three groups of linguistics variables: grammar, temporal, and phonological. I counted all grammar errors on each 2-minute speech sample and compiled them by error categories such as subject/verb agreement, dropped final-s, and missing article. To evaluate temporal variables on the set of failed and passed performances, I calculated speech rate, articulation rate, total pause time, mean length of runs, and disfluencies. The phonological variables included counting all sound errors in each speech sample. I measured suprasegmental level features by counting thought-groups per utterance and prominence per thought-group. I used descriptive statistics and dependent T-tests to test statistically the hypothesis that the results on the grammatical, temporal, and phonological speech variables on the set of passed assessments would be significantly different from the results on the set of failed assessments. I expected the speech analysis results on grammatical, temporal, and phonological speech characteristics to coincide to some extent with the rater judgments of grammar, fluency, and pronunciation, respectively. In addition, I was testing the

assumption that the change in the overall oral proficiency score from a failed to a passed level would reflect some improvement in the oral performance of the speech variables.

Finally, in Part C, I turned to a qualitative approach by conducting multi-case studies on each of the 10 international graduate students' assessments. It had become evident in completing Parts A and B of the data analysis that I needed a more detailed, encompassing, and individual level analysis. In Part A, the MANOVA was based solely on the rating scale scores but could not incorporate the evaluators' handwritten notes about the performance. The speech analysis showed addition limitations from the small sample size and in collapsing the resulting from 10 individuals into mean scores. The small sample size with such a large number of linguistics variables restricted the type of statistical analysis possibilities. In addition the failed dependent T-tests and variable ranges revealed complexities in the phenomena beyond what mean scores could not address. Basically, I was motivated by the results of Part B to explore more in depth the range of performances on the linguistic variables and the interrelation among the variables between each individual's failed and passed speech sample.

Part A: First Investigation of the Raters' Judgments

Part A Rationale

Rater judgments often serve as the deciding factor in high-stakes speaking assessments for non-native English speakers (Anderson-Hsieh et al., 1992; Derwing et al., 2004). The purpose of the first part of my investigation was to identify the changed scores between the failed and passed assessments based on the rating scale judgments of

the five tasks and four criteria. My aim was to contrast the acceptable and unacceptable assessments to find how the tests differed based on the rater judgments.

Because past research has established that fluency in a second language depends on the nature of a speaking task (Derwing et al., 2004; Egbert & Petrie, 2005), there was reason to explore whether improvement on a specific assessment task had influenced the change from a failed to a passed oral proficiency result. Similarly, the assessment criteria of pronunciation, grammar, fluency, and comprehensibility were analyzed to determine which had influenced the overall score improvement in this high stakes oral proficiency assessment.

Part A Context and Data Source

The first investigation focused on rater judgments obtained from the prospective ITA oral proficiency assessments at a large publicly supported Research 1 university. Approximately 100 to 200 international graduate students from at least ten different countries participate in oral proficiency screening every semester. These international graduate students, like many across the United States, are self-selected or sent by their departments to have their oral English skills evaluated in order to become eligible for departmental assistantships for financial support as well as academic teaching and/or research experience (Saif, 2006). Typically, these students have learned English in their native countries and have come to the United States to pursue graduate degrees in a wide range of disciplines. Generally speaking, each student has scored above 550 on the paper-

based TOEFL (or 213 Computer Based Test) and above a combined verbal and math score of 1000 on the GRE to meet university and departmental entrance requirements.

The English Oral Proficiency Test (EOPT) is worth a total of 30 points, although results are reported on a 300-point scale by multiplying the score by 10 (Appendix C). Students who score above 25 on the EOPT are eligible to take a position as a teaching assistant. Those who score below 25 but above 23 are required to take a semester-long course on oral communication skills. Individuals who score below 22.5 do not pass the oral proficiency test but have the opportunity to retake the assessment up to five times. Table 3.1 shows a breakdown of the assessment results for Fall 2005 and Spring 2006 semesters, the period from which the oral proficiency assessments for this study were taken.

Table 3.1 EOPT Results Summary for Fall '05 and Spring '06

	Total Number	Pass	Conditional Pass	Fail
Fall 2005	208	132	44	32
Spring 2006	95	51	21	23
Totals	303	183	65	55

During the 2005-2006 academic year, 60% of the international students taking the test passed on their initial attempt, while 21% conditionally passed, and 18% did not pass the assessment. It should be noted that international students coming from a country where English is the primary language (e.g., Great Britain, Ireland, South Africa,

Australia, New Zealand) or where English was the primary language for instruction from elementary school through secondary level were exempt from the oral proficiency testing requirement.

The raw data for Part A consisted of the 10 rating scale judgments made during the international students' performance on five speaking tasks (summary of an article, pronunciation of terms, explanation of terms, graph description, and role-play). The rating scales cover four criteria of pronunciation, grammar, fluency, and comprehensibility (Appendix A). Note that not all four criteria are rated for each task, although the summary task does get rated on all four criteria. The rater evaluations were taken from a set of failed and passed assessments by the same 10 individuals between Fall 2005 and Spring 2006 semesters. The comparison was possible because of the institution's retake policy. A period of at least three months was required between assessments, with preparation for the exam and re-exam left entirely to the discretion of the student. Because the source of the data was archived, it was not possible to survey the international students directly regarding their test and retest preparations.

The Texas Intensive English Program, the entity in charge of administering and maintaining records on the EOPT, keep the following data on each examinee: last name, first name, date of birth, phone number, country of origin, native language, language of instruction, major, employing department, graduate advisor, score, and date of test. For each assessment, a cassette tape of the entire assessment and the two rater evaluation forms were kept on file. All tapes are labeled with the student's name and the date of the

assessment. Each rater evaluation form includes: student's name, department, date of test, and rater's name. A database is used to assist in scheduling and tracking students who retake the assessment each semester.

As previously mentioned, the data used in this study included only international graduate students who had not initially passed the EOPT but later retook it and obtained a passing level during the Fall 2005 and Spring 2006 semesters. These 10 students included five native Korean speakers and five native Mandarin speakers. Each of the 10 individual's scores, testing dates, date of birth, major, and total times assessed are listed in Table 3.2 below. Five individuals retook the exam only once in order to attain a passing score. From the remaining five students, three had taken the test twice and two had taken it three times to reach the required level of English proficiency.

Table 3.2 Demographic Information on Prospective ITAs

#	Score	Date	Native Language	M/F	Birth Year	Major	Total # Assessments
1	215	8/19/05	Korean	M	1976	Mechanical Engineering	2
	250	1/6/06					
2	210	8/18/05	Chinese	M	1980	Chemistry	2
	250	5/22/06					
3	210	4/14/05	Korean	M	1974	Electrical Engineering	3
	250	8/23/05					

#	Score	Date	Native Language	M/F	Birth Year	Major	Total # Assessments
4	200	8/18/05	Chinese	M	1982	Physics	3
	250	5/22/06					
5	175	1/10/06	Chinese	F	1979	Microbiology	3
	260	5/22/06					
6	215	8/17/05	Chinese	M	1978	Chemistry	4
	250	1/9/06					
7	220	4/13/06	Korean	M	1970	Journalism	2
	250	5/22/06					
8	220	11/10/05	Korean	F	1980	Kinesiology	2
	255	5/22/06					
9	220	5/23/05	Korean	M	1976	Computer Science	2
	280	8/19/05					
10	220	5/23/05	Chinese	M	1981	Biology	4
	255	11/8/05					

Part A Measures

This study made use of the English Oral Proficiency Test (EOPT) to evaluate the oral speaking skills of international graduate students who hoped to become ITAs. Each assessment was conducted by two experienced English as a Second Language (ESL) instructors who served as the EOPT raters. The tests were tape recorded and lasted approximately 20 minutes. Five different speaking tasks judged on four different

assessment criteria comprised the measure of oral proficiency on a 30-point scale. As mentioned earlier, the students' final scores were reported on a 300-point scale by multiplying the overall possible 30 points by 10.

The primary data for Part A of the study were the two EOPT rater evaluations (Appendix A) from the set of failed and passed assessments by the 10 individuals listed in Table 3.1. The evaluation forms listed five speaking tasks with ten total rating scale criteria scored from zero to three. The speaking tasks included: explanation and summary of an article, pronunciation of terms, explanation of terms, interpretation of a graph, and a roleplay simulating a teaching assistant situation. The judges rated each task on one or more of the following four criteria: pronunciation, fluency, grammar, and/or comprehensibility (Appendix B). Each criterion was judged on a 0-3 rating scale.

Each assessment began with a few minutes of warm-up conversation during which the instructor-rater asked questions to generate small talk. On the first task, each student was given four minutes to summarize a short, field-specific article that they had been given after arriving at the testing center and had had 30 minutes to prepare prior to the start of the assessment. Raters evaluated this task on pronunciation, grammar, fluency, and comprehensibility. Next, the student read aloud from a list of 40 discipline-specific terms. The instructor raters judged the student's performance on pronunciation. On the third task, the student was asked to define two of three given words from the previously read list, and the rater judged comprehensibility. After this, the raters gave the student one minute to look over a graph and then asked him or her to describe it.

Grammar and comprehensibility made up the evaluation criteria. For the last speaking task, the raters gave the student a roleplay scenario regarding classroom information. The examinee explained the information as if he or she were a teaching assistant and had to present the information to a class of university students. Raters judged the performance on pronunciation and comprehensibility (Appendix D).

Part A Design

The design for the first investigation involved a comparison between the set of failed versus the passed performances from the same ten individuals. I ran three repeated measure multivariate analysis of variance (MANOVA): (1) a comparison of differences in the five speaking tasks, (2) a comparison of differences on the four criteria of pronunciation, fluency, grammar and comprehensibility, and (3) a comparison of these same criteria in the summary task. The purpose was not to establish that failed versus passed scores differed, as this was already established. Instead, I conducted the analyses to discern which tasks and which criterion had caused the overall score improvement to a passing level.

In Part A, time was the within-subjects factor and there was no between-subject factor. The dependent variables consisted of the mean ratings from two raters on the five tasks in the first analysis and the mean ratings on four criteria (pronunciation, grammar, fluency, and comprehensibility) in the second analysis. The rationale for running two separate analyses, on the task and criteria, was due to the fact that the same rating scale judgments made up both the task scores and the criteria scores. I also ran a third

multivariate analysis of variance (MANOVA) with the four criteria scores on the summary task.

Part A Procedure

In Part A, the mean ratings available from the 20 rater evaluation sheets served as the primary data. In order to prepare the data for analysis, I took the following steps. First, I located the rater sheets for each failed and passed assessment in the file systems and copied them. Next, I averaged the 0-3 rating scale judgments by the two raters and entered these into a spreadsheet. I organized the scores into a mean overall score, mean score on each task, and mean score on each of the four assessment criteria of pronunciation, grammar, fluency, and comprehensibility. I then completed the same process for the passed assessments. I uploaded these mean rater scores from the failed and passed set of assessments into SPSS and used these to conduct three separate repeated measures multivariate analysis of variance (MANOVA). I tested the following hypotheses to investigate where change occurred between the set of failed and passed speaking assessments based on the raters' judgments.

Part A Hypothesis and Data Analysis

Hypothesis 1.A.

Mean rater scores comparing the failed and passed oral proficiency assessments on the same 10 individuals will differ significantly. The mean rater scores on the passed oral proficiency tests will be higher than scores on the failed oral proficiency test.

Hypothesis 1.B.

Mean rater scores on the five EOPT tasks (summary, terms, explanation, graph, role play) on the failed and passed assessments will differ significantly. The mean rater scores from the passed oral proficiency tests will be higher than scores on the failed oral proficiency tests.

Hypothesis 1.C.

Mean rater scores on the four EOPT criteria (pronunciation, grammar, fluency, comprehensibility) on the failed and passed oral proficiency assessments will differ significantly. The mean rater scores on EOPT criteria will be higher on the passed oral proficiency tests than scores on the failed oral proficiency tests.

Hypothesis 1.D.

Mean rater scores on the four EOPT criteria (pronunciation, grammar, fluency, comprehensibility) for the summary task will differ significantly. The mean rater scores on the summary task criteria will be higher on the passed oral proficiency tests than scores on the failed oral proficiency tests.

I tested Hypothesis 1.A through an omnibus *F* statistic in a repeated measure multivariate analysis of variance (MANOVA) using SPSS (Version 11.0). Univariate *F* tests investigated Hypothesis 1.B, the statistical difference on the task dependent variables from the set of failed and passed assessments. I conducted a second repeated measures multivariate analysis of variance (MANOVA) with criteria as the dependent variables. The univariate *F* statistic in this analysis tested Hypothesis 1. C. The third

multivariate analysis of variance (MANOVA), based only on the summary task, evaluated Hypothesis 1.D. through the univariate *F* statistic. A significance level of .05 was used for all analyses.

I expected that the first investigation results would establish (a) if differences in the task performances influenced the change in overall score, (b) whether change in criteria scores contributed to the overall test score improvement, and (c) if all four criteria on the summary task changed between the failed and passed performances. The results provided grounds for choosing one task on which to conduct the speech analysis in Part B of the study. In addition, the results of Part A and B were later combined in Part C where I compared and contrasted the rater judgments and speech analysis results in for each of the 10 case studies.

Part B: Analysis of Speech Production Variables

Part B Rationale

The purpose of the speech analysis was to explore a connection between the rater judgments on the assessment criteria of grammar, fluency, pronunciation, and comprehensibility, with the linguistic variables performed on the speech samples. A body of research has investigated linguistic variables in relation to the language constructs of accent, intelligibility, comprehensibility, and fluency (Anderson-Hsieh & Koehler, 1988; Kormos & Denes, 2004; Munro & Derwing, 1998, 2001). The temporal measures of speech rate, mean length of runs, and disfluencies, for example, have been established as predictors of perceived fluency levels (Kormos & Denes, 2004). In

addition, a limited but growing number of studies have provided evidence for the significant role of prosodic features, specifically intonation, stress, and thought-groups in relation to perceived fluency levels and successful communication (Pickering, 2001; Wennerstrom, 2000).

Based on past findings, I wanted to test the expectation that grammatical, temporal, and phonological variable results would be weaker on the set of failed assessments compared to the results on the set of passed assessments. The context of ITA oral proficiency assessment provided a fertile ground for connecting underlying linguistic variables with language constructs via the rating scale criteria. Listener judgments on the specific assessment criteria documented the perception of competence in specific skills. Matching grammatical, temporal, and phonological speech characteristics to the assessment criteria of grammar, fluency, pronunciation, and comprehensibility could provide indications for what influenced the change in the rater judgments for these 10 oral proficiency test performances.

Although convenient for assessment, rating scales provide no means for extrapolating which specific linguistic feature cause favorable or unfavorable judgments of a specific language construct or assessment criterion, like fluency (Wennerstrom, 2000). In other words, a number on a rating scale offers no information regarding what in the speech signal caused the perception or reaction by the rater. The investigation allowed me to supplement the rating scale judgments with the actual linguistic variables produced during the assessments.

Based on results of Part A, the summary task statistically differed on grammar, fluency, and comprehensibility. Therefore, evidence was provided for making a comparison on linguistic variables evaluating grammatical, temporal, and phonological characteristics produced on the failed and passed assessments performances. In addition, the repeated measure type design built in a control for variation because each individual's performance on the failed test was compared to his or her own later performance on the passed test from the same field-specific summary task.

Part B Measure

The data for Part B came from the same English Oral Proficiency Test (EOPT) performances as described in Part A. Due to the large number of speech variables to be counted; however, I analyzed only 10 utterances from approximately 2-minute speech samples from the summary task for Part B. As previously mentioned, for this speaking task, the students had prepared and then gave a verbal summary of a field-specific article. The students had 30 minutes to prepare the summary prior to the start of the assessment and could refer to their notes during the 4-minute limit for the verbal summary task. Of the five speaking tasks on the EOPT, I chose the summary for the speech analysis because (a) it represented 40% of the overall score; (b) it was the only task that included all four assessment criteria of pronunciation, fluency, grammar, and comprehensibility; and (c) the first investigation validated it as a viable task because the failed and passed scores were statistically different.

Part B Design

The objective of Part B was to quantify the linguistic variables on 2-minute speech samples of each failed and passed oral proficiency assessments from the 10 international graduate students. I conducted the investigation through a mixed method design in which grammatical, temporal, and phonological characteristics were coded, counted, and evaluated using descriptive statistics. Based on mean scores, I ran dependent T-tests to test the statistical differences between the failed and passed speech analysis results.

My first step was to isolate 10 utterances isolated from the summary task. This was approximately 2-minute from a 4-minute speech recording. All pauses above .33 seconds were measured along with overall time to produce the utterances. I tallied the total number of syllables, disfluencies (uhs, ums, repetitions, restarts, and repairs), and segmental errors. I calculated speech rate, articulation time, and the mean length of the runs. I evaluated the accuracy of thought-groups and prominence (stress). I also counted and categorized the grammatical errors on the speech samples. I asked five native speakers, who were first and second year graduate students in an educational psychology department, to complete the summary task. I conducted the same speech analysis on their speech samples to use not as a standard but as a point of reference for skill performance levels.

I compared the speech analysis results by failed, passed, and native speaker group averages. I first calculated and compared the descriptive statistics including the mean,

range, and standard deviation on all variables between failed and passed groups. I ran nine dependent T-tests, with Bonferroni adjustment, to compare the differences in grammatical, temporal, and phonological measures between the set of failed and passed performance. Table 3.3 lists the specific variables compared on the set of failed and passed assessment results.

Table 3.3 Dependent T-Tests on Speech Analysis Results

Speech Production Categories	Dependent T-Tests
Grammatical	Failed versus Passed
Temporal	
Speech Rate	Failed versus Passed
Mean Length of Runs	Failed versus Passed
Total Pause Time	Failed versus Passed
Uhs	Failed versus Passed
Repetitions	Failed versus Passed
Phonological	
Segmental errors	Failed versus Passed
Thought-group counts	Failed versus Passed
Prominence counts	Failed versus Passed
Average # words/thought-group	Failed versus Passed

I also ran the same independent T-tests comparisons listed in Table 3.3 with Bonferroni adjustment to control for inflated Type I error rate, on the set of passed assessments and the five native speaker results.

Part B Procedure

The procedures for Part B included three major steps: (a) data clean-up and preparation, (b) coding and scoring, and (c) reliability check on 20% of each variable coded by two raters in addition to myself.

Data clean-up and preparation. First, I digitized each of the 20 cassette recordings using a cassette deck (JVC TD-R462), a Macintosh computer, and Soundtrack Pro (Version 1.0.3) software. Next, I cut the 4-minute speech samples from the summary task from the original 20-minute assessment recordings. I saved each of the 20 4-minute summaries as a wave file and then transcribed each summary. I used Final Cut Pro (Version 5.0.4) software to measure the total speaking time and all the pauses greater than .33 seconds (10 frames), the smallest unit that could be measured consistently across all samples due to the quality of the original recordings. I measured the length of pauses through the in/out function keys using Final Cut Pro software. As I took these measurements, I identified the number of frames (30/second) within the transcription files. Because the analysis was based on 10 utterances isolated from the original 4-minute speech samples, I used the pauses to assist in the identification of T-units and intonation units defined below. I also measured the total time it took to produce the 10

utterances using Final Cut Pro and used this measure to calculate speech rate and articulation rate.

Coding and scoring. In the next section, I review how each speech variable was defined and measured.

(1) Utterances. Because the speech samples were spontaneously spoken summaries, I needed a method to identify a unit of analysis. An accepted method of parsing speech in discourse analysis is through the use of minimal terminable units, commonly called T-units (Hillocks, 1986). According to Hillocks, a T-unit is "a main clause with all of its appended modifiers, including subordinate clauses" (1986, p. 64). First, I parsed the speech samples simply by reading the transcription and double spacing where I thought T-units occurred. Because T-units are based on grammatical structures, however, and the speech samples were oral and not written, I conducted a second parsing using intonation units (1996). When I could discern one of the four criteria of a rise, fall, pitch change, or lengthening, then I modified the determination of where a T-unit began and ended as indicated by the speaker's performance on these intonation properties (Tench, 1996). I began with the transcripts of the T-units and then adjusted the T-units as indicated by the intonation heard on the recording in the oral assessment. I played the recordings in Final Cut Pro, thus providing me with a visual wave form that assisted in the identification of intonation unit markers. When I could not clearly establish an intonation unit or T-unit due to lack of clear indicators on the intonation markers and/or grammatical construction, I made a judgment as to the boundaries of the T-unit. In other

words, when I could not discern a main clause from the candidate's words, I based the parsing of the utterance on an estimated T-unit. This situation occurred in less than 5% of the 200 utterances.

(2) Syllables and syllable structure. With the primary unit of analysis established, I counted syllables using the method in Anderson-Hsieh, Johnson, and Koehler (1992). First, I placed a dot under each syllable in the transcript and counted the dots. Next, I carefully followed the transcript while each speech sample recording played to make adjustments in the counts based on what the student actually produced. In addition, I wrote vowel or consonant epenthesis (insertions) onto a transcript copy. For example, when "of" was pronounced as "ov-u," I counted two syllables and noted the addition of the syllable on the transcript. I marked deletions, that is, any missing syllables, on the transcript by placing an "x" over what was not pronounced. In one case, "transparency," a four syllable word, was pronounced without the last syllable. I counted this word as three syllables with a deletion noted. Finally, I tallied the total number of syllables actually produced.

(3) Speech rate and articulation rate. With the total number of syllables and total speaking time established, I could calculate speech rate and articulation rate. Using Griffith's (1991) recommendations from pausology research, I calculated speech rate as the total number of syllables divided by total time to produce the utterances. Articulation rate, the total number of syllables divided by total time minus total pause time, was also

calculated for each of the 20 speech samples as well as for the five native speaker samples (Griffith, 1991; Kormos & Denes, 2004).

(4) Mean length of runs (MLR). Following previous research (Derwing et al., 2004; Kormos & Denes, 2004), I included mean length of runs as a temporal measure. I tallied the total number of syllables produced between pauses of at least .33 seconds for each of the 10 utterances. Although pauses above .25 seconds have been established as the most reliable cut off (Kormos & Denes, 2004), .33 was the smallest unit consistently measurable with the quality of recordings in this study.

(5) Disfluencies. Griffith (1991) categorized disfluencies as both filled and unfilled pauses. I divided filled pauses into two categories, those of "uh, uhm, mmm," and repetitions, restarts, and repairs. Because I could not always distinguish between repetitions, restarts, and repairs, I counted these in a single category. Partial words uttered and repeated words or phrases were all counted under the category of *repetitions*. I used total pause time as the measure for disfluency variable of unfilled pauses.

(6) Segmentals. Using Final Cut Pro software, I identified the phonetic inaccuracies on the failed and passed speech sample utterances. I reduced each sound files' speed to 80% and played it repeatedly until I could note all deviations in phonetic accuracy. When I heard a sound inaccuracy, I wrote the International Phonetic Alphabet (IPA) symbol for the intended sound over the orthographic transcription. Total segmental errors represented all vowels and consonants that I perceived to deviate strongly from native speaker pronunciation.

(7) Grammar. Next, I evaluated each of the 200 utterances for grammatical accuracy. I counted and categorized all grammar errors. The grammar categories included prepositions, articles, verb placement, subject/verb agreement, verb tense, missing verbs, as well as final-s deletions and additions. Other categories that I coded included inaccuracies in the use of connectors, word choice or odd expressions, plurals, demonstratives, and parts of speech.

(8) Thought-groups. Based on the marked pauses of greater than .33 seconds, I identified chunks of language within the utterances or T-Unit/intonation units. These units captured a smaller level of production breaks in the speech samples. Recall from the review of literature this fundamental unit of intonation has been previously termed *tonality* (Halliday, 1967), *tone group* (Ladd, 1996), *intonation unit* (Tench, 1996), *tone unit* (Brazil, 1997), *intonational phrase, sense group,* and *breath group.*

(9) Prominence. Stress is a primary component of prosody. Acoustically, stress is duration, intensity, and loudness applied to a syllable (Cutler & Ladd, 1983; Snow, 2001). The stress phenomenon occurs within a thought-group through *prominence*. On the prominent syllable, at least five possibilities exist: rise, fall, rise-fall, fall-rise, and neutral (Halliday, 1989). Because of the difficulty in impressionistically identifying just primary stress, all discernable syllables with prominence were coded.

Reliability. In order to insure accuracy in the data analysis process, a total of five ESL instructors conducted reliability checks on 20% of all the variables as defined above with teams of checking different variables of the data. I coded variables on 200 non-

native utterances and two of the five experienced ESL professionals evaluated each variable on 40 of the utterances. The speech samples were carefully selected so that both failed and passed performances were evenly represented in the reliability checks.

Part B Hypotheses and Data Analysis

Because grammar, fluency, and comprehensibility differed between the set of failed and passed assessments in Part A, I expected that accuracy in grammar, temporal measures, and phonological characteristics would differ between the two speech performance recordings.

Hypothesis 2.A.

Mean grammatical error counts from the set of failed and passed summary task will differ significantly. It is expected that there will be fewer grammatical errors on the passed than the failed performances.

Hypothesis 2.B.

Mean temporal measure scores from the performance on the failed and passed assessments will differ significantly. It is expected that the passed assessments will exhibit faster speech rate and fewer disfluencies than exhibited in the failed performances.

Hypothesis 2.C.

Mean phonological characteristics, that is, accuracy on segmentals, thought-groups, and prominence from the failed and passed summary task will differ significantly. It is

expected that there will be more accuracy on segmentals, thought-groups, and prominence on the passed than the failed performances.

Hypothesis 2.D.

Results on the passed assessments will not be statistically different from the baseline native speaker results on temporal measures (speech rate and disfluencies) and phonological characteristics (thought-groups and prominence).

First, I investigated all hypotheses by inspecting the descriptive statistics for the linguistic variables. I examined means, standard deviations, and score ranges of grammar errors, temporal measures, and phonological characteristics. I ran dependent T-tests to test the grammatical, temporal, and phonological differences in results between the set of failed and passed assessments. I ran independent T-test to test the differences between the set of passed assessment speech analysis results against the native speaker results.

Part C: Multi-Case Studies

Part C Rationale

The statistical analyses in Parts A and B, the rater judgment investigation and the speech analysis, revealed to me limitations in the quantitative approach. I saw richness in the data that was not accounted for in the statistical analysis. In the first investigation, for example, only raters' numerical judgments were included with no means to handle the handwritten comments they had included on the rating sheets regarding specific criteria and overall performance.

Although the speech analysis result in Part B went beyond the method of simple rating scale judgments and showed overlapping ranges between the failed and passed performance variables, the analysis could not incorporate directionality of the variables or the interrelation among the variables. For example, in terms of fluency, we would expect a lesser amount of total pause time to be "good" and a higher amount to be "bad." With this line of reasoning, I expected total pause time results from the set of passed assessments to be less than total pause time on the set of failed assessments. However, there were individuals whose total pause time increased on the passed test and contrary to expectations, seemed to have had a positive impact on the oral proficiency performance. Basically, the Part B analysis investigating the group comparison between the set of failed and passed assessments was weakened by this unexpected variation and directionality within the linguistic variables.

In addition, the speech variable means in Part B could not reveal or incorporate the interrelation among the variables. For instance, results of the speech analysis showed that speech rate, a predictor of fluency, remained relatively stable on both the failed and passed assessments, but a decrease in total pause time with a drop in the number of repetitions functioned together to influence the improvement to a passing score. In sum, the complexity of the data and challenges in understanding the change in oral proficiency performances caused me to turn to a qualitative approach to address the research question investigating how the failed and passed assessment performances differed.

Qualitative research, according to Strauss and Corbin (1998), has three major components. It includes using data from various sources, procedures like coding that allow the researcher to interpret and organize the data, and written or verbal reports. For each individual, I had rater judgments and comments, tape recorded tests, coded speech variables, and my personal observations as an experienced ESL professional. With each individual as a separate case, I could address my research question from a qualitative perspective (Stake, 1995). My question, and one often confronted in ITA training and ESL in general, was: what interactive linguistic variables influence a native listener's perception of comprehensibility in the speech from a non-native English speaker? I expected a case study analysis investigating the failed and passed conundrum at the individual level would reveal interrelations among the grammar, fluency, and pronunciation criteria on the one hand and the more encompassing assessment criterion of comprehensibility.

Speaking assessments define oral proficiency by the speaking tasks included on the assessment and the criteria on which the tasks are evaluated (Griffith, 1991, Luoma, 2004). Criteria such as pronunciation, fluency, and comprehensibility are commonplace because of their recognized role in communication. They each independently subsume important theoretical and practical areas. For example, what is the nature of pronunciation and its role in spoken language interaction (Pennington & Richards, 1986)? What aspects of pronunciation, namely segmental or suprasegmental, should be taught? A past line of research on accent, intelligibility, and comprehensibility (Munro &

Derwing, 1999, 2001) has connected phonological features of speech to the perception of being understood. The important role of prosodic features has been pointed out but research methods using rating scales have impeded more direct connections between a listener's perception and specific linguistic features.

Extensive research in second language acquisition has focused on explaining fluency in a second or foreign language. Research findings have identified speech rate, mean length of runs, and disfluencies as predictor variables of fluency (Derwing et al., 2004; Kormos & Denes, 2004; Wennerstrom, 2000). However, how fluency variables interact with other components to influence perceptions of comprehensibility is not well understood. Kormos (2004) asserted that "...fluency research suffers from the lack of studies that investigate a combination of linguistic, temporal, phonological and interactional variables" (p.146). In other words, fundamental areas of language learning, namely syntax, phonology, and fluency, have historically been studied within defined parameters. Taking on the construct of comprehensibility, however, forces the researcher to embrace a more multi-componential or multi-system approach to understand the factors involved.

My rationale in employing a multi-case study analyses was to extend existing knowledge on the interrelation among linguistic variables, namely, syntax, phonology, and temporal measures as contributors to the perception of native listeners' comprehensibility of a non-native speaker. In sum, the multi-case studies would offer a means to analyze at the individual level how the failed and passed assessments differed

based on the combined information of rater judgments and comments, recorded test, results from the speech analysis on linguistic variables, as well as my professional impressions and interpretations. Such an approach facilitated my ability to explore (a) how rater judgments on grammar, fluency, and pronunciation related to overall comprehensibility ratings, and (b) whether the grammatical, temporal, and phonological speech analysis results related to the respective assessment criteria of grammar, fluency, pronunciation, and comprehensibility. Each case study comparing the failed and passed oral proficiency performances then was instrumental in demonstrating how native listeners perceived and rated the non-native speech performance in the context of oral proficiency assessment for international teaching assistants.

Description of the Multi-Case Studies

Rater judgment and comment data. I used three sources of data from each rater beyond the overall scores: (1) the numerical judgments on each assessment criteria of pronunciation, grammar, fluency and comprehensibility from the field-specific summary task, (2) the rater comments on these same criteria and task, and (3) the overall performance comments written at the bottom of the evaluation sheet (Appendix A). Each individual case included this set of data in quadruple; that is, there were two rater evaluations on the failed assessment as well as on the passed assessment performance.

I interpreted the rater comments to reveal what the ESL instructors considered noticeably pertinent to the specific assessment criteria. I used the comments as an indication of speech production variables that influenced the rating scale criteria

judgments. In other words, the comments allowed me a window into what was phonologically, grammatically, and temporally noticed and/or potentially distracting to the evaluators. The comments shed light on the raters' perspectives on the characteristics related to the assessment criteria.

To get more information on the rater judgments, I compiled their hand written comments on each criterion as well as on the overall performance. This allowed for an individual case unit of analysis as well as group performance analysis by rating scale criterion. For example, on the failed assessment, both raters marked "2" but Rater 1 noted the following under pronunciation "stress/r/vowels/sounds/," while Rater 2 left no comments. By compiling comment data on each criterion, I could learn that there were, for example, 78 comments for the failed assessments, with 50% related to pronunciation. I will describe this information in the result section. The overall comments were inspected as additional information in understanding differences between the failed versus passed performance.

Because the ultimate goal was to answer how the failed performances differed from the passed performances, the overall comments were considered an additional source of data, especially on the failed test, as to how the performance in the eyes (or should I say ears) of the rater had been unacceptable. On the passed set of assessments, the comments offered an indication of what was done well or how skill levels could be rated acceptable in spite of areas needing continued improvement.

Speech analysis data. The speech analysis variables can be grouped into three general categories of information: (1) grammatical, (2) temporal, and (3) phonological characteristics from failed and passed speech sample taken from the summary task. These linguistic variables were defined in the Part B procedures. Generally, each two-minute speech sample of 10 utterances was reviewed for grammatical accuracy. I identified grammar errors and then marked them using track changes in Microsoft Word. From the errors marked, I categorized the types of grammatical errors to facilitate a comparison on both the number and type of grammar mistakes found on each speech sample.

From the speech analysis in Part B, temporal measures were compiled. I listed in a spreadsheet timing variables comprised of the number of syllables, total time of the speech sample, speech rate, total pause time, articulation rate, mean length of runs, and disfluencies. As explained in Part B, speech rate was calculated by taking the total number of syllables produced over the total time. Articulation rate differed from speech rate because pause time was taken out of the total time. Recall mean length of runs as the number of syllables produced between pauses of .33 seconds or greater. This variable gave an indication of the stretches of speech produced. Finally, I counted variables of disfluencies, that is, filled and unfilled pauses. I tallied the number of fillers such as "uhs" as well as repetitions, the number of repeated syllables. Table 3.4 depicts the spreadsheet organization comparing these temporal variables between the failed and passed oral proficiency performances.

Table 3.4 Data Sample for the Temporal Variables

Assessment	Byoung-Hyun Failed	Byoung-Hyun Passed
Score	215	250
Syllables	314	172
Syllable Structure	0.000	0.006
Total Time	133.57	74.8
Speech Rate	2.35	2.30
Total Pause Time	46.10	30.87
Articulation Rate	3.59	3.92
MLR	5.52	8.77
Uhs	27	0
Repetitions	30	8

In the phonological analysis for the case studies, I evaluated segmentals, thought-groups, prominence, and the average number of words per thought-group. As described in Part B, I listened to each speech sample and marked all segmental errors on the transcript. I identified thought-groups as the chunks of language between pauses of greater than .33 seconds. I counted the number of thought-groups in each individual's 10 utterances to get an average number per utterance. I marked prominence on a transcript while listening to the digitized recording of the summary task. I also counted the mean number of words in each thought-group to get a sense of instances of prominence across the number of words. I added these results to tables similar to the one depicted in

Table 3.4. Finally, I compiled each individual's results on grammar counts, temporal measures, and phonological variables into a spreadsheet.

Guiding questions and procedures. The qualitative case study approach offered me a way to analyze at the individual level how the failed and passed assessments differed based on the combined information of rater judgments and comments, recordings, speech analysis results, as well as my personal impression as an ESL professional for 10 years and accent modification trainer for five years. Within each case, I could explore (a) how rater judgments on grammar, fluency, and pronunciation related to overall comprehensibility ratings; (b) whether the grammatical, temporal, and phonological speech analysis results related to the respective assessment criteria of grammar, fluency, pronunciation, and comprehensibility; and (c) additional factors not previously considered but that might emerge during the case study. After a detailed review of the data with these questions in mind, I made interpretations regarding the change in oral proficiency performances between the failed and passed assessments. Following is a description of the general protocol I used for each individual's case study.

I evaluated the data by first reviewing the rater judgments on each criterion - pronunciation, grammar, fluency, pronunciation, and comprehensibility as well as agreement or disagreement between the two raters. The numerical rater judgments circled were used for their original intent, that is, as an evaluation of skill level on pronunciation, grammar, fluency, and comprehensibility. They were also used to explore: (a) the interaction among the rating scale judgment, specific assessment

criterion, and relevant results from the speech analysis; (b) how the judgments on pronunciation, fluency, and grammar might have contributed to the comprehensibility judgments; and (c) how change in performance could be explained based on "a" and "b."

Next, I looked at results from the speech analysis I had made using on the same groupings of grammatical, temporal, and phonological speech characteristics. By reviewing the rater judgments in relation to the speech analysis results, I could attempt to connect actual oral production performance to rating scale levels. For example, how many grammar errors were acceptable in a Level 2 rating? With the speech analysis results on grammar, I could look at the number and types of errors in relation to the rating level. Next, I compared both the number and types of errors between the failed and passed performances. The objective was to investigate how the actual grammar errors in the performance influenced the raters' judgments on the grammar criterion.

I followed a similar process for temporal and phonological measures. Because predictors of fluency include speech rate, mean length runs, and disfluencies (Kormos 2004, Wennerstrom, 2000), I compared whether the data supported this previously established finding: did speech rate increase on the passed assessments and how did the other temporal measures differ between the failed and passed assessments. Starting with an inspection of the rater judgments and comments, I then inspected the results of the phonological measures on segmentals, thought-groups, and prominence. The phonological data from the speech analysis offered a window into the potential influences of segmental and prosodic features on the raters' perceived understanding of the speech.

During each step in the analysis, I attempted to extrapolate how the rater judgments and their comments in relation to the linguistic variables could explain the improvement to a passing level at Time 2.

CHAPTER 4: RESULTS

Chapter 4 presents the study results in order of the three-part investigation. First, I give a brief overview of the results. Next, I describe the results from the first investigation looking at the contrast between the failed and passed assessments based on instructor-rater judgments. Then, I report the speech analysis results. Finally, I present three representative case studies to demonstrate the complexity found in explaining how each individual's failed assessment performance differed from the passed assessment.

Part A: Results of the Instructor Rater Judgments Analysis

The purpose of the first investigation was to ascertain, solely based on the original instructor rater judgments from the EOPT, how 10 international students' failed speaking assessments differed from their eventual passed assessments. Results from Part A indicated that, based on the rating scale judgments, scores on four of the five tasks improved from the failed to the passed oral proficiency performances. The following four tasks supported the research hypotheses predicting higher ratings on passed assessments: field-specific summary, explanation of terms, graph interpretation, and classroom role-play. Scores on the pronunciation of terms task did not improve from failing to passing. Additionally, results from the first investigation revealed that scores on the assessment rating criteria of pronunciation, grammar, fluency, and comprehensibility were rated higher on the passed set of assessments than on the set of failed assessments. When reviewing the rating criteria from only the summary task, however, the pronunciation criterion scores did not attain a statistically significant level

of change. The statistical results are presented below starting with (a) the mean ratings grouped by task and then by criteria; and (b) results from the repeated measures multivariate analysis of variance (MANOVA) conducted using the SPSS statistical package.

Out of 30 possible points, the overall assessment scores for the set of 10 passed assessments ranged from 25 to 28 (M = 25.5, SD = 1.4) and were, as expected, consistently higher than the failed scores range of 17.5 to 22.0 (M = 21.5, SD = .92). A repeated measure multivariate analysis of variance (MANOVA) with task scores as the dependent variables resulted in a significant difference between the two sets of tests, F (1, 9) = 30.48, $p < .05$. This was expected based on the test results but not necessarily guaranteed due to the small sample size.

More relevant to the primary research question, the cumulative scores on three of the five tasks increased approximately half a point on the passed assessments. The failed versus the passed task mean differences on the explanation of terms, graph interpretation, and summary were .55, .53, and .52 points, respectively. In contrast, pronunciation of terms and the role-play task changed to a lesser extent with mean differences of .35 and .25 between the two assessments. The mean difference of .52 points on the summary task most strongly influenced the changed speaking performance result. Recall that the summary task rating scores are comprised of pronunciation, grammar, fluency, and comprehensibility ratings. Because the overall assessment score is comprised of a total of 10 rating scale judgments, the summary task score makes up 40% of the total score.

Table 4.1 shows the task means and task mean differences (based on 0-3 Likert ratings) between the failed and passed assessments on each of the five tasks. The weight, the percentage each task represents on the overall score, is also shown.

Table 4.1 Mean Scores by Tasks

Tasks	Failed Means	*SD*	Passed Means	*SD*	Mean Differences	Weight
Summary	2.08	0.18	2.59	0.18	0.52	40%
Pronunciation of Terms	1.90	0.39	2.25	0.42	0.35	10%
Explanation of Terms	2.00	0.24	2.55	0.37	0.55	10%
Graph Interpretation	2.13	0.29	2.65	0.29	0.53	20%
Role-Play	2.30	0.16	2.55	0.20	0.25	20%

Based on results of the repeated measure multivariate analysis of variance (MANOVA), four of the five tasks were statistically different between failed and passed assessments. With the exception of the pronunciation of terms task, the univariate results showed significant differences on each task. The *F* statistics and significance values at $p < .05$ are summarized in Table 4.2.

Table 4.2 Mean Differences on Tasks

Tasks	Mean Differences	*Df*	*F*	*p*
Summary	0.52	(1,9)	91.47	0.000**
Pronunciation of Terms	0.35	(1,9)	3.12	0.111

Tasks	Mean Differences	*Df*	*F*	*p*
Explanation of Terms	0.55	(1,9)	22.22	0.001**
Grammar	0.53	(1,9)	18.99	0.002**
Role-Play	0.25	(1,9)	15.00	0.004**

* $p<.05$, ** $p<.01$

The mean differences of the rating scale criteria for pronunciation, grammar, fluency, and comprehensibility changed by .25, .58, .35, and .57, respectively. Based on mean differences and criteria weight, rating improvement on grammatical accuracy and comprehensibility levels were the stronger rating scale criteria scores influencing the changed overall results for these 10 prospective international teaching assistants. In Table 4.3, the rating scale criterion mean values, including weights (the percentage influence on the overall EOPT score), are displayed.

Table 4.3 Mean Scores by Rating Scale Criteria

Criteria	Failed Means	*SD*	Passed Means	*SD*	Mean Differences	Weight
Pronunciation	1.93	0.18	2.18	0.14	0.25	30%
Grammar	2.15	0.34	2.73	0.30	0.58	20%
Fluency	2.30	0.26	2.65	0.41	0.35	10%
Comprehensibility	2.16	0.27	2.73	0.15	0.57	40%

The MANOVA with the rating scale criteria scores as dependent variables showed statistical difference on the rating scale criteria between the failed and passed

assessments, $F(1,9) = 14.50$, $p < .05$. Table 4.4 summarizes the univariate results showing significant differences between the rating scale criteria scores from the failed assessments versus the same criteria from the passed assessments.

Table 4.4 Mean Differences on Criteria

Criteria	Mean Differences	*df*	*F*	*P*
Pronunciation	0.25	(1,9)	9.92*	0.012*
Grammar	0.58	(1,9)	19.76*	0.002**
Fluency	0.35	(1,9)	10.76*	0.01*
Comprehensibility	0.56	(1,9)	27.69*	0.001**

*p <.05, **p<.01

Finally, the scores obtained only on the summary task were investigated in a third MANOVA. Results indicated a significant difference $F(1,9) = 29.1$, $p < .05$ between the failed and passed performance of the summary task. Three criteria, grammar, fluency, and comprehensibility, were statistically different between failed and passed assessments. The analysis showed that 85% of the variance was accounted for by comprehensibility, 64% by grammar, and 54% by fluency. The mean ratings on pronunciation; however, were not significantly different. Table 4.5 presents the mean differences and univariate test statistic results of the summary criteria differences between the set of failed and passed assessments.

Table 4.5 Mean Differences on the Criteria from Summary Task

Summary Criteria	Mean Differences	*df*	*F*	*P*
Pronunciation	0.25	(1,9)	5.00	0.052
Grammar	0.55	(1,9)	15.78	0.003**
Fluency	0.35	(1,9)	10.76	0.010*
Comprehensibility	0.90	(1,9)	52.07	0.000**

* $p<.05$, ** $p<.01$

In summary, the first analysis indicated that mean scores on the failed and passed oral proficiency assessments were statistically different. The overall score increase of the passed tests was based on a half point increase in ratings on four of the five tasks (summary, pronunciation of terms, explanation of terms, graph explanation, role-play) and all four of the EOPT's rating scale criteria. Only the pronunciation of terms task did not reach a statistical level of significance between acceptable and unacceptable proficiency. Scores on the assessment criteria of pronunciation, grammar, fluency, and comprehensibility between failed and passed assessments were statistically different. On the field-specific summary task, the rating scale criterion of pronunciation was not significantly different on this task. Increased scores on grammar, fluency, and comprehensibility influenced the score improvement from failed to passed.

Part B: Speech Analysis Results

Results of the speech analysis did not support the hypothesis that passed mean scores on grammatical, temporal, and phonological characteristics would out perform the

failed mean scores of these same features during the oral performance of the field-specific summary task. Results from five native speakers who completed the same type of summary task are included as a baseline for evaluating the performance of the non-native speakers.

Grammatical Errors

Inter-rater reliability on grammatical error counts among two ESL professionals and myself was measured using inter-class correlation coefficient and reached an acceptable level of agreement at .70. The number of grammatical errors from the failed assessments ranged between 3 and 23. In contrast, the passed assessment counts of grammar errors ranged from 8 to 16. Contrary to expectations, more grammatical errors were counted on the passed assessments (M=11.3, SD 2.56) than on the failed assessments (M= 9.9, SD= 5.60). Six individuals out of 10 had more grammatical errors on their passed assessment samples than on their failed assessment speech samples. These findings were incongruent with the first analysis results showing an improvement in rater judgments on the criterion of grammar.

Temporal Variables

Based on past research and the first investigation, reflecting improved rating scale levels on fluency judgments for the passed assessments, performance on the temporal measures on the failed and passed assessments were expected to differ. This hypothesis was not supported by the data in the current study. Relative consistency in total time of the speech samples indicated that the non-native speaker and the native speaker speech

samples were similar. Total time averaged 108.82 (*SD*=23.90) seconds for the native speakers, 113.58 seconds (*SD*=24.55) on the failed set of assessments and 105.93 (*SD*=18.74) seconds on the passed samples from the 10 international graduate students. Based on inter-class correlation coefficient, inter-rater reliability on syllable counts was .99. An inspection of the descriptive statistics did not show clear improvement on temporal measures. Table 4.6 shows the small differences between failed and passed performances as well as the native speaker baseline measures for speech rate, total pause time, articulation rate, *uhs*, and repetitions.

Table 4.6 Temporal Variable Results with Native Speaker Baseline

		Failed Non-Native *N*=10	Passed Non-Native *N*=10	Native *N*=5
Speech Rate	Mean	2.47	2.38	3.95
# syllables/total time (sec)	*SD*	0.49	0.33	0.86
	Range	1.89-3.33	1.96-3.00	2.68-4.99
Total Pause Time	Mean	32.47	32.02	24.29
Seconds	*SD*	17.38	6.370	14.27
	Range	12.13-60.13	22.17-39.27	11.23-48.70
Articulation Rate	Mean	3.40	3.47	4.99
#syllables/total time-pauses	*SD*	0.36	0.43	0.69

		Failed Non-Native *N*=10	Passed Non-Native *N*=10	Native *N*=5
	Range	2.76-3.93	2.71-4.19	4.21-5.88
Uhs, uhms	Mean	11.60	12.5	15.2
Counts	*SD*	7.12	9.85	5.16
	Range	2-27	0-32	6-27
Repetitions	Mean	20.90	12.90	7.6
Counts	*SD*	12.67	13.64	3.78
	Range	2-35	0-48	5-14

Speech rate, the variable most consistently claimed to predict fluency, changed minimally between the failed and the passed speech samples. Results for the failed assessments ranged from 1.89 to 3.33 syllables per second (*M*=2.47, *SD*=.49) whereas for the passed assessment the range was from 1.96 to 3.00 (*M*=2.38, *SD*=.33) syllables per second. Total pause time also differed minimally with failed assessments averaging 32.47 seconds (*SD*=17.38) and passing assessments averaging 32.02 seconds (*SD*=6.37) of pause time in the 2-minute samples. As for disfluencies, filled and unfilled pauses, only repetitions showed a clear difference in counts. For the failed assessment, there was an average of 20.9 (*SD*=12.67) repetitions, repairs, or restarts in contrast to the passed assessment when there were only 12.9 (*SD*=13.64), a difference of eight errors. The number of filled pauses, counts on "*uhs, ums, mmm*," were similar, tallied to range between 2 to 27 (M=11.6, SD=7.12) and 0 to 32 (M=12.5, SD=9.85) on the failed and

passed assessments, respectively. An inspection of these descriptive statistics comparing the failed and passed speech analysis results by first language groups revealed similar performances across the variables. Repetition counts, however, showed that international graduate students with Chinese as their first language made six fewer repetitions on the passed assessment when compared to the failed assessments, whereas those with Korean as their first language made 10 fewer repetitions on the passed assessments.

Table 4.7 Temporal Variable Results by First Language Groups

		Native Speakers *N*=5	Chinese Speakers *N*=5		Korean Speakers *N*=5	
			Failed	Passed	Failed	Passed
Speech Rate	*M*	3.95	2.52	2.47	2.43	2.29
#syllables/total time (sec)	*SD*	0.86	0.5	0.4	0.54	0.26
	Range	2.68-4.99	1.89-3.04	1.96-3.00	1.90-3.33	2.04-2.73
Total Pause Time	*M*	24.29	33.03	31.84	31.92	32.2
Seconds	*SD*	14.27	19.78	6.37	16.95	7.12
	Range	11.23-48.70	12.43-60.13	24.47-39.37	12.13-52.60	22.17-39.20
Articulation Rate	*M*	4.99	3.39	3.47	3.41	3.48
#syllables/total time-pauses	*SD*	0.69	0.34	0.23	0.43	0.6

		Native Speakers *N*=5	Chinese Speakers *N*=5		Korean Speakers *N*=5	
			Failed	Passed	Failed	Passed
	Range	4.21-5.88	2.92-3.73	3.26-3.84	2.76-3.93	2.71-4.19
Uhs, uhms	*M*	15.2	10.6	16.2	12.6	8.8
Counts	*SD*	5.16	6.42	11.69	8.38	6.9
	Range	6-18	2-19	0-32	7-27	0-16
Repetitions	*M*	7.6	23.6	17.6	18.2	8.2
Counts	*SD*	3.78	12.21	18.44	13.93	4.81
	Range	5-14	6-35	0-48	2-34	2-15

Phonological Variables

Failed and passed differences on segmental errors, mean length of runs, thought-groups, average number of words per thought-group, as well as average number of prominence per thought-group were minimal. Descriptive statistics presented in Table 4.8 show a comparison between the failed, passed, and native-speaker values. Results on thought-groups, mean length of runs, and average number of words per thought-groups showed that native speakers had fewer stretches of speech within each utterance, but that each stretch of speech was significantly longer than for the non-native speakers. Native speakers averaged 10 words per thought-group, almost twice that of the four to five words produced by the non-native speakers. Based on inter-class correlation coefficients,

the inter-rater reliability on segmental and prominence counts was .20 and .91, respectively.

Table 4.8 Phonological Variable Results with Native Speaker Baseline

		Failed Non-Native *N*=10	Passed Non-Native *N*=10	Native *N*=5
Segmentals	*M*	12.1	10.8	n/a
Counts	*SD*	7.38	5.73	
	Range	3-26	3-23	
Mean Length Runs	*M*	8.34	7.21	17.18
syllables/thought-group	*SD*	3.14	1.730	3.7
	Range	5.24-13.44	5.01-9.78	11.41-20.89
Thought-Groups/Utterance	*M*	4.09	4.18	3
pauses > .33	*SD*	1.34	0.82	0.57
	Range	2.10-5.90	2.70-5.50	2.10-3.60
Average # Words/TG	*M*	5.23	4.54	10.68
Counts	*SD*	1.80	1.10	2.9
	Range	3.15-8.63	3.29-6.59	7.03-15.09
Average # Prominence/TG	*M*	1.93	1.63	2.25
Counts	*SD*	0.60	0.38	0.58
	Range	.79-2.69	1.30-2.48	1.45-2.99

Part C: Multi-Case Study Findings

"Problems in all areas obscure comprehensibility." (EOPT rater comment)

Introduction of Case Studies

In conducting individual case studies on 10 participants, I aimed to discern how linguistic competency influenced the rater judgments on grammar, fluency, and pronunciation and contributed to the raters' overall impression of comprehensibility. Specifically, I wanted to know if the score improvements on pronunciation, grammar, and fluency were equivalent to the increase in comprehensibility ratings. Three scenarios emerged from the individual failed versus passed score evaluations: two individuals' pronunciation, fluency, and grammar rating improvements from failed to passed equaled their rating increase in comprehensibility; six individuals' score improvements on the criterion of pronunciation, fluency, and grammar were greater than their increase in ratings on comprehensibility; two individuals' increase in comprehensibility judgments were higher than their score improvements on pronunciation, fluency, and grammar ratings. Table 4.9 shows the rating scores on each of the assessment criterion at the time of the failed and passed assessments from three individuals representing the three scenarios just described. Judgments from both raters on the failed and passed assessments are listed as well as score improvements when the rating scores increased from the failed to the passed oral proficiency test. I selected one case from each of the scenarios to describe in this section.

Table 4.9 Rater Judgments by Criteria for Three Prospective ITAs

Students	Xu			Cha			Byoung-Hyun		
Assessment Criteria	Fail	Pass	Diff.	Fail	Pass	Diff.	Fail	Pass	Diff.
Pronunciation Total	3	4	+1	4	4	0	4	5	+1
Rater 1	1	2		2	2		2	2	
Rater 2	2	2		2	2		2	3	
Grammar Total	5	6	+1	4	6	+2	5	5	0
Rater 1	2	3		2	3		2	3	
Rater 2	3	3		2	3		3	2	
Fluency Total	5	6	+1	4	5	+1	5	5	0
Rater 1	2	3		2	3		2	3	
Rater 2	3	3		2	2		3	2	
Comprehensibility Total	2	5	+3	4	6	+2	4	6	+2
Rater 1	1	2		2	3		2	3	
Rater 2	1	3		2	3		2	3	

Xu's Case - Failed Assessment

Xu was an international graduate student from China who was majoring in physics. In August of 2005, Xu failed his first assessment by receiving a 20 overall score. He failed again in January of 2006 when he repeated the assessment and received a 22 score. On his third attempt in May of 2006, he achieved a passing score of 25.5. For the summary task, Xu summarized an article on Newton's law of motion for the first

assessment and on the passed assessment he described how the concept of light developed. I selected Xu's case for description because his comprehensibility improvement from the failed to passed performances equaled the score increases in rater judgments on the criteria of pronunciation, fluency, and grammar. Table 4.9 shows the score results on Xu's first failed and later passed assessments.

Rater judgments and comments. On Xu's initial failed assessment, both judges agreed that his comprehensibility was a level one, defined by the scoring key description as "generally not comprehensible because of frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control." On the other three criteria of pronunciation, grammar, and fluency, however, the raters disagreed. Rater 1 assigned a Level 1 rating in pronunciation, defined as "frequent phonemic errors and foreign stress and intonation patterns that causes the speaker to be often unintelligible." Rater 2 was more generous on pronunciation by assigning a Level 2, describing his speech as "some inconsistent phonemic errors and foreign stress and intonation patterns, but speaker is generally intelligible." Rater 1 noted under pronunciation, "intonation-too loud; pitched very high." Rater 2 offered no comments.

Rater 2's higher ratings, when compared to Rater 1, were reflected in grammar and in fluency judgments as well. At Time 1, Rater 1 judged Xu's grammar to be a Level 2, defined as "generally good control in all constructions but with grammatical errors that do not interfere with overall intelligibility." Rater 2 scored Xu's performance a Level 3 rating, identifying grammar errors to be "sporadic" and "minor." Fluency ratings differed

similarly where Rater 1 judged Xu's fluency to be a Level 2, defined as "some nonnative pauses that do not interfere with intelligibility" while Rater 2 chose a Level 3 rating, stating that "speech is smooth and effortless, closely approximating that of a native speaker."

Based on the pronunciation, grammar, and fluency ratings, it seemed that Rater 1 perceived Xu's skills to have impeded intelligibility while Rater 2's judgments consistently leaned toward seeing his skills as not interfering with intelligibility. Ironically, in the overall comments, hand-written at the bottom of the evaluation sheet, Rater 2 noted "extreme monotone (high nasalization) and nonnative stress makes him very difficult to understand. There is no stress within sentences and no pauses to indicate a sentence has ended. He also needs to slow down." This comment was inconsistent with the pronunciation and the fluency judgments, which rated Xu's skills as not interfering with intelligibility. Rater 1 wrote, "needs to work on moderating volume and speed and achieving a more native-English intonation in order to become more comprehensible." This comment explicitly connected the rater's low comprehensibility judgment to the linguistic feature of intonation.

In sum, the review of Xu's rater judgments and comments on his failed assessment showed somewhat contradictory information between the two raters as well as between the rating scale judgments versus the raters' handwritten comments. Although the raters agreed on a Level 1 comprehensibility, Rater 2 scored Xu a Level 2 on pronunciation and Level 3 on grammar and fluency. In spite of the higher rating,

Rater 2 noted that he was very difficult to understand in the comments at the bottom of the evaluation sheet. The rater specifically pointed to the fact that Xu lacked appropriate stress and pauses to identify sentences and that he needed to slow down. Rater 1's score of Level 1 on pronunciation and on comprehensibility was more consistent with the written explanation that non-native intonation obscured comprehensibility. Based on this information, I was curious to investigate Xu's speech analysis results, especially his temporal and his phonological variables. Both raters' comments suggested non-native stress and intonation patterns, and one rater felt Xu spoke too fast.

Speech analysis results. On Xu's first oral proficiency assessment, he produced 364 words in describing Newton's law of motion on the field-specific explanation and summary task. The amount fell close to the failed group mean that ranged between 215-586 words. The utterances used in the speech analysis are presented below. Recall that the ten utterances (T-units or intonation units) used in the analysis started after the introduction within the four-minute speech, and not at the beginning to prevent looking at possibly memorized or more rehearsed introductions.

10 Utterances from Xu's Failed Assessment

1. for example if we pull (11) the doors we add a we we had a force on the door (29)
2. uh there uh another very important ingredients is the mass (12)
3. the mass means (16) the mass means (13) uh (24) the the mass is a measure of how difficult to change the object's velocities (27)

4. there are three kinds of (20) newnewton's laws but this articles focus on the first laws (14)
5. sim simply to say the first law means that (10) the (17) the motion of the object will not change if there is no force (1;04)
6. and a uh detailed a di (12) a detailed ddddimesion is that an object moving with constant (12) vesolities (velocities) continues to move with the s the sasame speed (14)
7. and if if the object is at rest if there is no net net force it will still (17) be at rest (19)
8. uh there are two situations in the meaning of net force (10)
9. one is that there is no force on the object (20)
10. the another mean the there another situation is that (16) uh the force act on the object sums to zero (26)

I conducted a speech analysis on these utterances by coding and compiling the grammatical, temporal, and phonological characteristics.

(1) Grammar. Results from the grammar check on the failed assessment showed that Xu made 11 grammar mistakes on the two-minute speech sample. Xu's failed performance had grammatical inaccuracies in the categories of final-*s*, articles (a/the), and verbs. In Utterance 1, for example, Xu said "*the doors*" where he likely meant to say "a door" in creating a conditional sentence, explaining, "*if we pull on a door, this generates a force on the door*." He also used "*the*" in Utterance 4 where "*the object's*

velocities" was likely intended to mean a generality of "an" object's velocity. At the end of Utterance 4, Xu also added an unnecessary "s" in describing Newton's first *law,* not *laws*.

The verb errors included a subject/verb agreement, two verb tense errors and two instances of missing verbs. Utterance 4 contained an instance of a missing verb where Xu said "*this article's focus (missing verb-is) on the first laws*." In Utterances 1 and 10, the verb tenses were not clearly discernable by what was spoken. In Utterance 1, Xu probably meant to say "for example, if you pull on a door, you put a force on the door;" however, his phrasing did not include the two-part verb, *pull on*. He used "*the door*" twice, making it unclear what door he was initially talking about. Then he seemed to search for the verb to express "puts force on the door." Utterance 3 also contained a missing verb where Xu likely meant to say, "the mass is a measure of how difficult (it is) to change the object's velocity." Xu also said "*vesolities*" instead of the singular "velocity" in Utterance 3. In Utterance 6, he said "*vesolities*" instead of "velocity," which was counted as a wrong word/odd expression error although it could be argued to be a pronunciation error.

(2) Temporal variables. The 10 utterances isolated for the speech analysis consisted of 251 syllables spoken in one minute and 23 seconds with 12.43 seconds of total pause time. Speech rate was calculated at 3.04 while articulation rate, accounting for total pause time, was 3.57 syllables per second. Xu's speech and articulation rates were both faster than the failed group average of 2.47 and 3.40 syllables per second

average, respectively. In relation to the native speaker average, Xu's results were slower than the 3.95 and 4.99 syllables per second averages on speech rate and articulation rate. Xu's total pause time of 12.43 seconds was half the average for native speakers at 24.29 seconds. These results related directly to the rater comments addressing speed and pausing. Although Xu's speech rate was slower than the native speaker average, his small amount of total pause time to mark chunks of speech presumably influenced the perception of speaking to fast and the inability of the raters to discern sentences.

In terms of stretches of spoken speech produced, Xu's mean length of runs averaged 12.77 syllables. This was higher than the failed group average of 8.34 but still less than the native speaker average of 17.18 syllables. Xu uttered 7 "*uhs*" and 32 repeated syllables on the failed speech sample. Utterances 2, 3, 6, and 8 include examples of "uhs." Utterance 3 shows seven syllables counted as repetitions where Xu repeatedly restarts his sentence by stating "*the mass means / the mass means / uh / the*" before deciding on "*the mass is a measure of how difficult to change the object's velocities*." Although not counted as a temporal factor, Xu's failed speech sample also contained 24 syllables of epenthesis or insertions. These extra syllables have been added to the transcription of Utterance 3 below.

3. the mass means(zu) // (16) the mass(zu) means(zu) // (13) uh // (24) the the mass is a measure of how difficult(te) to change(e) the object's(zu) velocities (27) //

(3) Phonological variables. Xu's speech sample contained 28 segmental errors. All but one sound error were mistakes on voiced /ð/ and voiceless /θ/ "th." In other words, Xu mispronounced "*the, there, that*" in his field-specific summary task speech sample. The /æ/ in *example* was also mispronounced and sounded more like "egzompul."

In terms of chunks of language, Xu broke his utterances into an average of 2.41 thought-groups. This was fewer than the overall failed mean for the assessments of 4.09 thought-groups per utterance and fewer than the three thought-groups per utterance averaged by the native speakers. The thought-groups contained an average of 2.41 instances of prominence distributed within them, which was higher than both the 1.93 failed group average and the 2.25 average for the native speakers. Xu's prominence displayed a speech habit of stressing almost every word. The transcription of Utterance 3 with prominence marked in capital letters demonstrates this. Prominence marked on the entire transcript can be seen in Appendix E.

3. // ↗ the MASS means → // (16) the MASS MEANS → // (13) uh → // (24) the the MASS is a MEAsure of HOW difficult to change the object's veLOCities ↘ // (27)

Researcher impressions. In my impressionistic review of Xu's failed speech sample, I wrote the following: "It sounds like he's yelling – loudness and intensity instead of pitch change!" In listening repeatedly to the speech sample, the perceived loudness of Xu's speech stemmed not only from actual volume but also from his

tendency to overuse word stress without any overall pattern of intonation. He had a tendency to stress every word and generated it by using loudness and intensity on the stressed syllables instead of pitch change. Instead of using intonation, which subsumes word stress into a pattern of phrasal intonation strung together to create discourse intonation, Xu did not change pitch (fundamental frequency) across his chunks of language. His speech, therefore, was loud because he used volume instead of pitch change and because of his overemphasis on word level stress. From my review of his failed recording, I also noted that "he never falls in his intonation pattern," and this made the listener feel as though she or he must stand at military attention during the whole speech sample. Basically, his lack of pitch change across thought-groups made his intonation sound flat and monotone.

My impressions were consistent with the rater comments except in terms of stress. Recall that Rater 2 noted, "Extreme monotone (high nasalization) and nonnative stress makes him very difficult to understand. There is no stress within sentences and no pauses to indicate a sentence has ended. He also needs to slow down." By "stress," I believe Rater 2 was referring to the absence of intonation but perhaps lacked enough training in this area to distinguish the speech phenomenon and terminology. The "no pauses" was confirmed in the temporal analysis showing that Xu's speech sample included only 12 seconds of total pause time, half that of a native speaker.

Overall interpretation. Xu scored 20 out of 30 on the failed assessment. We saw how, although the raters disagreed on pronunciation, fluency, and grammar, they agreed

that Xu was a Level 1 on comprehensibility. The rating scale descriptions ascribed this level of speech as having "frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control." The speech analysis confirmed some of these issues. For example, Xu's speech sample contained 32 syllables from repetitions, 24 epenthesis (syllable insertions), 28 segmental errors, and 11 grammatical errors. The 56 extra syllables in a 2-minute sample likely contributed to the rater's perception of comprehensibility level.

Total pause time, however, was an issue in the opposite direction, as might have been expected. In other words, instead of the frequent number of pauses common in less fluent speech, Xu's speech suffered from the lack of pauses to the extent that sentences could not be discerned. His overall articulation rate, defined as the total number of syllables over total time minus total pause time, was 3.57 syllables per second compared to the native speaker average of 4.99 syllables per second; thus, although his speech was fast in comparison to the overall mean of 3.40 for failed assessments, his speech was not as fast as the native speakers' average.

The comprehensibility rating of Xu's speech likely also resulted from the lack of pitch change to create intonation and from his overuse of word stress. His word level stress was made with loudness and intensity rather than with pitch change or movement in pitch, resulting in the need to "moderate volume" as pointed out by Rater 1.

In sum, the speech analysis identified grammar errors, temporal measures, and phonological factors that seemed to have impacted the rating scale judgments.

Specifically, I believe Xu's short total pause time, extra syllables, and overuse of word stress in relation to short mean length of runs and thought-groups that lacked intonation negatively influenced the raters' perceptions of comprehensibility.

Xu's Case – Passed Assessment

Rater judgments and comments. On the passed assessment, the two raters agreed on pronunciation, grammar, and fluency but disagreed on their ratings of comprehensibility. In contrast to the failed assessment ratings, both judges on the passed assessment gave Xu a Level 2 rating for pronunciation, defined as "some inconsistent phonemic errors and foreign stress and intonation patterns, but speaker is generally intelligible." They agreed that Xu displayed Level 3 characteristics in the areas of grammar and fluency, that is, he had "sporadic and minor grammatical errors," and his speech was "smooth and effortless, closely approximating that of a native speaker." On overall comprehensibility, however, Rater 1 gave Xu a Level 2 ranking while Rater 2 scored him a 3. According to the rating scale descriptions, their perceptions differed in whether Xu was "generally comprehensible with errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing" or "completely comprehensible in normal speech with occasional grammatical or pronunciation errors." Under pronunciation, Rater 1 noted "intonation & stress unnatural." The same rater described the overall performance by noting, "His voice gets louder in the middle of the sentence & other strange places." Rater 2 commented, "Some pronunciation problems

especially stress & intonation, but he's generally comprehensible," in spite of the Level 3 rating, which defined his skills as "completely comprehensible."

Speech analysis results. On the passed assessment, Xu produced 346 words on the total speech sample describing the historical development of the concept of light for the summary task.

10 Utterances from Xu's Passed Assessment Speech Sample

1. uh (10) as we know (11) the uh light (17) is a electroma (12) uh light is a electromag (16) magnetic (18) uh wave huh (26)
2. and light is characterized by its wavelengths (14) but this concept (14) uh (15) but the development uh but the development of this concept (25) uh experiences several uh centuries (28)
3. about four four cen about four centuries years about four centuries ago (28) ummm Ga Galileo uh (1;11) uh performed experiments (17) to measure the speed of light (13)
4. of course his result is not accurate (24)
5. and then (13) new then the great a very good a very famous scientist is the New Newton (12) uh did several other experiments (14)
6. and (15) he found (1;13) light (15) is actually a mixture of all colors (1;11)
7. and he also suggested (23) light travels (13) uh trav uh he also suggest that light travels like an article uh (16) uh no no sorry (17) light light travels like a (19) particle (2;16)

8. uhm but other scientist (1;10) had a different opinion (23)

9. and (1;06) uh a famous scientist uh with the same (11) era (1;05) Wiggens (24)

10. he suggested light travels in the form of waves (26) and in 1801 (22) another scientist (11) Yang (10) confirmed (23) that (10) uh confirmed that light (17) uh (11) has some (13) feature of wave (1;15)

(1) Grammar. On Xu's 2-minute passed assessment speech sample, I counted eight grammar errors. These errors were scattered across several categories: one preposition, three final-*s*, and two article errors. Verb errors included one subject/verb agreement and one verb tense error. He stated "of course his result is not accurate" in explaining Galileo's experiments on light when he meant to say *his result was not accurate*. Xu also stated "…*this concept uh experiences several centuries*" when he presumably meant to say "this concept *developed over* several centuries." In Utterance 4, the use of present tense "*his result is not accurate*" might be confusing for a listener since Galileo's experiments clearly occurred in the past.

(2) Temporal Variables. Xu's speech rate on the passed assessment was 2.53 syllables per second, slightly higher than the mean for passed assessment of 2.38, but still lower than the native speaker average of 3.95 syllables per second. Xu's speech rate contrasted significantly with articulation rate of 3.56 syllables per second when total pause time of 33.73 seconds was excluded. Looking at overall speech rate, Xu's speech seems to have slowed down compared to the failed assessment sample. By inspecting

articulation rate, however, it was evident that rate had not changed but total amount of pause time had increased. In relation to the native speakers' 4.99 syllables per second, Xu's articulation rate of 3.56 syllables per second was slower.

In terms of disfluencies, Xu included 21 instances of "*uhs*" and 48 repeated syllables in the two-minute speech sample. The number of "*uhs*" was slightly higher than the average number of 15 instances by native speakers produced when completing a similar task. More prevalent were the number of repetitions found. Xu often restarted his phrases or rephrased his utterances in formulating what he wanted to express. Utterance 3 contained three syllables of "repetitions," showing this speech habit. In Utterance 7, we can see a combination of "*uhs*" and 10 syllables of repetitions.

1. // **uh** // (10) as we know // (11) the **uh** light // (17) is a electroma // (12) **uh** light is a electromag // (16) magnetic // (18) **uh** wave huh // (26)
3. // about(e) ***four four cen about four centuries years about*** four centuries ago // (28) ummm ***Ga*** Galileo uh // (1;11) uh performed(e) experiments // (17) to measure the speed of light // (13)
7. // and he also suggested // (23) light travels // (13) ***uh trav uh he also suggest that light travels*** like an article // (9)uh // (16) uh no no sorry // (17) light light // (6) travels like a // (19) particle // (2:16)

(3) Phonological variables. In terms of phonological production, Xu made 11 segmental errors on the passed speech sample. These included two consonant errors, the /ð/ in *the* and the final /n/ in *Newton*, in addition to 10 vowel errors. Xu had mispronunciations with the schwa; for example, saying /ə/ in the first syllable of *development* and the /ɪ/ sound in *accurate, scientist,* and *opinion*. He also pronounced *era* more like /ɨrə/ instead of /ɛrə/.

Compared to the native speaker average of three thought-groups per utterance, Xu produced an average of 4.5 thought-groups per utterance. Within each thought-group, 1.67 instances of prominence were counted across an average of 4.55 words. These numbers show that Xu broke his utterance into an average of four chunks and stressed about one and one half times in each of these stretches of speech. Xu's prominence was similar to the passed assessment average of 1.63 but much lower than the 2.25 average of the native speakers.

Researcher impressions. On the passed speech sample performance, Xu sounded like he was trying to be very careful to control his speech. He had fairly clear thought-group formation, and he appeared to incorporate some pitch change, especially in falling at the end of phrases. He had trouble, however, pronouncing the six-syllable word "*electromagnetic,*" a relevant term to his topic on the development of the concept of light. In addition, the high number of repeated syllables seemed to indicate a lack of automaticity in formulating his ideas into accurate grammatical structures and then producing them temporally and phonologically.

Overall interpretations. Xu's passed assessment case is interesting in light of the combined results of the rater judgments and the speech analysis. The rating scale judgments on pronunciation, grammar, and fluency, consistently placed him at Level 3. The speech analysis, however, showed a number of vowel errors but a fair production of thought-groups. Although eight grammar errors were counted and he produced 48 syllables of repetitions, Xu was rated a Level 3 on grammar and fluency, respectively. Looking at these results, it seems that either the judges were generous in their use of the ratings scale or Xu's errors were not considered detrimental enough to change the rating level. Vowel errors, for example, did not deem Xu's performance unintelligible, a Level 1. Similarly, the number of grammar errors across several categories could still be rated as sporadic and minor, indicated by a Level 3 rating. Most surprisingly, however, is the fact that even with 48 syllables of repetitions and 15 extra epenthesis (insertions), Xu's fluency was rated "smooth and effortless." Alternatively, perhaps these errors in relation to other competencies demonstrated in Xu's performance outweighed these errors. These issues will be examined in the failed versus passed assessment comparison.

Xu's Case- Failed versus Passed Assessment Comparison

Recall that Xu's case was selected because the failed to passed improvement in rater judgments on pronunciation, fluency, and grammar equaled the 3-point score improvement on comprehensibility. I suspect, however, that because of the rater disagreement on the failed assessment performance, it is unlikely that the equal improvement on grammar, fluency, and pronunciation was indicative of the 3-point

improvement in comprehensibility. Instead, I believe Xu's improved comprehensibility rating could be attributable to sustained speech rate with increased total pause time that created more discernable thought-group formation. In addition, the reduction in prominence and the addition of falling intonation patterns made him more understandable to the native speaker judges.

In sum, Xu's total time, total pause time, repetitions, and thought-groups increased on the passed assessment. His speech rate, segmental errors, grammar, prominence, and average number of words per thought-group decreased from the failed to the passed test performances. On the failed assessment speech sample, he had longer stretches of speech without enough pause time to mark thought-groups, as exemplified in the speech analysis and explicitly written in the rater comments stating there were no pauses to indicate where a sentence ends. Because of the important role thought-groups play in meaning making, this combination of factors detrimentally influenced the perception of comprehensibility. The increase in comprehensibility ratings occurred with specific production changes: total pause time increased, the number of thought-groups increased, while prominence and average number of words decreased, positively impacting Xu's speech.

Looking across the rater judgment and speech analysis results on Xu's failed and passed performances there exist areas both of clear and unclear changes. For example, based on the rater judgments, Xu improved in all areas: pronunciation, fluency, grammar, and comprehensibility. The speech analysis showed, however, that only certain aspects

of pronunciation and fluency changed. No significant grammar differences were found. According to the phonological analysis results, Xu's segmental error count did not change, but his increased thought-groups and decreased prominence were an improvement in relation to the combined speech attributes.

The temporal measure results did not show a rate increase between Xu's failed and passed performances. When total time and total pause time were taken into consideration, that is, by using articulation rate instead of speech rate, it is possible to establish that Xu's speed of speech were the same on his failed and passed performances, 3.57 versus 3.56 on the failed and passed samples respectively. Xu's disfluencies, however, did change. On the passed assessment sample, he produced a 22-second increase in pause time, as well as 14 more *uh*s and 14 more repetitions. In addition, the drop in mean length of runs and average number of words per thought group showed that stretches of speech decreased. The change from 2.1 to 4.5 thought-groups, with half the number of average words per thought-group, apparently improved Xu's comprehensibility rating. I believe the change in total pause time, thought-groups, and prominence variables contributed to Xu's passed assessment improvement in comprehensibility ratings.

This case study provided evidence pointing to the influence of the interrelation between temporal measures and phonological characteristics in explaining the perception of comprehensibility improvement in Xu's oral proficiency assessments.

Cha's Case – Failed Assessment

The second case study included the assessment of Cha, an international graduate student from China who was majoring in chemistry. She took her first oral proficiency screening in August of 2005, a second time in January of 2006 when she received a 24, and finally in May of 2006 attained a score of 25. On the first assessment, she summarized an article on air pollution, whereas for the passed assessment she addressed the topic of toxic materials. Like six out of the 10 cases, Cha's increase in comprehensibility ratings from a failed to a passed level was less of a difference than her improvement in rater judgments on pronunciation, grammar, and fluency.

Rater judgments and comments. Cha's failed performance was rated a Level 2 by both evaluators on all the rating scales of pronunciation, grammar, fluency, and comprehensibility (see Table 4.9). The rater comments, however, did not reflect the same type of consistency. In spite of the "generally intelligible" Level 2 rating, Rater 1 noted "stress, intonation and syllable enunciation" under pronunciation. In the overall comments, he noted "Serious problems in all areas; needs to work on English and speak up!" Rater 2 also indicated pronunciation issues by writing, "needs to open mouth wider and project more – hard to hear." This rater's overall comment on Cha's performance was that "she should take advantage of Austin's English-speaking environment to practice speaking."

Cha was interrupted two times by Rater 1 during her performance of the summary task. First, the rater asked her to repeat the title, apparently because of her pronunciation,

but also perhaps because she was somewhat soft-spoken. A few lines later, the rater stopped her and requested that she not read. Next to the summary task criteria, Rater 1 noted "cont. to read after 2 reminders! Spoke for about 1 minute total when she finished what she had written on her note sheet!" The rater's reaction seemed to indicate his impression that Cha could not summarize without reading and that she lacked the ability to speak for the required four minutes without relying on her note sheet.

Speech analysis results. In actuality, Cha spoke for 2 minutes and 3 seconds with 242 words on air pollution for her field-specific summary task. The amount of language produced fell at the lower end of the 215-586 range for the set of failed assessments. A description of her grammar, temporal measures, and phonological characteristics follow the 10 utterances used in the speech analysis.

10 Utterances from Cha's Failed Assessment Speech Sample

1. as you know the air is mostly nitrogen and oxygen (29) and (11)
2. this article is talking about air pollution (1;04) and uhh (1:02)
3. uh (15) the article can be divided into two major parts (22)
4. one is (10) pollution (11) sor... (12) pollution source (17) and the other is (11) how to maintain clean air (2;11)
5. first of all (13) as the industrial growth (1;01) growth (22) auto (10) mobiles (10) uh (1;01) family choice and uh (1;08) uh (1;10) using of (13) petroleum (17) cause people to have burning eyes and brethbreathing problems (1;09)

Interruption time: (12;27)

(can I stop you for just a moment, we don't want to hear you read, can you just explain what you read in the article.) Ok. So (19) hummm (3;21)

6. so (18) the organization is the major part of (28)the pollution source (1;07)
7. and (20) in (1;03) in ad in addition one of the most common forms of (10) part of air con (14) pollution is (1;07) particles (22) and uh (1;13)
8. the particles may be (28) sea salt elemental elemental (11) elemental carbon (19) or even small metal particles (2;15)
9. and uh next how to deal with our (11) pollu (11) polluted air (2;24)
10. industrial emissions of particles (25) can be prevented by treating your emissions (12) with (15) physical methods such as uh (24) filtration by (1;14) using fil using using some filters (23) and uh (20) separation by (2:25) hmm by chemical reaction (23) or absorption (21) by adding some (19) uh special solids (1;29)

(1) Grammar. Cha had seven grammatical errors on her failed speech sample. These included: two final-*s*, 2 articles, one verb, one connector, and one wrong word/odd expression. In Utterance 5, for example, Cha stated, "*family choice and uh uh using of petroleum cause people to have burning and brethbreathing problems*." She omitted the final "s" on *choice* and *cause*. In addition, she said "*using of petroleum*," instead of "*use*." Utterance 6 contained a wrong word choice in "*organization*" in which she might have meant to say *civilization is the major pollution source (or source of pollution).* Her use of "*the*" organization, "*the major*" and "*the pollution*" appeared to demonstrate an

overuse of the article but was only counted as one error because the structure and meaning of the sentence is unclear. Next, in Utterance 9, Cha used the infinitive form, *to deal*, instead of making a rhetorical question using *do* in "*how (do we) deal with our pollu polluted air*."

5. // first of all // (13) as the industrial growth // (1;01) *(extra article)*growth // (22) auto // (10) mobiles // (10) uh // (1;01) family choice *(dropped final-s)* and uh // (1;08) using of // (13) *(verb tense)* petroleum // (17) cause *(dropped final-s)* people to have burning eyes and brethbreathing problems //(1;09)

(2) Temporal variables. Cha's 10 utterances contained 277 syllables with 47 seconds of total pause time. Due to the pauses, her speech rate of 2.24 syllables per second varied considerably from an articulation rate of 3.61 syllables per second although both were slower than the native speaker averages of 3.95 and 4.99 respectively. In terms of stretches of speech produced, Cha's mean length of runs was 6.21 syllables compared to the set failed assessment average of 8.34 and much lower than the 17.18 native speakers' average. Her disfluencies included 12 "*uhs*" and 29 repetitions. Utterance 5 shows two repetitions when "*growth*" was said twice, and she started to say "*breathe*" but changed to "*breathing*" and produced "*brethbreathing problems*." Utterance 8 added eight repeated syllables in producing "*elemental*" twice before uttering the phrase "*elemental carbon*." Five extra syllables from "*fil using using*" in Utterance 10 were also counted in repetitions.

(3) Phonological variables. I counted eight segmental errors including six vowel inaccuracies in the following words: *parts, petroleum, breathing, organization, separation*, and *reaction*. The remaining two sound errors were in the production of /θ/ in two instances of "*with*." Cha generated an average of 5.1 thought-groups per utterance with 1.49 prominence per thought-group. The thought-groups contained an average of 3.80 words. Utterance 4 below demonstrates Cha's tendency to break her utterances into short thought-groups with one or two instances of prominence.

4. // ↗ ONE is // (10)
   polLUtion → // (11)
   sor... → // (12)
   polLUtion SOURCE ↘ // (17)
   and the OTHer is ↘ // (11)
   HOW to maintain clean air → // (2:11)

Researcher impressions. Cha's failed performance sounded flat and choppy. In my notes I wrote "she sounds like she's announcing something," meaning her speech gave a very disengaged impression. When she was interrupted by Rater 1 and asked not to read her summary, Cha said "*ok*" and then continued sounding the same as when she started the summary. In addition to the amount of speech, rate of speech, scattered grammar errors, and perception of reading, Cha mispronounced several content words. In other words, elements important in the lexicon and necessary in understanding her description, such as "*petroleum, breathing, separation,* and *reaction*," contained sound errors.

Overall interpretation. Cha's failed performance resulted from consistent rater judgments across all four rating-scale criteria. Her skills were all rated a Level 2, indicating generally good intelligibility, generally good control of grammar, and enough fluency for intelligibility, contributing to her being "generally comprehensible." In spite of the "good" rating-scale judgments, the rater explicitly mentioned Cha's loudness, stress, intonation, and syllable enunciation. The global comments of serious problems in all areas and recommendation to take advantage of the Austin speaking environment clearly express the raters' impression of insufficient skills overall.

Cha's speech analysis revealed that she was only able to produce 242 words on her total summary task, a relatively small amount for the expected four-minute summary on air pollution. In addition to the little amount of language produced, Cha's speech showed more pauses, repetitions, and thought-groups than the failed group average – each demonstrating a weak level of competence. She also had less mean length of runs, less prominence per thought-group, and fewer mean number of words per thought-group than the mean for the overall failed average. Again, these results indicated weak skills in terms of the quantity of language was produced and her ability to highlight the key points. In sum, based on the combination of skills and ratings, Cha seems to have performed sufficiently to achieve a Level 2 rating but at the same time clearly lacked competence in oral proficiency.

What Cha's case also highlighted was the fact that based on the instrument's scoring system and rating-scale criteria description, it is possible to receive a Level 2 on

each criterion, indicating a generally intelligible performance, but to fail the overall assessment. In fact, on her failed assessment, Cha received a 2 level rating from both raters on each task and criterion with the exception of comprehensibility on the role-play task, where she was rated a Level 3, completely comprehensible.

Cha's Case – Passed Assessment

Rater judgments and comments. On the passed assessment, Cha's criterion scores revealed consistent rater agreement on pronunciation, grammar, and comprehensibility, but not fluency. She was judged a Level 2 in pronunciation, reflecting "some consistent phonemic errors and foreign stress and intonation patterns, but speaker is generally intelligible." Her grammar was judged to have minor and sporadic errors, a Level 3. The raters also agreed that her overall comprehensibility was a Level 3, that is, "completely comprehensible in normal speech with occasional grammatical or pronunciation errors." On fluency, Rater 1 judged Cha a Level 3, smooth and effortless, while Rater 2 judged her skills to be a Level 2, "some nonnative pauses that do not interfere with intelligibility." Besides the fluency rating, Rater 2 noted "hesitation, repetitions." Both raters noted her mispronunciation of *lead*, but Rater 2 also noted intonation and "*th*" under pronunciation.

Speech analysis results. Cha produced 429 words in one minute and 46 seconds in summarizing her field-specific topic of toxic materials.

10 Utterances from Cha's Passed Assessment Speech Sample

1. it's very important for us to know the (14) proper properties or (13) some infor information about toxic (16) toxic uh materials (1;18)
2. uh when you (12) paint the wall (18) by some pigments (25) uh the (12) the pigments uh could be could (13) could contain a lot of (10) lead (15) compounds (22)
3. and you know the lead is very (13) toxic (24) to our body (18) so (14)
4. when the concentration is higher than 20ppm (18) it could be toxic (10) to our body (26)
5. and so because of this reason (18) uuuuh this this kind of (11) uh pigments (11) are now forbidden (15) in many kinds of (13) paint (13) pigments (23) products (1;03)
6. and another toxic (23) compound is methanol (1;19)
7. and (11) methanol is uh toxic to our eyes (1;01)
8. if you drink a bottle of (18) uh wine which is (12) is not good (20) it could contained uh methanol and alcohol (17) yeah (18) so (1;02)
9. uh (16) if you if the the wine is (18) uuuh contain more than (14) 40 percent (19) of (16) methanol (1:06) it could be damage it could damage our eyes (26) and uhm (2;23)
10. uuuh you cannot see clearly after 10 hours (1;24) of (17) drinking this wine (25) and (1;09)

(1) Grammar. Cha had nine grammar errors scattered throughout her passed speech sample. These errors fell into the following categories: 1 preposition, 2 final-*s*, 1 article, 1 verb form, 1 s/v agreement, 2 verb tense, 1 connector, and 1 plural error. In Utterance 2, Cha described painting a wall "*by*" a pigment instead of "*with*." Utterance 3 contained both an unneeded article and a connector error. There were verb errors in Utterance 8, 9, and 10. Cha said a bottle of wine "*could contained*," using an unneeded "ed" ending. In Utterance 9, she missed the final-s when stating "*the wine contain*..." and also made a verb error in stating "*it could be damage*." Lastly, in Utterance 10, she explained that "*you cannot see clearly*" when she likely meant the conditional "you would not be able to see clearly 10 hours after drinking this wine."

(2) Temporal variables. On the passed speech sample, Cha produced 236 syllables in 1 minute and 46.53 seconds with 35.73 seconds of total pause time. Her speech rate, calculated at 2.22 syllables per second, was slightly lower than the passed group average of 2.38 syllables per second. Her total pause time was slightly more than the 32.02 passed average and about 10 seconds more than the native speaker average. Regarding filled pauses, Cha's speech sample contained 13 "*uhs*" and 19 repetitions. Utterance 1, for example, contains one "*uh*" and six repetitions from restating "*proper*" in *properties*, "*infor*" in *information*, and the word *toxic*. Utterances 2 and 9 contain similar repetition errors.

(3) Phonological variables. For the passed assessment, nine segmental errors were found in Cha's speech sample. Her major error, as noted by both raters, was in pronouncing *lead* - /lɛd/ - as /lid/. Other errors were in pronouncing the final "k" in *toxic* and both the vowel and dipthong in *compound*. Note that the mispronunciation of *compound* related not only to the sounds but also to proper length of the two syllables. The primary word stress on the first syllable makes /***CAM****pound/* with the first syllable pitched higher and stretched longer. Without accuracy in either sound or syllable length, it was difficult to recognize this key word.

Cha broke her utterances into an average of 4.8 thought-groups with 1.37 instances of prominence. There were 3.29 average words per thought-group. Her performance on thought-groups was higher than both the overall passed average (4.18) and native speaker average (3.0). Her results on prominence per thought-group and the average number of words were lower than the overall passed average (1.63) and native speaker average (2.25). These results both likely negatively influenced the rater judgments. For example, Utterances 2 and 5 below demonstrate Cha's tendency to break her speech into relatively short thought-groups with one or two instances of prominence within each.

2. // uh ↑ WHEN YOU → // (12)
   paint the wall → // (18)
   by some PIGments → // (25)
   uh the → //(12)
   the PIGments uh could be COULD ↑ // (13)
   could contain a LOT of → // (10)

LEAD ↑ // (15)
compounds ↓ // (22)

5. // and SO beCAUSE of THIS reason → // (18)
uuuuh THIS THIS kind of → // (11)
uh PIGments → // (11)
are now forBIDden → // (15)
in MANY kinds of → // (13)
paint → // (13)
PIGments ↓ // (23)
PRODUCTS ↓ // (1:03)

Researcher impressions. On the passed assessment, Cha sounded like she was describing a story to a friend. In fact, she included a relevant example about how toxic materials might enter into the body while doing a lab experiment. I list these utterances used in the descriptive story below. Because they were not among the ten utterances selected for the speech analysis, however, they are not numbered.

a. one major (11) way for the (15) the meh substance (19) going to our body is (15) skin contact (1:01)

b. uh (16) if you (16) if you are doing **a** experiment (20) that (17) you don't put on the gloves (28) to protect yourself (23) the chemicals can contact your hands (16) and um (1:27) and going to your blood and damage your body (18) so(1:02)

c. uh remember to put on your gloves (11) bebefore you start a experiment (1:23)

d. and (10) another way (12) is by drinking or eating (13) theee (26) uuh toxic (18) toxic (11) um substance (21) such as (11) hummm (1:08)

e. some some students like to (19) put a cup of a cup of coke (15) on the experimental desk (20) and uh(20) he is (10) doing the (12) mixing solutions (16) but he forgot (10) which cup is (1:01)

f. you know so it could be very dangerous so remember don't drink (15) coke or some waters in in the (19) in the lab (27).

Overall interpretation. Cha's passed assessment appeared to stem from the consistent rater judgments of a Level 3 rating on grammar and overall comprehensibility. One rater also considered Cha's fluency a Level 3. More information from the raters' perspective are lacking because no additional comments were noted at the bottom of the evaluation sheet.

Results from the speech analysis in terms of grammar, temporal, and phonological characteristics showed that Cha had grammar errors scattered across various categories, a relatively slow speech rate with 19 syllables of repetitions, and rather short chunks of speech with only one or two instances of prominence.

Impressionistically, Cha's strength on her passed assessment was her rhythm and intonation combined with her conversational tone. I think the interjection of a specific lab-relevant example also showed comfort and confidence in the test setting and positively influenced the raters.

Cha's Case – Failed versus Passed Assessment Comparison

In comparing the data between Cha's failed and passed assessments, there are several outstanding aspects. First, Cha was able to produce much more language on the

passed assessment, 187 more words to be specific. In addition to the amount of language, her speech changed from sounding disengaged to sounding as if she was talking to a friend. Based on the speech analysis, Cha's grammatical, temporal, and phonological characteristics changed minimally. The biggest difference was that she dropped 11 seconds of total pause time and reduced her number of repetitions by 10 syllables.

On all the judgments from the failed assessment, the two raters were in 100% agreement. On the passed assessment, the raters disagreed on the fluency rating for the explanation and summary task. On the four other speaking tasks, the two raters disagreed on Cha's comprehensibility level on the interpretation of graph task. Rater 1, who assigned a Level 2 in fluency, rated comprehensibility in the graph interpretation task a Level 3. On the failed assessment, Rater 2 also assigned a Level 2 on this criterion. Therefore, the failed to passed rating improvement occurred from two points on grammar, one point on fluency, and two points on comprehensibility within the summary task. From these results, we can discern why Cha's overall comprehensibility rating increased less than her improvement across pronunciation, grammar and fluency. The second rater on the passed assessment noted "hesitation, repetitions," suggesting the rater was apparently more influenced by Cha's fluency. Due to the lack of additional rater comments, little else can be interpreted.

In comparing the failed and passed speech analysis results, Cha produced much more language on the second performance, as revealed in the overall word count from the transcript. There were 429 words counted on the passed compared to 242 on the failed

assessment. The most notable change in the speech variables was a drop in pause time from 47 to 35.73, as well as a 10-syllable drop in repetitions from 29 to 19. The comparison of speech rate (2.24 to 2.22) and articulation rate (3.61 to 3.33) revealed a lack of significant improvement in speed of speech. There were also very minimal differences in mean length of runs, thought-groups, and prominence between the failed and passed performances.

In spite of improvement in the two temporal variables from the speech analysis, fluency was the area that received inconsistent ratings. Apparently, one rater was more influenced by the number of repetitions and hesitations, or the raters had different interpretations of the criterion. Two utterances from the Time 2 performance definitely show instances of longer than normal pauses: In Utterance 9, Cha pauses over one minute between the clause "*if the wine*…(pause), *it could damage*..." The same utterance also contained eight syllables of repetitions in "*you if the*" and "*it could be damage*." Again in Utterance 10, she paused over a minute (1:24) in the middle of the utterances.

9. // uh // (16) if you if the the wine is // (18) uuuh contain more than // (14) 40 percent // (19) of // (16) methanol // (1;06) it could be damage it could damage our eyes // (26) and uhm // (2;23)

10. // uuh you cannot see clearly after 10 hours // (1;24) of // (17) drinking this wine // (25) and // (1;09)

Another outstanding factor between the performances that I found in listening to the recordings was the interruptions by one rater and intonation contours. On the failed assessment, when Cha stated the title of her article, she was instantly asked by Rater 1, "Can you say that again, please." The title of the article was apparently *Air Capricious Canopy*.

The exchange went like this:

Cha: Uh good afternoon ladies and gentlemen

Cha: uh I I am going to tell you about an article which is titled uh air depretious canopy

Rater 1: Can you say that again, please.

Cha: Oh ok air "represhous" canopy

After about eight utterances, Cha is again interrupted by the same rater:

"Can I stop you for just a moment, we don't want to hear you read, can you just explain what you read in the article."

Without video, it is impossible to see if Cha was actually reading. The voice recording, however, clearly reveals a flat and disengaged intonation pattern. Disengagement is cued by flat intonation, and I would guess that her increase in overall pitch movement across the thought-groups on the passed assessment was qualitatively different than on the failed assessment performance.

A significant change on the passed performance was this change in intonation contours and the more descriptive summary that included an added example. Although

not captured in the thought-group/prominence measures, it was clearly distinguishable when listening to the sound files impressionistically that pitch movement across the thought-groups on the passed assessment was more native-like. These two limitations, lack of video and instrumental analysis of fundamental frequency, will be discussed more thoroughly in Chapter 5.

In sum, I believe Cha's improvement from failed to passed on the oral proficiency assessment was influenced by what she described, in terms of the amount of language produced, the inclusion of a relevant example, and her more engaging intonation patterns. The increase of 187 words demonstrated more competence in production. She showed more comfort and confidence by adding into her summary a story about the dangers of drinking coke in the laboratory. Although not revealed in the variables measured, my impressionistic review discerned a significant change in intonation patterns throughout Cha's performance. I believe these to be significant reasons in the change from a failed to a passed oral proficiency test result.

Byoung-Hyun's Case – Failed Assessment

For the third case, I discuss the performance of Byoung-Hyun, a Korean graduate student majoring in mechanical engineering. Byoung-Hyun's first assessment was taken in August of 2005, on which he scored 21.5, and five months later in January of 2006, he passed the assessment with a 25. Byoung-Hyun's field-specific summary task topic on the failed assessment was the physics of motion, while on the passed test he described internal combustion engines. Byoung-Hyun's case was selected because rater judgments

improved only one point on pronunciation but two points on comprehensibility. In other words, Byoung-Hyun's results were the opposite of Cha's. His comprehensibility rating increases were greater than the combined increases in pronunciation, fluency, and grammar judgments. His speech analysis also demonstrated a particularly interesting interrelation among speech variables.

Rater judgments and comments. The two raters on Byoung-Hyun's failed assessment agreed on a Level 2 pronunciation and comprehensibility rating on the field-specific summary task. His pronunciation was judged to have "some consistent phonemic errors and foreign stress and intonation patterns, but [the] speaker is generally intelligible." His comprehensibility was defined as "generally comprehensible with errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing." Rater 1 judged grammar and fluency a Level 2, while Rater 2 assigned a Level 3 on the same criteria. The raters agreed on Level 2 for all other rating scale judgments on all tasks except for the comprehensibility rating from the role-play task. On the rating scale, Rater 1 ranked Byoung-Hyun's skill a Level 3 in contrast to Rater 2 who chose Level 2. Ironically, Rater 1's overall assessment comments were, "problems in all areas obscure comprehensibility. Also needs to work on volume and fluency." Rater 2 noted, "verbal hesitancy, repetitions, lack of organization indicated the need for additional language practice."

On the summary task, Rater 1 wrote "stress/r/vowel/final sounds," indicating some of the perceived pronunciation errors. Next to grammar, Rater 1 noted articles,

prepositions, and subject-verb agreement errors. On the side of the evaluation sheet, "mostly read – very brief" was handwritten. Rater 2 left no comments by the summary task criteria.

Initially curious about Byoung-Hyun's case was the inconsistency between the rating- scale judgments versus comments. Rater 2 judged fluency to be a Level 2 rating, indicating "some nonnative pauses that do not interfere with intelligibility" but then in the overall comments noted verbal hesitancy and repetitions as reason for additional language practice. Even more incongruent was Rater 2's consistent Level 2 ratings on all criteria, generally defined as intelligible or generally comprehensible in conjunction with the comment that "problems in all areas obscure comprehensibility." Based on the inspection of rater information, I wanted to evaluate Byoung-Hyun's failed performance in terms of a speech analysis of the grammatical, temporal, and phonological characteristics.

Speech analysis results. On the failed assessment, Byoung-Hyun summarized a physics related article describing how machine components transmit motion. His description contained 362 words. Following are the 10 utterances that were used in the speech analysis where he was explaining how springs, gears, and ratchets generate motion.

10 Utterances from Byoung-Hyun's Failed Assessment Speech Sample

1. for spring (23) we can learn from this article (29) uuuh (26) the various var various kinds of springs and their uh structures (1;08)

2. for a rachet (15) we can (14) uh learn each structure and each function (13) which is preventing (11) the motion from being reversed (2;20)
3. I want to ask you to see the figures of each component (23) described in this article (14) carefully (27) to make it easy to understand their structures and functions (1;14)
4. for the gear (16) it has teeth (11) uhh it can it can trantransit momotion (24) by uh (24) using the teeth (1;01)
5. and (1;28) uh for the chem (13) there is a (11) moving part (13) rotating part (15) there is one um arm (19) uh but uh because the motion is transited the motion transitted from the moving part to the arm (19) it can uh it can it can transfer the motion (1;23)
6. uh (13) for the linkage there (17) at least three part of (12) uh three part of components (19) uh (1;12)
7. it can use it can be used can be used uh (1;28) uh to (1;02) transfer the rotating motion to uhh (1;00) uhhh (1:07) uhh (17) up down motion (20) uh (21)
8. and (13) for the spring (18) I think you can (29) you you already know the structure of the spring but/ (1;14) uh (10) we don't have (10) we don't just have that (13) uh (13) uh (11) spiral uh shape uh (11) spring (17)
9. there are lots of springs (20) like the play spring (1;01) like this (22) uh (1;10)
10. for the rachet (11) actually I'm not familiar with rachet (21) before reading this article (18) but (32) umm (17)

(1) Grammar. In the two-minute speech sample, five grammatical errors were counted on the transcript of Byoung-Hyun's summary. Two final "s" errors were made, along with one missing article, one verb tense error, and a missing verb. Utterance 6 shows an example of a missing verb error where "*are*" was omitted and final-*s* on "*part*" was dropped.

(2) Temporal variables. Byoung-Hyun's speech rate was calculated at 2.35 syllables per second with 46.10 seconds of unfilled pause time. Accounting for total pause time, articulation rate was 3.59 syllables per second. Both speech rate and articulation rate were slower than the overall failed assessment average (2.47 and 3.40 syllables per second) and the native speaker averages (3.95 and 4.99 syllables per second). Byoung-Hyun's other disfluencies included filled pauses of 27 uhs and 30 repeated syllables. Byoung-Hyun's repetitions included mostly restarting his utterance as demonstrated in Utterances 4 and 7.

The mean length of runs was 5.52 syllables, produced between pauses of .33 seconds, indicating relatively short stretches of speech. This was less than one third the length (17.18 syllables per thought-group) produced by native speakers. Utterance 5 shows Byoung-Hyun's tendency to produce short chunks of speech.

(3) Phonological variables. On Byoung-Hyun's failed assessment speech sample, ten segmental errors were counted. Eight were voiced or voiceless "th" mistakes. He consistently used a "d" when saying words with voiced "th" such as *this*, *the*, and *their*.

Voiceless “th” errors, /θ/ were found in the initial position of *think* and final position of *teeth*. In addition, Byoung-Hyun mispronounced the short /ɪ/ in “*this*.”

In terms of stretches of speech between pauses, Byoung-Hyun broke each of his utterances into an average of 5.8 thought-groups. As indicated by the mean length of runs, Byoung-Hyun seemed to lack the ability to produce consistently longer chunks of language without pausing or repeating. His short thought-groups included an average of 3.78 words with 1.33 instances of prominence. The native speaker samples averaged 10.68 words with 2.25 instances of prominence. Utterance 4, 5, and 6 below show this pattern.

4. // → for the GEAR ↗ // (;16)
   it has TEETH ↗ // (;11)
   uhh IT can IT can TRANTRANsit moMOtion → //(;24)
   by uh → //(;24)
   USing the TEETH → // (1;01).

5. // → And // (1;28)
   uh for the CHEM ↗ // (;13)
   THERE is a → // (;11)
   MOVing part → // (;13)
   ROtating part → // (;15)
   there is one um ARM → // (;19)
   uh but uh because the MOtion is transited the motion transitted from the Moving part to the ARM // (;19)
   it can uh it can it can transFER the MOtion ↘ // (1;23)

6. // → Uh // (;13)
   for the LINKage there → // (;17)

at least three part of → // (;12)
uh THREE part of components → // (19)
uh → // (1;12)

Researcher impressions. Three prominent factors were outstanding on Byoung-Hyun's failed performance. First, after about three sentences of the summary task, Byoung-Hyun stopped and said, "*That's it*." After encouragement from the evaluators, he continued. His speech, however, contained a noticeable number of repetitions in restarting or repairing his phrasing. This number of repeated syllables was disruptive both in terms of flow and expression of ideas. The description also seemed very simple in grammatical structure and ideas communicated. For example, in seven of the 10 utterances, Byoung-Hyun started out by saying, "*For X*...":

1. "For Spring…"
2. "For a rachet…"
4. "For the gear…"
5. "And / (1;28) uh for the chem…."
6. "Uh / (13) for the linkage…"
8. "And / (13) for the spring…"
10. "For the rachet…"

Overall interpretation. In sum, Byoung-Hyun's failed rating was likely influenced by stopping in the middle of the performance, showing lack of language competence to complete the task regardless of whether it was related to test anxiety, lack

of language confidence, unfamiliarity with the topic, or task difficulty. His rate of speech, pause time, and repetitions demonstrated weakness in fluency. His short phrasing and simple description likely also influenced the raters' judgment, disclosed as "needing additional language practice" and "problems in all areas obscure comprehensibility."

Byoung-Hyun's Case – Passed Assessment

Rater judgments and comments. On Byoung-Hyun's passed assessment, he summarized an article on combustion engines using 255 words. The raters agreed on a Level 3 comprehensibility rating. They disagreed, however, by one point on pronunciation, grammar, and fluency ratings. Rater 1 scored Byoung-Hyun's performance a 2level on pronunciation and Level 3 on grammar and on fluency. Rater 2 indicated a better performance on pronunciation through a Level 3 rating but marked a Level 2 on grammar and fluency. Rater 1 noted "articles" next to the grammar criteria but left no other comments. Rater 2, who rated Byoung-Hyun a Level 3 on pronunciation, wrote "nice-careful" beside this criterion. This same rater wrote, "very good tho slow," beside the comprehensibility criteria. In the overall assessment comments, Rater 2 also noted "some misuse (overgeneralization) of passive structures. Speak more energetically!"

The rater judgments assigned a Level 3 on comprehensibility, indicating "completely comprehensible in normal speech with occasional grammatical or pronunciation errors." In terms of pronunciation, grammar, and fluency, the raters

seemed to have been influenced differently by Byoung-Hyun's performance because their ratings differentiated by one point for each of these criteria.

Speech analysis results. Byoung-Hyun used 255 words to describe internal combustion engines in his field-specific article task. The 10 utterances used for the speech analysis were extracted from just over 74 seconds of the summary.

10 Utterances from Byoung-Hyun's Passed Assessment Speech Sample

1. there are two kind of internal combustion engines (2;06)
2. one is gasoline engine (12) and the other is diesel engine (2;05)
3. in gasoline engine (13) we use (21) gasoline (23) asa asa fuel (2;06)
4. one of the most significant differences of these engine (13) with diesel engine is the method of ignition (1;22)
5. in gasoline engine (1;04) spark ignition method is used (2;17)
6. the advantage of gasoline engine (10) is (10) light (16) and (16) value (1;24)
7. however (16) it doesn't show (23) high efficiency (20) because of low (19) comcompression ratio ratio (1;26)
8. that's why gasoline engine is used for small cars (2;10)
9. in diesel engine (21) we use we use (11) diesel oil (12) as fuel (1;15)
10. the ignition method in diesel engine is compression ignition (2;25)

(1) Grammar. On Byoung-Hyun's speech sample from the passed assessment, 13 grammatical errors were counted. Nine of the 13 errors fell into the dropped final *–s* category. There were also three missing articles and one subject-verb agreement mistake

counted. Final-*s* was consistently dropped on the term "engine(s)" in Utterances 1, 4, 5, 6, 8, 9, and 10. Byoung-Hyun missed the use of an article in Utterance 2 in stating, "*one is (a) gasoline engine (12) and the other is (a) diesel engine*. The subject verb agreement error was in Utterance 8 where he said *"that's why gasoline engine is used*…"

(3) Temporal variables. On Byoung-Hyun's passed assessment, his speech rate was calculated at 2.30 syllables per second while articulation rate was 3.92 syllables per second. Byoung-Hyun's rate of speech fell below the native speaker average of 3.95 speech rate and 4.99 articulation rate averages. His total pause time was close to the passed group average of 32.02 seconds but higher than the native speaker average of 24.29 seconds within their speech samples.

In the category of filled pauses, Byoung-Hyun produced zero "*uhs*" and only eight repetitions in the 10-utterance speech sample. On these variables, his performance contrasted to the performance by the native speakers, who produced an average of 15.2 "*uhs*" and 7.8 repetitions. He produced 8.77 syllables per second between pauses of .33 seconds, showing the mean length of runs close to the failed average of 8.34 syllables per second but significantly less than the 17.18 native speaker average.

(3) Phonological variables. Nine segmental errors were counted on Byoung-Hyun's passed set of utterances. Four of the sound errors were "th," and the other five were vowel sound inaccuracies. The "th" errors were in pronouncing *these* and *method*. The vowel errors included the /ə/ in *method* multiple times and the /æ/ in *advantage*.

Byoung-Hyun produced an average of 2.6 thought-groups per utterance with 1.88 instances of perceived prominence across an average of 5.20 words per thought-group. Although his thought-group count was close to the native speaker average of three thought-groups per utterance, his average number of words was half the 10.68 average produced by the native speakers.

Byoung-Hyun's prominence counts in Utterances 1 through 4 are shown below using capital letters within each thought-group.

1. // ↗ THERE are two kind of intERnal combustion engines →// (2;06)

2. // ↗ ONE is gasoline engine → // (;12)
   and the other is DIEsel engine ↘ // (2;05)

3. // ↗ IN GASOline engines → // (;13)
   we use → // (;21)
   gasoLINE → // (;23)
   asa asa FUEL ↘ // (2;06)

4. // ↗ ONE of the MOST SIGnificant DIFFerences of this engine → // (;13)
   with DIEsel engine is the METHod of ignition ↘ // (1;22)

Researcher impressions. My impression of Byoung-Hyun's passed performance speech sample was that he seemed very deliberate and controlled in trying to make his speech clear and fluent. I also noted that he checked for understanding by explicitly asking the raters, "Can you understand what I presented so far?" He also used organizers as shown in Utterance 1 and Utterance 4. The expressions "*the most significant*

different" and "*the advantage of gas engines is*" show examples of effective formulaic language.

1. // There are two kinds of internal combustion engines // (2;06)

4. // One of the most significant differences of these engine // (13) with diesel engine is the method of ignition // (1;22)

Overall interpretations. I believe Byoung-Hyun's passed score came from several factors. Rater 1 scored each of the comprehensibility criteria on four different tasks a Level 3, that is, "completely comprehensible in normal speech with occasional grammatical or pronunciation errors." The same rater also rated Byoung-Hyun's grammar and fluency a Level 3. Rater 2 scored Byoung-Hyun's pronunciation and comprehensibility criteria on the summary task a Level 3. Pronunciation was rated a Level 3 on the pronunciation of terms task although rated Level 2 in the role-play.

By examining the speech production characteristics, Byoung-Hyun produced only 255 words but performed well in creating thought-groups per utterance and instances of prominence. I believe that his organizing language and comprehension check added to the impression of oral proficiency competence.

Byoung-Hyun's Case – Failed versus Passed Assessment Comparison

Rater judgments and comments. Based on the difference in rater judgments between the failed and passed tests, Byoung-Hyun's performance on the summary task improved by two points in comprehensibility and one point on the pronunciation criteria.

The comparison of grammar errors showed that although more were found in the passed assessment, the type of error, 9 of 13, was predominantly final-*s* and missing articles. The consistency in a 2 level rating indicates that although errors were present, the rater still perceived the speaker as having "generally good control in all constructions with grammatical errors that do not interfere with overall intelligibility." It is possible that the final-*s* errors were perceptually received differently by the different raters, which could explain the difference in rater judgments on grammar, which varied one point on the failed performance.

A comparison of the temporal measures indicated that Byoung-Hyun produced less language (362 vs. 255 words) in less time and with less pauses on his passed assessment. The number of repetitions decreased drastically from 30 to 8 while mean length of runs increased from 5.52 to 8.77 syllables per thought-group. This change was also reflected in the intonation measures. The reduction in number of thought-groups with an increase in average words per thought-group show an increase in stretches of language produced. Prominence also increased slightly from 1.33 to 1.88 per thought-group. Although an improvement, the variable seems to capture an improvement relative to the increase in stretches of language produced. Emphasis on almost two syllables out of a 9-syllable run of speech would more closely emulate native speaker production (assuming they were content words). I suspect Byoung-Hyun's prominence application was a stable variable and simply showed an increase due to his improvement in producing longer language chunks.

Rater judgments on grammar and fluency did not change between the failed and passed assessments. There were, however, more grammar errors on the passed performance and significantly fewer repetitions. It is somewhat surprising that the fluency rating did not improve, given that disfluencies have been found to predict perceptions of fluency. This could be attributed to rater calibration, as Byoung-Hyun's fluency rating initially was scored a 3 by Rater 1 even though hesitancy and repetition were pointed out in the comments. The generous ratings on the failed assessment might have created a ceiling effect, which could explain why fluency judgments did not improve in spite of improvement in temporal measures shown in the speech analysis.

The change in pronunciation ratings could not be attributed to improvement in segmental production captured in the speech analysis because the number of inaccuracies was very similar on both performances. The rater change in comprehensibility and pronunciation, however, does seem to have been influenced by the improvement in intonation variables, specifically in thought-groups and prominence. Byoung-Hyun's thought-groups per utterances changed from 5.8 to 2.6, with 1.33 and 1.88 change in prominence. The improvement in mean length of runs from 5.52 to 8.77 and increase in the average number of words per thought-group from 3.78 to 5.20 add evidence to the increase in stretches of language. Because speech rate and articulation rate remained relatively equal, Byoung-Hyun's score improvement cannot be attributed to simply faster speaking.

The reduction in repetitions likely contributed to improved rater perceptions of pronunciation and comprehensibility. It is possible that two primary interferences in the failed assessment were number of repetitions and restricted ability to produce more than five-syllable stretches of speech, capping the perception of comprehensibility. Repetitions might be considered inherently distracting, but they are also disruptive to the grammatical, temporal, and phonological systems. Utterance 7, "*it can use it can be used it can be usually used*," demonstrates the working out of the sentence structure out loud. When the repetition occurs, there is interference in constructing a thought-group and adding prominence needed for meaning making. As demonstrated in Utterance 7, a 15-syllable thought-group was produced, but the repetitions generated prominence on the same word three times. Because prominence is used to highlight content or information, the repetitions created "empty" prominence that would likely detract from meaning making and the perception of comprehensibility.

7. // ↗ it CAN use it CAN be used it CAN be USually used uh → // (1;28)

Finally, without a doubt, there were additional factors influencing the failed and passed performances. Byoung-Hyun stopped on the initial assessment after only a few minutes and said, "*That's it*." In addition to the reduction in repetitions and longer stretches of speech, Byoung-Hyun on the passed performance demonstrated more sophisticated performance by checking for comprehension and using organizing statements. He stopped and asked, "*Can you understand what I have presented so far?*" with rising intonation. In contrast to the failed performance where he began each

explanation with "*For the cam,….for the linkage, ….for the rachet…*," in the passed performance Byoung-Hyun stated, "*There are two kinds of internal combustion engines*" as well as adding "*the most significant different*" and "*the advantage of gas engine is…*" Although content knowledge or ability to teach is proclaimed not to be a component of the assessment, oral proficiency cannot be totally divorced entirely from the type of discourse constructed.

In fact, in the overall assessment comments, one rater on the failed assessment stated "problems in all areas obscure comprehensibility; needs work on volume and fluency." The speech analysis demonstrated how repetitions were indeed disruptive to "all areas;" that is, repeating words impaired grammatical structures, temporal measures, and phonological accuracy, specifically the production of thought-groups and placement of prominence. We can interpret the rater as having been influenced by the grammatical, temporal, and phonological aspects of this international student's speech. The second rater noted that verbal hesitancy, repetitions, and lack of organization indicated the need for additional language practice. This rater seemed to have been most distracted by these factors.

On the passed assessment, the raters agreed on the Level 3 of comprehensibility in spite of disagreeing consistently between ratings of Level 2 or Level 3 on pronunciation, grammar, and fluency. The passed assessment raters judged pronunciation to be a Level 3, while commenting, "misuse of passive structures" and "Speak more energetically!" Apparently, energy level was perceived and deemed felt to be important. The rating

judgment on pronunciation, however, did not reflect this lack, either because it was not considered to be part of the assessment criteria or because it was not influential enough to downgrade the rating. It is possible that the "energy" level would be considered a "tone of voice" issue, but it could also be explained by a common characteristic of Korean speakers: narrow pitch range or not enough prominence being discernable by the listener. Prominence changed slightly, but not to the extent needed in the improvement of mean length of runs and average number of words. A fundamental frequency analysis would be needed to determine whether the intonation improved and whether it was within a narrower pitch range than is common for a native speaker.

CHAPTER 5: DISCUSSION

The goal of this study was to examine 10 prospective international teaching assistants' oral proficiency assessments on which each individual had initially failed but within the same year retook the assessment and received a passing score. This set of archived data provided a unique opportunity, through a failed/passed comparison, to address the nature of oral proficiency components as defined through a currently used ITA oral proficiency screening instrument. The aim of the investigation was to explore how a candidate's grammatical, temporal, and phonological characteristics produced in the test performance impacted the instructor rater judgments on the assessment criteria of pronunciation, grammar, fluency, and comprehensibility.

This study contributes to several lines of research relevant to the practical issue of measuring comprehensible speech as well as theoretical inquiry on the nature of oral proficiency components. The results of the case study analysis address a question Munro and Derwing (1999) raised in asking which particular aspects of non-native English speech influences a listener's perception of comprehensibility. I investigated this through a grammatical, temporal, and phonological speech analysis from a candidate's test performance and the instructor rater judgments on grammar, fluency, pronunciation, and comprehensibility from the assessment instrument. Through these data, I could explore how deviation in segmentals and the prosodic features of thought-groups and prominence influenced a native speaker's judgment of pronunciation as Anderson-Hsieh, Johnson, and Koehler (1992) called for. Using results from the speech analysis in a case study

analysis afforded a window into how a combination of the grammatical, temporal, and phonological variables, and how their interaction influenced the perceptions of second language fluency and comprehensibility. Kormos and Denes (2004) pointed out the need for this type of investigation in fluency research.

I carried out the investigation using a three-part mixed method design to explore (a) the instructor rater judgments on the assessment's five task and four rating scale criteria, (b) results of a grammatical, temporal, and phonological variable speech analysis, and (c) how the specific speech variables might have influenced rater judgments on the assessment criteria of pronunciation, grammar, fluency, and comprehensibility in a case by case analysis. Each step focused on a failed/passed assessment comparison made possible through a unique set of archived data for five Chinese and five Korean graduate students undertaking the oral proficiency screening process at the Research 1 institute where the study was undertaken.

In Part A, I ran several MANOVA investigations to determine how the set of failed and passed assessments differed according to the rating scale scores on the five tasks and the four assessment criteria. Based on the results of the first investigation, I selected one speaking task from which to extract a 2-minute speech sample in order to explore the candidate's grammatical, temporal, and phonological variables within both the failed and passed assessment performances. When in Part B mean score differences on linguistic variables from the set of failed and passed assessments resulted in no differences suggesting an insufficient method and unit of analysis, I turned to individual

case studies to focus each fail/pass analysis at the individual level and to encompass all data at hand for each individual's two assessment performances. Within each case study, I was able to analyze the instructor rater judgments and comments, results from the speech analysis in relation to the assessment criteria scores, as well as additional testing factors found in the failed and passed oral proficiency assessments for these 10 prospective international teaching assistants.

In this chapter, I discuss the research findings on the instructor raters' perceptions of comprehensibility in the context of oral proficiency assessment for international teaching assistants and the limitations of the study. I also address directions for future research on the interrelation among speech variables on listener perception of comprehensibility and the use of pronunciation as a speaking assessment task and criterion.

Discussion of Study Findings

Differentiating Failed/Passed Oral Proficiency Assessments

My initial investigation on the rating scale judgments from the set of failed and passed oral proficiency assessments revealed that scores on four of the five tasks differentiated the performances. The set of passed assessment scores could be attributed to improvement on the following tasks: summary of an article, explanation of terms, interpretation of graph, and roleplay. Scores on the task of pronouncing a list of field-specific terms did not increase between the failed and passed assessment performances. By comparing the set of failed/passed scores on the assessment criteria of pronunciation,

grammar, fluency, and comprehensibility, I found that the set of passed assessments resulted from improved rater judgments on all the criteria. As previously explained, because the summary task had been evaluated on all four of the assessment criteria, it was selected for the speech analysis. Yet, I found that rater judgments on grammar, fluency, and comprehensibility differentiated the failed from the passed performances on the summary task but pronunciation ratings did not differ statistically.

The results of the speech analysis in Part B failed to support the hypotheses that the set of passed oral proficiency assessments had resulted from better performance on grammatical, temporal, and phonological variables. The investigation, however, highlighted the need to look at the data on a case by case basis in order to analyze the interrelation among the rater judgments and comments, the speech analysis variables, as well as how the relation between these had changed on the failed and passed evaluations. Consequently, I added Part C to the study where I conducted 10 case studies on each individual's data. The results of the case study investigation showed three areas of findings (1) individual explanations for the difference between failed and passed assessments, (2) an interrelation among speech variables that seemed to influence the instructor raters' perceptions of comprehensibility, and (3) measurement issues in relation to the EOPT, the oral proficiency assessment instrument on which the data were based and the pass decision had been made.

Individual results between the failed and passed performances. In each case study investigation, I attempted to answer for each candidate how their failed and passed

oral proficiency assessment performances differed for the purpose of better understanding what impacted the instructor rater judgments in scoring the evaluations. For this set of 10 prospective ITAs, the improvement to a passed assessment could be attributed to two main factors that generally appeared to have positively influenced the instructor rater judgments across the 10 case studies: (a) an improved interrelation among several speech variables and/or (b) discourse complexity. In this section, I describe how the interrelation among three variables, the temporal variables of total pause time, and mean length of run, with the phonological variable of thought-groups, led to improved rater perceptions of comprehensibility.

As described in Chapter 4, when analyzing Xu's case study, I found that an increase in total pause time, reduction in mean length of runs, as well as prominence per thought-groups positively influenced the change from failed to passed speech samples from the summary task. In the second case study described in Chapter 4, Cha's results showed minimal differences in the speech variables between her failed and passed assessments outside of a 10-syllable reduction in repetitions. The impressionistic review of her speech samples, however, revealed several key changes between the assessments. Cha produced almost twice as much language on the summary task for the passed test than on the failed test. Although not reflected in the speech analysis variables, Cha demonstrated a more sophisticated passed performance by including a relevant and amusing anecdotal example as well as improved intonation patterns. In contrast to Cha, Byoung-Hyun's case showed that fewer total number of words produced on the summary

task with increased accuracy in production was also possible. Byoung-Hyun's performance improvement seemed to stem from the drastic reduction in repetitions, and total pause time in combination with an increase in mean length of runs. The comparison showed that Byoung-Hyun did not say more in terms of quantity, in fact he said less, but demonstrated improved temporal and phonological features. I also believe Byoung-Hyun's textual organization and comprehension checking demonstrated discourse sophistication and likely influenced the raters.

Interrelation among speech variables. The most significant finding in this study in terms of instructor rater perceptions of the comprehensibility criterion was evidence of an interrelation among total pause time, mean length of runs, and thought-groups that appeared to have a positive impact on the rating scale judgments. Recall that non-fluent speech is replete with stops and therefore, as fluency develops, second language speakers typically have less total pause time in their speech. In addition, more developed speakers are able to produce longer stretches of speech without stopping, and consequently reduced total pause time and increased mean length of runs have become common markers of fluency in second language studies. As described in Chapter 2, thought-groups refer to the chunks of language typically to which intonation patterns are applied and are marked by pause boundaries. The native speakers in this study averaged three thought-groups per utterance with 10.68 words per thought-group (see Tables 4.6 and 4.8). Because thought-groups are a feature of prosody, this phonological measure is not commonly included in studies on fluency. As Kormos and Denes (2004) pointed out,

"...fluency research suffers from the lack of studies that investigate a combination of linguistic, temporal, phonological and interactional variables" (p.146). Using the speech analysis results in the case studies addressed this need and revealed and interrelation among total pause time, mean length of runs and thought-groups.

Xu and Byoung-Hyun's cases demonstrated how performance on specific temporal and phonological variables functioned together to influence rater perceptions of comprehensibility. In Xu's failed case, his rate of speech with lack of pauses, long mean length of runs, and strong prominences per thought-group combined to make him unintelligible, as indicated by the Level 1 rating on comprehensibility from the failed assessment. The interrelation among these variables was further verified on the passed speech sample in that Xu's speech rate remained the same on both the failed and passed speech sample but his total pause time more than doubled on the passed speech sample. His mean length of runs shortened in addition to a reduced count of prominence. These changes in speech variable performance seemed to have made a positive impact on the comprehensibility judgments, which improved from a Level 1 rating from both raters on the failed performance to rating Levels of 2 and 3 by the two evaluators on the passed performance.

Byoung-Hyun's case demonstrated a similar interrelation among the same speech variables of total pause time, mean length of runs, and thought-groups, but in the opposite directions. Recall from that in Byoung-Hyun's failed performance, one instructor rater's comment noted that all areas impeded comprehensibility and pointed out verbal

hesitancy. Byoung-Hyun's speech analysis results did indeed confirm a high amount of total pause time and 30 syllables of repetition. In addition, he had short mean length of runs that created a high number of thought-groups per utterance with little prominence. The change in these variables on the passed speech sample showed that total pause time and repetitions decreased while mean length of runs increased and thought-groups per utterance decreased. Byoung-Hyun's comprehensibility ratings on the summary task increased from a Level 2 on the failed performance to a Level 3 on the passed performance with both raters in agreement each time.

In several of the 10 case studies, the speech analysis on the 2-minute sample of the summary task showed that a drop in total pause time occurred between the failed and passed performance. In one of these cases, the reduction in total pause time occurred with an increase in mean length of runs and fewer thought-groups per utterance. The longer stretches of speech meant fewer stops, thereby creating a drop in the number of thought-groups, marked by pauses within utterances. The opposite combination also occurred in two cases as exemplified in the description of Xu's case. Xu, a native Mandarin speaker, as well as another student from Korea had less than half the amount of total pause time than a native speaker on their failed performances. Both of their speech analysis results from passed performances revealed not only an increase in total pause time but also shorter mean length of runs and an increased number of thought-groups per utterance. I believe these examples to be evidence of an important interrelation between

total pause time and mean lengths of runs as predictors of fluency and connecting them to thought-groups, a fundamental feature of prosody upon which intonation is applied.

My findings extend the understanding of the role of temporal variables in combination with phonological variables influencing listener perceptions of comprehensibility. First, speech rate, a common predictor of fluency, did not differentiate the failed and passed oral proficiency performances. All the speech samples in this study were within 1.89-3.33 syllable per second compared to the native speaker samples which ranged between 2.68 and 4.99 syllable per second. As Griffith (1991) pointed out from pausology research, articulation rate of the speech samples was useful because it accounted for total pause time. Recall that speech rate differs from articulation rate in that the later accounts for total pause time. By analyzing the combined sources of information from the changed speech analysis results and instructor rater judgments between the failed/passed speech samples, I discovered the role of total pause time and mean length of runs with thought-groups coincided with improved rater judgments on overall score improvement to a passing level.

Fluency research has shown that pause time and mean length of runs affect rater impressions of perceived fluency (Kormos & Denes, 2004). In addition, findings from conversation analysis research has shown that frequency and type of hesitation phenomena, frequency of repair, amount and rate of speech, and other interactive features contribute to perceived fluency (Wennerstrom, 2000). It has also been found that listeners' perceptions of comprehensibility have a clearer tie to fluency judgments than

do their assessments of accentedness (Derwing et al., 2004). Wennerstrom (2000) found a .85 correlation between SPEAK ratings of fluency and comprehensibility and claimed "that the fluency score is a strong component of overall language ability in the perception of raters" (p. 107).

Based on the evidence from my data, Wennerstrom's (2000) results showing a high level of correlation between fluency and comprehensibility judgments could be due to the interrelation among pauses, mean length of runs, and thought groups. I believe frequent pauses inhibit the production of thought-groups, the fundamental prosodic unit on which intonation is applied. As revealed in the case studies, mean length of runs or stretches of speech that are too short or too long impeded the formation of these meaningful units of English. The lack of appropriate pause placement, therefore, inhibits the creation of though-group formation, the verbal handing out of syntactic chunks of information within an utterance, expected by a listener. I believe this extends our understanding of why a certain level of fluency is needed to generate comprehensible speech. The temporal variables of total pause time and mean length of runs punctuating speech in terms of the chunks of language produced are required for the application of intonation that adds meaning to a phrase. In sum, the temporal variables of pause time and mean length of runs with the prosodic feature of thought-groups in this study demonstrated how the temporal and phonological systems of spoken language converge in comprehensible speech. The findings showed that an interaction among these variables influenced the instructor raters' perceptions of comprehensibility.

Measurement Issues on the EOPT. The case studies addressed at the individual level the difference in failed and passed oral proficiency performances but also brought to my attention two measurement issues. First, because each performance was rated by two judges, disagreement between raters highlighted potential problems in inter-rater reliability. Second, there was incongruence between some of the instructor rater judgments and their handwritten comments that made it seem as though the raters' interpretation of the criteria descriptions differed from how they were actually using the rating scale levels to score the oral proficiency performances.

Interrater reliability on the EOPT instrument, based on a year of assessments, was calculated at .78 inter-class correlation coefficient. The judgment variation between the pair of raters judging the failed performances as well as the pair on the passed assessment challenged my analysis in examining the change in scores across the failed and passed assessment criteria. Because I had initially grouped the individual cases in terms of how their results on pronunciation, fluency, and grammar scores related to comprehensibility ratings, I did not find clear results implicating these judgments as equally contributing components of comprehensibility. On Xu's failed assessment, for example, the raters agreed on only one out of four criterion judgments. Ironically, the criterion agreed upon on the failed assessment, comprehensibility, was the only criterion of four not agreed on by the two evaluators on the passed assessment (see Table 4.5). The change between failed and passed comprehensibility was due to the one point rater disagreement on the passed assessment. Cha's case demonstrated the highest level of rater agreement while

Byoung-Hyun's was somewhat mixed. Due to variation in the rater judgment, therefore, I do not believe the change in rating scale scores of pronunciation, fluency, and grammar were representative of the score change on overall comprehensibility.

I also found incongruence between the rating scale numerical values representing the EOPT's pre-defined criterion levels and the content of the evaluators' hand-written comments. Xu, for example, was judged a 2 on pronunciation and a 3 on fluency but one rater noted that lack of stress and pauses made him "very" difficult to understand. In spite of being "very" difficult to understand, the rater gave Xu a rating that has descriptions of pronunciation and fluency as "generally intelligible" and "smooth and effortless." One possible reason for the inconsistency between the rating and comment might have been in this case due to the lack of pauses, not an excess, that made Xu difficult to understand. With a 0-3 rating scale, it was not possible to downgrade the temporal aspect of speech in the reverse direction to identify that Xu's speech was fluent in the sense of producing continuous stretches of speech, but not comprehensible because the listener could not identify thought-groups. There were consistent examples in the data where the rating scale score defined a skill as "intelligible" but the handwritten comments identified the lack of skill competence. The results of the 10 case studies raised questions regarding interrater reliability, interpretation, and application of the rating scales as well as rater calibration on the instrument used in this study.

Pronunciation in Oral Proficiency Assessment for ITAs

Having summarized and discussed my main findings, I turn now to a discussion of pronunciation in oral proficiency assessments. Research has demonstrated that oral proficiency testing involves both the evaluation of a construct as well as a method of assessment both of which contribute to variance (Chalhoub-Deville, 1995). Pronunciation, as a component of oral proficiency and as an assessment criterion, demonstrates this issue of variance through the complexity in (a) defining the phonological aspects of speech that impact rater perceptions of comprehensible speech, and (b) measuring these variables accurately and consistently. As demonstrated in the review of literature, *pronunciation* encompasses both segmental and suprasegmental aspects of speech. By suprasegmental, I refer to prosodic features defined as but not limited to intonation, rhythm, and stress. These phonological variables serve an integral role in non-native English speaker communication, but less firmly established is the influence of each feature in rater judgments of pronunciation and comprehensibility, or even grammar and fluency as assessment criterion. Anderson-Hsieh, Johnson, and Koehler (1992) pointed out the psychometric limitation in judgments of pronunciation from speech samples used in the SPEAK Test because they failed to identify whether the native speaker judges reacted equally to deviance in all major areas of pronunciation or whether areas carried different weights in influencing the scores that were assigned. In addition, it has also been found that oral proficiency ratings are context-specific with regard to tasks and rater groups. It is not uncommon, for example, for instructor raters to

focus more on the form, while non-teacher raters focus more on content of what is performed. According to Chalhoub-Deville (1995), a rater's orientation to an assessment criterion and its relevant features may be deep-seated and may involve substantial reinterpretation of the criteria not intended by the test developers.

In the current study, pronunciation as an assessment task and criterion on the EOPT showed mixed results in this set of 10 failed and passed oral proficiency assessments. The pronunciation of terms task failed to differentiate the set of 10 failed and passed tests. Pronunciation as a rating scale criterion applied to judging the field-specific summary, pronunciation of terms, and roleplay tasks, however, was statistically different between the set of failed and passed assessments. However, pronunciation as a criterion on just the summary task did not differentiate the set of failed and passed performances. In other words, the instructor rater judgments on pronunciation contributed to overall test score changes between the failed and passed assessments but on the summary task, pronunciation ratings did not differentiate between the failed and passed performances.

Findings provided evidence of the discontinuity among the EOPT rating scale judgments on pronunciation and what the instructor raters noticed through their handwritten comments about pronunciation. Although the number of instructor rater comments dropped from 78 on the failed set of assessments to 38 on the passed set, a higher percentage of pronunciation comments were found on the evaluations sheets from the passed assessments, with 58% of the comments from the passed assessment

identifying aspects of pronunciation in contrast to 50% of the comments from the failed set of assessments. These handwritten comments included notes on errors of specific sounds, words that were mispronounced, as well as non-native like syllable structure, stress, and intonation patterns observed during the assessment performance.

On Xu's passed evaluation, both raters left similar observations that he had unnatural stress and intonation. On Cha's failed assessment, one rater noted "stress & intonation/syllable enunciation," and on the passed assessment, one rater again noted "intonation" as well as "hesitancy, repetitions." Byoung-Hyun's failed assessment received a 2 level on both pronunciation and grammar but one rater wrote "stress/r/vowels/finals sounds" beside the pronunciation criterion and also identified articles, prepositions, and subject-verb agreement beside the grammar criterion. Again, we see that although a level 2 is defined as generally intelligible, the rater concluded in the overall comments that "problems in all areas obscure comprehensibility" and that he needed to work on volume and fluency.

In terms of segmentals, there seemed to have been a qualitative difference in how mispronunciations affected the raters. Five of the 10 individuals in the case studies had more segmental errors on their passed than failed assessment speech sample. In some cases, the high segmental count was due to consistent mispronunciation of one or several sounds such as "th" (voiced and/or voiceless) or vowel sounds, which were not made note of in the rater comments. Through analyzing the speech analysis results by individual case studies, I noticed that the evaluators seemed to note more often distracting

errors or those that impeded understanding the topic of speech. If the sound error was made on the summary task topic, the error was documented in the instructor rater comments. In one case study, a student repeatedly referred to his topic as /ɛdz/ when he was talking about the disease AIDS. Another example of an error on an important concept word was Cha's pronunciation of "lead" as /lid/ instead of /lɛd/ in describing toxic materials, which both instructor raters made note of in their comments.

On the set of passed assessment performances, over half of the comments were related to pronunciation. In addition, the contrast between an "intelligible" rating scale score but contrasting rater comments drew attention to whether the ESL instructor raters could "understand" and judge performances as comprehensible even with a high degree of deviation in a candidate's phonological system because of their continued exposure and experience with non-native speech. It is conceivable that inexperienced or non-teacher raters would not have judged the speaking performances as comprehensible.

In light of contemporary assessment trends embracing authentic and communicative tasks (Bachman, 1991; Hoekje & Linnel, 1994), reading a list of isolated terms could be viewed as a less than optimal assessment task for prospective ITAs. Because pronunciation is embedded in communicative competence, I believe it is not necessary to be assessed as an isolated task. Instead, speaking assessments could increase validity by further developing pronunciation as a criterion. A holistic rating scale score of pronunciation in the evaluation of sounds, intonation, stress, and rhythm was insufficient for the EOPT results in this study. I believe pronunciation as a speaking

assessment criterion is too broadly defined and the measurement too vague for valid and reliable measurement within oral proficiency assessment instruments. Although fraught with complexities, grappling with a more thorough definition of phonological competence and especially the prosodic features in listeners' perceptions of comprehensibility has promise in allowing for a better evaluation of these phenomena within oral proficiency assessments.

I would argue that because errors in prosodic features have been found to be more influential in the perception of comprehensibility, a finer gradation of pronunciation as an evaluation criterion is needed. Such a change in this assessment criterion has the potential to allow (1) a more accurate measure of comprehensible speech, (2) to provide more detailed feedback to test-takers who fail, and (3) to increase pedagogical focus of these features in "washback" or the impact of testing on the instruction that leads up to them (Bachman, McNamara, 1996; Saif, 2006).

Limitations

First Investigation

Although useful to the current study's intended purpose of evaluating change in rater judgments between the failed and passed set of assessments, the first investigation was constrained psychometrically by the small sample size and population representing only five native Mandarin and five native Korean speaking prospective ITAs. I, therefore, interpreted the first investigation findings only as indices of the instructor

raters' perceptions of speaking ability as measured through the English Oral Proficiency Test for prospective ITAs at a large Research 1 university.

Speech Analysis

Results of the speech analysis failed to confirm all hypotheses about the speech analysis scores on grammatical, temporal, and phonological variables in comparing the passed and failed speech samples. I believe the lack of statistical results in Part B was ultimately due to insufficient measurement techniques and the statistical method used including the level of analysis. One factor causing the limited finding could have been due to inadequate measurement using only two-minutes from a 4-minute performance on only one of five tasks. Another factor pointed out in the rationale for Part C was the limitation of collapsing the data into mean scores across all 10 individuals. In other words, comparing the speech analysis results between the set of 10 failed and passed speech sample performances was insufficient because of the individual level of variation in the speech variables in addition to the interrelation among the variables unique to each individual's speech. Below, I discuss more specific measurement issues by speech variable category.

Grammar. In Part B, mere counts of grammatical errors were used, rather than type of grammatical error, thereby treating all grammatical mistakes as one variable and assuming that all errors had impacted instructor rater judges in the same way. From the case studies which included grammatical error categories, it seemed that raters were more influenced by verb tense errors than article usage. Although the counting and

categorizing of the grammatical errors on the speech sample allowed me to include a look at the role of grammatical accuracy played in the speaking task, discourse structure was not included. The measurement in this study failed to incorporate the complexity of sentence structures generated, the level of lexical use, organization markers, and instances of formulaic language use, for example. Using discourse analysis would have been more useful in evaluating these aspects of language that research has shown to impact communication (Hoekje & Linnel, 1994; McNamara et al., 2002).

Temporal and Phonological Variables. Coding and scoring the linguistics variables of speech samples was tedious and time consuming (Kormos & Denes, 2004). It took four months of steady work to review the 20 2-minute speech samples used in this study. The analysis, however, went beyond the rating scale judgment technique often used in second language research and offered more real-time data on the linguistic features that influenced the instructor rater judgments of grammatical, temporal, and phonological variables. Conducting the same speech analysis on native speaker speech samples also provided guidance in interpreting the results on each variable, especially because this group of prospective international teaching assistants would naturally be compared to native speakers by native-speaking undergraduate students.

As described in the rationale for Part C, collapsing the speech analysis results into mean scores failed to show directionality changes and interaction among the speech production variables. There were unexpected directional changes on the speech variables, specifically speech rate, total pause time, thought-groups, and prominences all

of which caused the fail/pass comparison of the speech analysis to be insufficient and inappropriate.

Although the speech analysis went beyond global rating scale judgments of the speech variables commonly used in past research studies (Anderson-Hsieh et al., 1992), the phonological category of variables was somewhat rudimentary when recognizing their complex multi-dimensional nature. This study only included two features of prosody, thought-groups and prominence, and I measured them impressionistically. Consequently, the phonological variables in this study were limited both in the number of features included and how they were evaluated. I continue this discussion in the next section on directions for future research.

Future Research

Impressionistic Versus Instrumental Analysis of Prosodic Features

This study included impressionistic measures of thought-groups and prominence. These were valid phonological variables to include in an oral proficiency speech analysis because of the role prosodic features play in meaning making. The production of a thought-group is a prerequisite to the application of an intonation pattern across a chunk of spoken language. Performance on prominence as demonstrated through pitch change, syllables duration, and loudness, is a key component in sentence stress and in creating the overall rhythm of American English. In spite of my earlier arguments in favor of impressionistic over instrumental methods in studying prosodic features, however, I have changed my stance on this issue. Although the impressionistic investigation allowed for

a larger sample size and greater total number of speech variables counted, my experiences in this study demonstrate a compelling need for the addition of instrumental analysis on prosodic features of native and non-native speakers of American English. Specifically, future research should combine impressionistic and instrumental analysis on prosodic features to measure pitch range, fundamental frequency (intonation contours), thought-groups, and prominence. In several cases, the non-native speaker speech was described as flat, and several were asked by the raters to stop reading their notes and instead, to speak without reliance on their notes. Although video might show whether or not the candidate was actually reading their notes instead, it is also possible that these individuals were using intonation patterns more common to reading practice. It is not uncommon for Korean native speakers to use a narrow pitch range, causing intonation movement across a thought-group to be less perceptible by a native listener. I have experienced this in practice, but realize the need for instrumental analysis to measure pitch range and intonation contours to test this hypothesis.

In spite of 10 years of experience in accent modification and training of non-native English speakers to use an American intonation pattern, I found it very difficult to distinguish nuclear stress from prominence. I had initially thought that I would be able to identify the intonation pattern, and therefore the utterance peak, nuclear stress. I switched to simply identifying prominence impressionistically by recognizing pitch change, loudness, and intensity. Instrumental analysis could verify these impressions and

further inform the acoustic properties that create the perception of prominence and nuclear stress.

I believe a valuable line of future research on prosodic features must include both instrumental and impressionistic studies on specific features of prosodic as Pickering (2001) exemplifies in a research study on tone choice. In addition more research is needed to encompass (a) how prosodic features interact with each other, (b) how the features interact with syntactic, temporal, and discourse level variables, and (c) how non-native English speaker prosody influences native listeners' perception of comprehensibility.

Conclusion

Performance testing has been a consistent strand in second language testing for the last 40 years (McNamara, 1996) in response to the demand for foreign students who study at Anglophone universities and the need to bring testing in line with theories of communicative competence. The demand for reliable and valid oral proficiency assessments for prospective international teaching assistants continues.

In this study, I compared a set of failed and passed oral proficiency assessments of 10 prospective ITAs for the purpose of better understanding the nature of oral proficiency components as defined by the measurement instrument. Specifically, I investigated aspects of what Bachman (1990) termed *grammatical knowledge*, the performance of grammatical, temporal, and phonological variables, and evaluated how these variables influenced rater judgments of the assessment criteria of pronunciation,

grammar, fluency, and comprehensibility. The goal was to understand how the production of the speech variables from a non-native English speaker impacted the native listeners' perception of comprehensibility within an oral proficiency assessment.

The study resulted in two substantive findings in showing an interrelation among temporal and phonological variables that influenced instructor rater judgments on comprehensibility. In addition, the case study analysis highlighted the pronunciation criterion as being ineffective in representing important phonological features of speech and in the raters' application of the assessment criterion. These results contribute to ESL research that has suffered from a lack of studies going beyond rating scale measurement of speech production variables and how these variables interact. In addition the results from the native speaker speech samples provided evidence of the skill level to which non-native English speaking TAs are naturally compared. Finally, connecting rater judgments with actual speech variables brought us closer to understanding the aspects of what influences perceptions of comprehensibility for prospective ITAs.

In spite of the advancements in speaking assessment research, Bachman (1991) proposed that more progress is needed in the application of theoretical models of language proficiency to the design and development of language tests. More specifically, McNamara (1996) pointed out that more research is needed on the scale-rater interaction as well as the effectiveness of rater training. Wennerstrom (2000) made the same observation regarding the fluency description on the SPEAK test, claiming "while such descriptions may be appropriate to instruct test raters to think holistically in the

judgments of fluency, we are left without a clear notion of what linguistic features actually correspond to these general characteristics" (p. 103). In other words, the rating scale descriptions based on overall impressions of the assessment criteria fail to illuminate what phenomenon compels a rater's judgment. Pronunciation as an assessment criterion of the EOPT showed similar limitations. In this study pronunciation was too broadly defined and too simplistic for valid and reliable assessment of this essential aspect of speech.

Further research is needed on the interrelation among grammatical, temporal, and phonological variables as well as how they combine to influence perceptions of oral proficiency and specific assessment criteria. I recommend instrumental and impressionistic analysis of prosodic features to increase our knowledge of how these important features function in communication. Finally, language testing should investigate a more complex evaluation for the criterion of pronunciation.

As Luoma (2004) stated "when people hear someone speak, they pay attention to what the speaker sounds like almost automatically. On the basis of what they hear, they make some tentative and possible subconscious judgments about the speaker's personality, attitudes, home region and native/non-native speaking status" (p. 10). In a time of globalization and World Englishes, I believe there is compelling reason for continued research on the perception of comprehensibility for non-native English speakers.

Appendices

Appendix A: English Oral Proficiency Test

International Teaching Assistant and Assistant Instructor
ENGLISH ORAL PROFICIENCY TEST

Name: _____ Date: _____

Student #: _____

Dept: _____ Rater: _____

	(LOW) .....
A. Warm-up	
B. Explanation and Summary of an Article	
_____ 1. Pronunciation	0.....1.....2.....3
_____ 2. Grammar	0.....1.....2.....3
_____ 3. Fluency	0.....1.....2.....3
_____ 4. Comprehensibility	0.....1.....2.....3
C. Pronunciation of Terms	
_____ 1. Pronunciation	0.....1.....2.....3
D. Explanation of Two Terms	
_____ 1. Comprehensibility	0.....1.....2.....3
E. Interpretation of a Graph	
_____ 1. Grammar	0.....1.....2.....3
_____ 2. Comprehensibility	0.....1.....2.....3
F. Classroom Announcement Role-Play	
_____ 1. Pronunciation	0.....1.....2.....3
_____ 2. Comprehensibility	0.....1.....2.....3

Total: _____ (x 10) = Final Score:_____ Category: _____

Comments:_____

_____

_____

Category 4:	250 — 300
Category 3:	230 — 249
Category 2:	200 — 229
Category 1:	0 — 199

TS:jss/930210\DRWI

Appendix B: ITA Assessment Scoring Key

Pronunciation

0 – Frequent phonemic errors and foreign stress and intonation patterns that cause the speaker to be usually unintelligible. (15-40 unintelligible words, part C)

1 – Frequent phonemic errors and foreign stress and intonation patterns that causes the speaker to be often unintelligible. (7-14 unintelligible words, part C)

2 – Some consistent phonemic errors and foreign stress and intonation patterns, but speaker is generally intelligible. (3-6 unintelligible words, part c)

3 – Occasional nonnative pronunciation errors, but speaker is always immediately intelligible. (0-2 unintelligible words)

Grammar

0 – Virtually no grammatical or syntactical control except in simple stock phrases.

1 – Some control of basic grammatical constructions but with major and/or repeated errors that interfere with intelligibility.

2 – Generally good control in all construction with grammatical errors that do not interfere with overall intelligibility.

3 – Sporadic minor grammatical errors

Fluency

0 – Speech is so halting and fragmentary or has such a nonnative flow that intelligibility is virtually impossible.

1 – Numerous nonnative pauses and/r a nonnative flow that interferes with intelligibility

2 - Some nonnative pauses that do not interfere with intelligibility

3 – Speech is smooth and effortless, closely approximating that of a native speaker.

Comprehensibility

0 – Overall comprehensibility is too low in even the simplest type of speech

1 – Generally not comprehensible because of frequent pauses and/or rephrasing, pronunciation errors, limited grasp of vocabulary, or lack of grammatical control.

2 – Generally comprehensible with errors in pronunciation, grammar, choice of vocabulary items, or infrequent pauses or rephrasing.

3 – Completely comprehensible in normal speech with occasional grammatical or pronunciation errors.

Appendix C: EOPT Proficiency Descriptions for Selected Score Ranges

SCORE RANGE	
300	The speaker is always comprehensible with perhaps occasional nonnative pronunciation errors or sporadic minor grammatical errors that do not interfere with intelligibility. Speech closely approximates that of a native speaker.
250 – 295	The speaker is almost always comprehensible with occasional nonnative pronunciation errors or sporadic minor grammatical errors that rarely interfere with intelligibility. Speech is smooth and effortless, and communication is very effective.
230 – 245	The speaker is usually comprehensible with errors in pronunciation, grammar, word choice, or pauses or rephrasing that generally do not interfere with intelligibility. Communication is generally effective.
200 – 225	The speaker is somewhat comprehensible with consistent, distracting errors in pronunciation, grammar, word choice, or nonnative pauses that sometimes interfere with intelligibility. The speaker struggles with the language needed to communicate his ideas.
Below 200	The speaker is generally not comprehensible because of frequent pronunciation errors and foreign stress and intonation patterns, lack of grammatical control, limited grasp of vocabulary, and numerous pauses and/or rephrasing that often interfere with intelligibility. Communication is not effective, and the listener is left confused.

userstiepterryitaprofdesc101701

Appendix D: EOPT Rater Instructions

TIEP ITA/IAI English Oral Proficiency Assessment

A. Warm-up.

(Get envelope from student. Insert cassette into recorder. Attach mike to machine and student. Turn the recorder on. Make sure the cassette recorder and the mike are *turned on* and red/green lights are flashing. If not, replace the microphone battery; replacement in the plastic baggie.)

Introduce yourself and the other rater. Brief chit-chat, such as:

"Hello, how are you? Where are you from? How long have you been in Austin?

THEN: *We need to get some information from you before we begin the test.*

What's your name?

*(**DO NOT ASK: How do you spell your name? Have you taken this test before?)*

What is your student ID number?

*What department are you in?**

*In which department do you plan to teach?"**

(*If a discrepancy, allow the student to choose the Pronunciation of Terms list.)

Complete the top of the rating sheet.

Read the following directions to the examinee:

"In this test, you will be able to demonstrate how well you speak English. Your teaching ability and your knowledge of your field will NOT be evaluated. The test has five different sections and special instructions will be given for each section."

B. Explanation and Summary of an Article.

Tell the examinee:

"In the first section of the test, you will be asked to explain and summarize the contents of the article you read during the last half hour. You will have up to 4 minutes to summarize the article. If you don't finish in 4 minutes, it's okay; there's no penalty for not completing the summary. We need only a sample of your speaking. You may use your notes for reference, but don't read them to us; speak to us as you would to a class. You may sit or stand, however you feel more comfortable. You may also use the chalkboard briefly if you wish for formulas or diagrams, but for this oral exam we prefer to have you speak rather than write. Speak as accurately about the article as you can. Do you understand? Please begin now."

*Refer to the yellow Scoring Key throughout the exam.

*If the examinee says very little, ask him/her to elaborate.

*Feel free to ask for clarification if you don't understand something.

*If the examinee is doing too much reading of notes or writing on the board, redirect him/her:

*"Excuse me, please remember to **speak to** us; don't just read to us. OR: Excuse me, please use the board only **briefly** for diagrams; we need to hear an oral summary of the article."*

*Stop the examinee after **4 minutes**; interrupt if necessary.

***Collect** the examinee's **notes** (white or yellow sheet) when he/she finishes; compare the article title to the actual oral summary. If there's a major discrepancy, report it to Ann or Estherlene after the test.

*If the examinee has tested before, check the title of the previous Graph(s) and Role Play(s) to avoid repetition.

C. Pronunciation of Terms:

Read to the examinee:

"In this section, you will be asked to read aloud a list of 40 words and phrases from your field. You will be scored for proper pronunciation only. First you will be given up to one minute to read the words silently. Now, look at the list and begin reading the words silently."

*Give the examinee the list. (Don't allow him/her to make marks on the list.)

*After one minute, say: *"Now begin reading the words out loud."*

*Collect the paper from the examinee.

D. Explanation of Two Terms.

Read to the examinee:

"*In this section, I will give you <u>three</u> terms from your field of study. You must choose <u>two</u> of these terms to define. Imagine that you are explaining the terms to a first-year class. Give a simple, basic definition, perhaps with an example. Be sure to speak as clearly and as accurately as you can. The three terms are _____, _____, and _____. Choose two of them.*

(Student chooses terms.) *Now, please define _____."*

*When the examinee finishes, say, *"Thank you. Now, define _____."*

*The examinee may use the board if he/she chooses, but don't suggest it.

*Don't lower the comprehensibility score in the event of incorrect definitions.

E. Interpretation of a Graph.

Read to the examinee:

"In this section of the test, you will see a graph. We will not see the graph. You will be asked to first describe the graph to us and then to explain or interpret the information presented in the graph. There are some questions at the bottom of the page for you to answer. You may look at the graph while describing it. Speak as accurately and in as much detail as you can. You now have up to one minute to study the graph silently."

*Give the examinee the graph.

*After one minute: *"Now begin describing the graph to us."*

*Examinees may NOT use the board during this section.

*Collect the graph from the examinee.

F. Classroom Announcement Role-Play.

Read to the examinee:

"In the last section of the test, you will see some class information and you will be asked to explain it. Imagine that you are the instructor for this class. You are meeting your students for the first time, and you need to explain this information to them. Your students do not have a copy of it. Be sure to include all of the information in your description. Do not just read the information printed, but play the role of the teacher; present it as if you were talking to a group of students. You will be able to look at this sheet while presenting the information. You now have up to one minute to study the information silently."

*Give the Classroom Announcement sheet to the examinee.

*After one minute, say:

"Remember to include all of the information in your description of the class. Please begin your description now."

*Examinees may NOT use the board during this section.

*Collect the Classroom Announcement sheet from the examinee.

G. Wrap-up.

Say to examinee: *"You're finished with the exam. Please return to room 212, where you began, and we'll bring your results to you in a few minutes. Thank you."*

*Turn off tape recorder. Disconnect mike from examinee.

*Complete rating sheets. Clearly circle individual scores and mark your total score, and be sure to write your name at the top of the sheet. Average the two scores and write the averaged score on one rating sheet and also on the envelope label (in red and circled).

*Write the titles of the Graph and Role-Play on the side of the examinee's Article Summary notes if he/she scored below 250 average.

*Record the student's name, ID#, individual rater scores, and averaged score on the blue ITA Rater Record sheet in your folder.

* Put the two rating sheets, the Article Summary notes, and the cassette in the examinee's envelope.

*One rater returns the envelope while the other keeps an eye on the room. Do not leave the testing materials unattended.

*Turn off the microphone at the end of the day.

*Return all testing materials to room 212B closet.

*Return the completed blue or green ITA Rater Record Sheet to Jim Stelling.

THANK YOU!

TIEP\a\acadcoor\ITA\RaterInstructions\072805

Appendix E: Xu's Prominence Transcripts

Failed Assessment
Total Time: 00:01:22;21

1. // → for eXAMple if we pull → // (11)
   the DOORS we HAD a we we had a FORCE on the DOOR → // (29)

2. // → uh there uh aNOTHer very important ingredients is the MASS ↗ // (12)

3. // ↗ the MASS means → // (16)
   the MASS MEANS → // (13)
   uh → // (24)
   the the MASS is a MEAsure of HOW difficult to change the object's veLOCities ↘ // (27)

4. // → there are THREE kinds of → // (20)
   newnewton's laws but THIS articles FOCus on the FIRST laws ↘ // (14)

5. // ↗ SIM SIMply to say the FIRST law means that → // (10)
   the → // (17)
   the MOtion of the object will NOT change if there is NO force → // (1:04)

6. // ↗ and a uh DEtailed a di → // (12)
   a DEtailed ddddiMENsion is that an obJECT moving with constant ↗ // (12)
   veSOLities (velocities) conTINues to move with the s the saSAme speed → // (14)

7. // ↗ and if if the obJECT is at REST if there is NO net net FORCE it will STILL // (17) BE at REST ↘ // (19)

8. // → uh there are TWO situations in the MEANing of NET force → // (10)

9. // ↗ ONE is that there is NO force on the obJECT → // (20)

10. // ↗ the aNOTHer mean the there aNOTHer situation is that → // (16)
    uh the force act on the object SUMS to ZEro → // (26)

Passed Assessment
Total time: 00:01:56:07

1.
// ↗ Uh → // (10)
as WE know → // (11)
the uh LIGHT ↗ // (17)
is a eLECtroma → // (12)
uh LIGHT is a eLECtroMAG → // (16)
magNETic → // (18)
uh WAVE huh ↘ // (26)

2.
// ↗ and LIGHT is CHARacterized by its WAVElengths → // (14)
but THIS concept // (14)
uh → // (15)
but the deVELopment uh but the deVELopment of this concept → // (25)
uh exPERiences several uh CENturies → // (28)

3.
// ↗ aBOUT four FOUR cen about FOUR centuries years about FOUR centuries ago // (28)
ummm GA GALileo uh → // (1:11)
uh perFORMed exPERiments → // (17)
to MEASure the speed of LIGHT → // (13)

4.
// ↗ of COURSE his REsult is not ACcurate → // (24)

5.
// ↗ and THEN → // (13)
NEW then the GREAT a VERY good a VERY famous SCIENtist is the NEW NEWton → // (12)
uh did several OTHer exPERiments → // (14)

6.
// → and → // (15)
he FOUND ↗ // (1:13)
LIGHT ↗ // (15)
is ACTually a MIXture of ALL colors → // (1:11)

7.
// and he ALso suggested → // (23)
LIGHT travels → // (13)
uh trav uh he ALso suggest that light travels like an ARticle uh → // (16)
uh no no sorry ↘ // (17)
LIGHT LIGHT travels like a → // (19)
PARticle ↘ // (2:16)

8.
// → Uhm but OTHer scientist → // (1:10)
had a DIFFerent opinion → // (23)

9.
// → and → // (1:06)
uh a FAmous scientist uh with the same → // (11)
era → // (1:05)
WIGgens → // (24)
HE suggested light travels in the FORM of WAVES → // (26)

10.
//→ and in EIGHTeen-o-one (1801) → // (22)
aNOTHer scientist → // (11)
YANG ↘// (10)
conFIRMED ↘// (23)
that → // (10)
uh confirmed that light → // (17)
uh → // (11)
has some → // (13)
FEAture of wave → // (1:15).

References

Abraham, R., & Plakans, B. (1988). Evaluating a screening/training program for NNS teaching assistants. *TESOL Quarterly, 22*, 505-508.

ACTFL. (2007). *American Council on Teaching of Foreign Language*. Retrieved June 18, from http://www.actfl.org/i4a/pages/index.cfm?pageid=3348

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. *Language Learning, 42*(4), 529-555.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*, 561-613.

Asher, R. E., & Simpson, J. (Eds.). (1994). *The Encyclopedia of Language ad Linguistics* (Vol. 6). Oxford: Pergamon Press Ltd.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4), 671-704.

Bachman, L., & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Birdsong, D. (Ed.). (1999). *Second Language Acquisition and Critical Period Hypothesis* (1 ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Brazil, D. (1997). ***The communicative value of intonation in English*****. New York, NY: Cambridge Univeristy Press.**

Bresnahan, M., & Kim, M. (1993). Predictors of receptivity and resistance toward international teaching assistants. ***Journal of Asian Pacific Communication, 4*****, 3-14.**

Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and affective response toward accented English. ***Language & Communication, 22*****, 171-185.**

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

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. ***Language Learning, 45*****(2), 251-281.**

Cruttenden, A. (1986). ***Intonation*** **(2nd Edition ed.). Cambridge: Cambridge University Press.**

Crystal, D. (1997). ***The Cambridge Encyclopedia of Language*****. Cambridge: Cambridge University Press.**

Cutler, A., & Ladd, D. R. (Eds.). (1983). ***Prosody: models and measurements*****. Heidelberg: Springer-Verlag.**

Derwing, T., Munro, M., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. ***Language Learning, 48*****, 393-410.**

Derwing, T., & Rossiter, M. (2002). ESL learners' perceptions of their pronunciation needs and strategies. ***System, 30*****, 155-166.**

Derwing, T., Rossiter, M., Munro, M. J., & Thomson, R. (2004). Second language fluency: judgments on different tasks. *Language Learning, 54*, 655-679.

Egbert, J. L., & Petrie, G. M. (2005). *CALL Research Perspectives*. Mahwah, NJ: Erlbaum Associates.

ETS. (2005). *TSE Details: Learners and Test Takers*. Retrieved Retrieved November 8, 2005, from http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=e9242d3631df4010VgnVCM10000022f95190RCRD&vgnextchannel=2f587f95494f4010VgnVCM10000022f95190RCRD

Flege, J. E. (1981). The Phonological basis of foreign accent: a hypothesis. *TESOL Quarterly, 15*, 443-455.

Flege, J. E., Munro, M. J., & Mackay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of Acoustic Society of America, 97*, 3125-3134.

Fox, A. (2000). *Prosodic features and prosodic structures*. Oxford: Oxford University Press.

Gass, S., & Selinker, L. (2001). *Second Language Acquisition: An Introductory Course*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Gorsuch, G. (2001). Testing textbook theories and tests: the case of suprasegmentals in a pronunciation textbook. *System, 29*, 119-136.

Gorsuch, G. (2003). The educational cultures of international teaching assistants and U.S. universities. *TESL-EJ, 7*(3).

Griffith, G. (1991). Pausological research in an L2 context: A rationale, and review of selected studies. *Applied Linguistics, 12*, 345-364.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Halliday, M. A. K. (1989). *Spoken and Written Language*. Oxford: Oxford University Press.

Hillocks, G. (1986). *Research on Written Composition: New Directions for Teaching*. Urbana, IL: Eric Clearinghouse.

Hoekje, B., & Linnel, K. (1994). "Authenticity" in language testing: evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*(1), 103-126.

Jones, R. H. (1997). Beyond "listen and repeat": pronunciation teaching materials and theories of second language acquisition. *System, 25*, 103-112.

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145-164.

Ladd, D. R. (1996). *Intonational phonology*. New York: Cambridge University Press.

Lennon, P. (1990). Investigating fluency in EFL: A qualitative approach. *Language Learning, 40*, 387-417.

Lennon, P. (2000). The Lexical Element in Spoken Second Language Fluency. In H. Riggenbach (Ed.), *Perspectives of Fluency* (pp. 25-42). Ann Arbor: University of Michigan Press.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring Second Language Performance.* New York: Longman.

McNamara, T., Hill, K., & May, L. (2002). Discourse and Assessment. *Annual Review of Applied Linguistics, 22*, 221-242.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*, 159-182.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 49 (Suppl. 1)*, 285-310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition, 23*, 451-468.

Myers, S. (1995). Using written text to teach oral skills: an ITA training class using field-specific materials. *English for Specific Purposes, 14*, 231-245.

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

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*(2), 233-255.

Pierrehumbert, J., & Hirshberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In R. Cohen, Morgan, J., & Pollack, M. (Ed.), *Intentions in Communication* (pp. 271-311). Cambridge, MA: The MIT Press.

Piske, T., Mackay, I., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics, 29*, 191-215.

Reed, B. S. (2006). *Prosodic Orientation in English Conversation*. New York, NY: Palgrave Macmillan.

Riggenbach, H. (Ed.). (2000). *Perspectives on Fluency*. Ann Arbor, MI: The University of Michigan Press.

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of international teaching assistants. *Canadian Journal of Applied Linguistics, 5*(1-2), 145-167.

Saif, S. (2006). Aiming for positive washback: a case study of International Teaching Assistants. *Language Teaching, 23*(1), 1-34.

Skehan, P. (1998). Processing perspectives on testing. In *A cognitive approach to language learning* (pp. 153-183). Oxford: Oxford University Press.

Snow, D. (2001). Transcription of suprasegmentals. *Topics in language disorders, 21*(4), 41-51.

Stake, R. (1995). ***The Art of Case Study Research*****.**

Tench, P. (1996). ***The Intonation Systems of English*****. New York: Cassell.**

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), ***Perspectives on Fluency*** **(pp. 102-127). Ann Abor: The Unversity of Michigan Press.**

Vita

Lin Alison McGregor was born on April 8, 1970, to Ralph and Ada McGregor in York, Pennsylvania. She attended Dover Area High School and received a Bachelor of Arts degree in Cultural Anthropology with a minor in German from Kansas State University. During her junior year, she was awarded a Kansas State University/Deutscher Akademischer Austausch Dienst scholarship for a one year study abroad program in Munich, Germany. In 1997, she obtained a Master of Arts in Liberal Studies, an interdisciplinary degree program combining the fields of English as a Second Language, Cultural Anthropology, and Communication at Wichita State University. During her graduate program, she worked with Harold T. Edwards, Ph.D., the speech pathologist who trained her in accent modification for non-native English speakers.

Alison has spent over 10 years working in the field of English as a Second Language (ESL) specializing in speaking and pronunciation skills for non-native English learners. Her teaching experiences have included teaching conversational English for a year in Taiwan, and three years in Japan teaching business English at Nippon Denso, Kawasaki Heavy Industries, and Mitsubishi Electric. Alison has also worked for over six years with non-native English speaking professionals residing in the U.S. on accent modification.

During Alison's five year doctoral program, she has had the opportunity to work for Graduate Studies, the Center for Teaching Effectiveness, the Educational Psychology Department, ESL Services, and Connection in Undergraduate Studies at the University of Texas at Austin.

Permanent address: 609 B. W. 35th Street, Austin, TX 78705

This dissertation was typed by Lin A. McGregor.