RATINGS OF L2 ORAL PERFORMANCE IN ENGLISH: RELATIVE IMPACT OF RATER CHARACTERISTICS AND ACOUSTIC MEASURES OF ACCENTEDNESS

by

OKIM KANG

(Under the Direction of Donald Rubin)

ABSTRACT

Judgments of nonnative speaker (NNS) oral language performance are notoriously susceptible to rater biases. While acoustically measurable elements of pronunciation are indeed relevant to comprehensibility, variance in oral proficiency scores due to rater background and linguistic stereotypes constitute trait-irrelevant error. This study tested the supposition that raters' background characteristics — including attitudes toward World Englishes — influence their rating of oral performances. In addition, a brief psychosocial intervention was undertaken in an attempt to mitigate the impact of potentially biasing rater characteristics. Seventy US undergraduate students rated the speaking and teaching proficiency of eleven international teaching assistants (ITAs). The ITA speech samples were acoustically analyzed on 12 measures of speech rate, pauses, stress, and intonation. Following the intial rating, one group of raters participated in a social-psychological intervention (training), in which undergraduates solved mystery puzzles with ITAs and interacted informally. Finally, 63 raters (including 29 trained) rated the same ITA speech samples a second time. All data collection from participant-raters was conducted online, including brief interviews as a qualitative supplement to the main results. Separate multiple regressions were employed to account for the variance in each dependent

variable (rater judgment scores), based on linear combinations of independent variables (rater and speaker characteristics). Rasch modeling also yielded measures of rater stringency. Analysis results revealed that about 20-30 % of variance in proficiency ratings were attributable to rater background characteristics. Rater native English speaker status, amount of contact with NNSs, prior teaching experience, and negative past experience in ITA courses affected student judgments of ITAs' accented speech. In contrast, 60-70% of the variance in ITAs' oral performance ratings was attributable collectively to objectively measured prosodic pronunciation factors, especially acoustic fluency. The intercultural sensitization intervention mitigated the impact of rater biases on ratings of ITA instructional competence. Among recommendations for screening and training raters warranted by this study is the notion that students who feel that their class grades have been harmed by NNS instructors should be disqualified as raters in high stakes speech assessment. Conversely, NNSs who wish to improve their perceived oral proficiency should work on avoiding filled and irregularly placed pauses.

INDEX WORDS: Rater bias, Oral Language Assessment, World Englishes, International Teaching Assistants, Prosody, Foreign Accent, Contact Hypothesis, Speech Perception

RATINGS OF L2 ORAL PERFORMANCE IN ENGLISH: RELATIVE IMPACT OF RATER CHARACTERISTICS AND ACOUSTIC MEASURES OF ACCENTEDNESS

by

OKIM KANG

B.A., Chungnam National University, South Korea, 1997

M.A., The University of Auckland, New Zealand, 2003

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOTOR OF PHILOSOPHY

ATHENS, GEORGIA

2008

RATINGS OF L2 ORAL PERFORMANCE IN ENGLISH: RELATIVE IMPACT OF RATER CHARACTERISTICS AND ACOUSTIC MEASURES OF ACCENTEDNESS

by

OKIM KANG

Major Professor: Donald Rubin

Committee: Linda Harklau
Seock-Ho Kim
Rebecca Callahan

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2008

DEDICATION

To the Memory of my Father and Mother

ACKNOWLEDGEMENTS

There are many people who should be thanked for their support during the preparation of this dissertation. First and foremost, I thank my supervisor and dissertation advisor, Dr. Donald Rubin, who has always supported and encouraged me during this dissertation journey. Don's role as a mentor has gone far beyond work on this dissertation. He has not only challenged, inspired, and guided me throughout my doctoral programs, but also helped me find my own professional identity in the field of language assessment and World Englishes. Without his support in my application for the Spaan Fellowship, for example, I would not have been able to undertake this study. Also, I really want to thank him for his tireless readings of my many drafts and helpful suggestions for revision. I know that Don will remain as a role model for the rest of my professional endeavours.

My special thanks go to my committee members: Dr. Linda Harklau, Dr. Seock-Ho Kim, and Dr. Rebecca Callahan because their suggestions and comments led me to refine and rethink my work. I sincerely thank Linda for providing me with valuable opportunities in mentoring ITAs, which is a grounded subject of this research. Her warm and kind concern for the timely completion of my dissertation has been invaluable. I would like to express my sincere appreciation to Dr. Kim for his crucial suggestions on statistical data analyses and for his thorough and detailed comments on this dissertation. I am also very grateful to Rebecca for her kind and insightful suggestions along the progress of this study.

My sincere thanks should be extended to Dr. Lucy Pickering, at Georgia State University, who provided me with practical knowledge about acoustic analysis methods. This study has greatly benefited from the collaboration of the ETS-funded project with her because the current study and the ETS-funded project were methodologically analogues, and Lucy's input in suprasegmental measures was essential. I would also like to express thanks to Dr. Ramon Littell, the University of Florida, for kindly offering me the consultations of the SAS mixed model analysis, to Dr. Tracey Derwing, the University of Alberta, for showing her interest in my research and giving me encouragement, and to Dr. Rachel Strom, the University of Texas, for sharing her curriculum for inter-cultural intervention meetings. Also, I wish to thank Dr. Louis McBee for her constant faith and support in my academic career. I am indebted to all of my good friends, Emily Gung, for helping me devise the online rating instrumentations, and Shelly Hovick for introducing this mixed random coefficient modelling for the alternative statistical analysis of my data. I sincerely appreciate the Spaan Fellowship Committee at the University of Michigan English Language Institute Testing and Certification Division for funding this study.

I cannot say enough to thank my husband, Jinhee Yi, for his unconditional and endless love. I know my entire doctoral program would not have been possible without his support, devotion, and trust. He always puts my work on his first priority, setting his duties aside, if necessary. I thank my father-like friend, Joe Webb, for his financial and emotional support. Finally, I hope my daughter, Dain Yi, can someday understand how thankful I am for her happy smiles and for her being with me always. Her happy face has always turned my tears to smiles. Last, but most gratefully, I wish to dedicate this dissertation to the memory of my late father and mother in heaven. Their fondest goal for my life was that I successfully complete my doctoral studies. I know their love continues with me as I work toward the further success.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER ONE: INTRODUCTION

Contemporary English language teaching and learning are being profoundly transformed by the recognition that English is increasingly used for international communication (McKay, 2002). So widespread is this use of English for international communication, that many such interactions—perhaps the majority of them—involve no L1 (native) speakers of English whatsoever. According to Kachru (1997), L1 speakers are actually a minority of English speakers, as there are far more non-native speakers of English in the world than there are native speakers. Communication in English between non-native speakers is far greater in frequency, amount, and aggregate significance than that between non-native speakers and native speakers (Yano, 2001). Therefore, it can be said that English is a fundamental tool for global communication in that people nowadays often share their everyday life experience in English. Being able to understand each other is the most important goal for these English speakers (McKay, 2002).

Despite the revolution in the users and functions of the English language, the norms for a particular minority of English language speakers continue to frame the dominant conventions for the assessment of English language oral performances (Kim, 2006). International tests of English language proficiency are unfair to speakers of non-standards forms of English, since these tests privilege standard forms (Hamp-Lyons & Davies, 2008). Because non-native speakers far outnumber native speakers (Crystal, 2003; Kachru, 1997), and because English is used for world-wide communication by native and non-native speakers alike, questions arise regarding whose standards should prevail in addressing English language proficiency. These questions include (1) what characteristics are required for raters to effectively evaluate non-native speakers' English

language oral proficiency and (2) to what extent are raters' impressions of "foreign" accentedness (and other oral language factors, for that matter) in L2 English performance assessment valid. Therefore, extensive research pertaining to the assessment of English language oral performances for non-native speakers is warranted. In this study, the issues of raters' perceptions in L2 oral performance assessments, and the relation of those perceptions to objective measures of accentedness, were closely examined.

Statement of the Problem

Oral proficiency tests in L2 learning contexts serve as "gatekeeping tests" as people move around the world in search of jobs and opportunities (Jenkins & Parra, 2003, p. 90). Since the role of speaking ability has become more central in language teaching with the advent of communicative language teaching (Nakamura, 1995), the assessment of non-native speakers' English language oral performance has become one of the most important issues in language assessment. However, as Canagarajah (2006) argues, our professional knowledge gets muddled by the new movements of globalization and World Englishes, which pose fresh questions that are yet to be addressed. A problem can arise particularly with regard to raters of oral assessment; i.e., it is not clear who is qualified to judge L2 oral performances in English. Ratings of speaking skills are indeed extremely susceptible to rater expectation and stereotype, because listeners can be so prone to rendering social inferences about speakers on the basis of just a few seconds of speech (Bradac, Cargile, & Hallet, 2001; Piché, Michlin, Rubin, & Sullivan, 1977). That is, ratings of speaker accent can be distorted by raters' social judgments and beliefs about speakers along dimensions that are quite extraneous to the assessment task and criteria. Consequently, this leads to rater bias related to different speech samples collected from specific purposes of oral

performances. For example, it is likely that raters who are scoring high-stakes English proficiency tests engage in different processes than those who are grading ordinary classroom performances of non-native speakers (NNS).

The goal in language assessment is to "reduce sources of variability that are external to the learner's language performance to the greatest possible degree in order to reflect the candidate's true ability" (Wigglesworth, 2001, p. 188). The sources of trait-irrelevant variability in the assessment of second language (L2) oral performance are diverse. Besides, since the judgment of all listeners is suspect, to determine "accuracy" of a rating requires some benchmarking against objectively measurable linguistic features. For example, suprasegmental elements of prosody, which are known to contribute significantly to comprehensibility and communication (Derwing & Munro, 2005), can be objectively detected via instrumentation (e.g., Levis & Pickering, 2004). Although subjective ratings of oral proficiency remain highly suspect, speech science has made progress toward identifying objectively measurable features of pronunciation such as intonation and speech rate that can serve as such benchmarks.

In general, test scores are influenced by task characteristics and raters (Bachman and Palmer, 1996). A great deal of research has been done on task effect on oral test performances (Chalhoub-Deville, 1995; Shohamy, 1983; Skehan & Foster 1999; Wigglesworth, 2001). Also, a considerable body of research on rater effects pertains to evaluations of written language (e.g., Erdosy, 2004; Shohamy, Gordon & Kraemer, 1992; Weigle, 1998). Very recently, Iwashita, Brown, McNamara, & O'hagan (2008) investigated relationships between spoken language features (grammar, vocabulary, pronunciation, and fluency) and scores awarded by raters. However, research in raters' effects in oral performances assessments has not been still thoroughly documented.

Some researchers have studied rater effects on oral proficiency test scores (e.g., Backman, Lynch, & Mason, 1995; Chalhoub-Deville, 1995; Lumley & McNamara, 1995; Lynch & McNamara 1998; Upshur & Tuner, 1999). One rater may be more or less lenient or severe in providing scores on oral performance assessment than another rater (e.g., Lumley & McNamara, 1995; Lynch & McNamara, 1988). There has been some evidence of rater leniency or bias in favor of certain groups of candidates (McNamara, 1996; see especially Chapters 5 and 8 on raters and ratings). Furthermore, the nature and amount of rater training can affect scoring, at least in L2 composition assessment (Cohen & Reed, 2001). Trained and untrained raters have been shown to disagree on scale points (Barnwell, 1989). Also, differences have been found in rater severity and criteria utilization between trained and non-trained raters in the assessment of learners' L2 oral ability in particular (Barnwell, 1989; Galloway, 1980; Hadden, 1991).

Nevertheless, there has been little study about a host of other issues related to raters in language assessment: (1) *Who can be a qualified rater to evaluate L2 oral performances*? For example, can non-native speakers discern L2 oral proficiency as well or better than native listeners? (2) *How do raters' characteristics impact on their assessment of L2 oral performances*? For example, are raters with a high degree of cross-cultural contact less biased than raters with limited cross-cultural contact? Therefore, research is needed to study the ratings of non-native speakers' English language oral performances provided by raters from different (cultural, linguistic, educational, and professional) backgrounds. This study additionally sought to determine the degree to which a particular type of rater training (based on intergroup contact) may reduce the impact of rater characteristics on speech ratings

Foreign-accented speech is defined as "non-pathological speech produced by second languge learners that differs in partially systematic ways from the speech characteristics of native

speakers of a given dialects" (Munro, 1998). Pronunciation aspects of L2 speech can be of various types, from difficulty in producing individual phonetic segments to prosodic differences in intonation, lexical stress or sentence forcus (Shah, 2002). Typically, accent comprehensibility is determined by listener dictation accuracy, and degree of accentedness is judged by panels of experts. It is now possible, however, for elements of accent to be detected by instrument and computer-assisted acoustical analysis (Pickering, 1999). This approach leads to a possible solution to the problem of subjectivity in rating any oral performances. Thus, certain acoustical features of non-native speaker (NNS) speech can now provide baseline parameters for "degree of accentedness." To the degree that conformity to native speaker comprehensibility needs constitutes a criterion for oral proficiency, these acoustical parameters measured via instrumentation can be considered legitimate proxies of true score components of speaking proficiency.

Few studies, however, have addressed the acoustic characteristics of non-native speaker's production (e.g., Flege & Port, 1981; Magen, 1998). Besides, most acoustic studies in the past focused in a univariate manner on vowel duration differences, changes in fundamental frequency as related to intonation, or a measure of rate of speech (Fledge 1991; Flege & Port, 1981; Schmidt & Fledge, 1996), or some pause phenomena in pausological research (for a review see Griffiths, 1992). Acoustic parameters combining rate, pausing, and intonation of of NNS' speech have not yet been well studied as a result. In addition, acoustic research has been usually interpreted in a linguistic sense of phonetics or phonology (e.g., Griffiths, 1992 for a study of speech rate), or in perceptual characteristics of NNSs (e.g., Dunkel, 1991; Voss, 1984 for pause structures that affect NNSs' comprehension, and Wennerstrom, 1998 for intonation aspects) , but has not been widely applied to the field of assessment of oral performances. Thus, the present

study investigated how selected acoustic, [1]suprasegmental (prosodic), speech characteristics of English-accented speakers influence native listeners' perception of accentedness. It attempted to look for relationships between those acoustic measures of accentedness — along with effects of rater background — and impressionistic assessments of raters in L2 oral performance.

This study was guided by the following research questions:

1. What is the relative impact of rater background characteristics on ratings of L2 oral performance?

    i. … on ratings of oral proficiency?

    ii… on ratings of instructional competence?

    iii… on ratings of comprehensibility?

    iv… on ratings of accent standardness?

    v… on ratings of speaker *superiority*?

    vi… on ratings of speaker *social attractiveness?*

    vii… on naïve raters' severity scores?

    a. What is the impact of rater native English language status?

    b. What is the impact of amount of rater formal training in language and linguistics?

    c. What is the impact of the amount of self-reported contact by raters with NNS friends and acquaintances?

[1] According to David Crystal's definition from *a Dictionary of Linguistics & Phonetics* (2003), suprasegmental is defined as a term used in phonetics and phonology to refer to "a vocal effect which extends over more than one sound segment in an utterance, such as a pitch, stress or juncture pattern" (p.446). Similarly, prosody is a term used in suprasegmental phonetics and phonology to "refer collectively to variation in pitch, loudness, tempo, and rhythm. Sometimes, it is used loosely as a synonym for 'suprasegmental', but in a narrower sense, it refers only to the above variables, the remaining suprasegmental features being labeled paralinguistic" (p.378). On the other hand, the term 'acoustic' analysis involves "the use of instrumental techniques of investigation" for the study of the physical properties of speech sound (p. 7). Therefore, in this study, even though the term, 'suprasegmental', was primarily used to refer to phonetic features selected for the instrumental analysis, the terms, 'prosodic' or 'acoustic', were interchangeably used as well.

d. What is the impact of the amount of rater's experience in teaching/tutoring English as a second language or foreign language?

e. What is the impact of raters' past negative experience in ITAs' classes ?

f. What is the impact of degree of rater 'reverse linguistic stereotyping'?

2. What is the impact of objectively measured suprasegmental characteristics of accented English on ratings of L2 oral performance?

i. … on ratings of oral proficiency?

ii… on ratings of instructional competence?

iii… on ratings of comprehensibility?

iv… on ratings of accent standardness?

v… on ratings of speaker *superiority*?

vi… on ratings of speaker *social attractiveness?*

vii… on speaker ability scores?

g. What is the impact of speaker's acoustical fluency factor?

h. What is the impact of speaker's irregular boundary markers?

i. What is the impact of speaker's hesitation markers?

3. To what extent does a course of training (a brief socio-psychological intervention function) affect ratings of L2 oral performance?

Potential Contributions of the Study

Rater effects in the scoring of oral proficiency examinations constitute measurement error, and yet speech assessment is consummately sensitive to listener expectations and social

stereotypes. Previous research has well documented rater differences in severity (e.g., McNamara, 1996) or salience of rating criteria (e.g., McNamara 1990, 1996). Statistical methods (based on Rasch models and G-theory; e.g., Lynch & McNamara, 1998) have been developed to statistically control for such rater deviations. Yet were it possible to ascertain individual attitudinal and experiential characteristics that predisposed raters toward greater or lesser accuracy in rating speech samples, corresponding methods of screening, selecting, and training raters could be devised.

The present study represents an innovative approach to assessing certain rater characteristics that are likely biasing factors in speech evaluations and comparing the impact of those "nuisance" rater effects with the impact of features of pronunciation which are legitimately components of "true score" variance. Consequently, the findings of this study can challenge—and improve--the validity of oral proficiency rating. In addition, the results of this study can have implications for the interpretations of the assessment of non-native speakers' English langue oral performances, for rater training, and for the evaluation of NNS' English language oral proficiency. In fact, collaborative projects among researchers in language assessment, World Englishes, and linguistic analysis are needed to better develop assessment criteria, and to implement assessment training. Thus, this study can provide helpful information to assist in that effort.

Moreover, the specific context from which speech samples are drawn is classroom communication between international teaching assistants (ITAs) and U.S. undergraduate students. Because of the significant role of ITAs in U.S. higher education, a substantial body of research and practical scholarship has been devoted to ITA proficiency in English (Smith, Strom, & Muthuswamy, 2005). This study also contributes to that body of research and practice

regarding relations between ITAs and their undergraduate students. The involvement of undergraduates and ITAs--including the inter-cultural intervention opportunities--in this study can contribute mutual benefits for both ITAs and undergraduates, as the two seek to gain a better understanding of the other's frame of reference. Consequently, the study can help (1) provide opportunities to improve undergraduates' comprehension of World Englishes, and (2) contribute to more focused training in English pronunciation and teaching strategies for international teaching assistants.

Organizational Structure

In the next chapter, the literature review includes material surveying overall trends in language assessment, oral language assessment and raters' background characteristics, listeners' attitudes and speech perception in general, undergraduates' attitudes toward ITAs in particular, and acoustic measures of accentedness and their relations with listeners' comprehensibility and intelligibility. Next, in Chapter Three, the methodology of the study is described. It includes hypotheses, instruments, speech analysis methods, and data collection procedures, and data analysis implemented to address this study's research questions. The results of the investigations are provided in Chapter Four, while Chapter Five summarizes the findings, integrates them with prior research, addresses implications and limitations of this study, and offers recommendations for future studies.

CHAPTER TWO: LITERATURE REVIEW

This chapter provides an overview of existing scholarship in the field of L2 oral language assessment. The first part of this section has discussed the historical trends in language assessment. The second part of this literature review has focused on raters in oral language assessments, looking at possible variability in the assessment process associated with different rater characteristics. The third part of this chapter has reviewed issues of rater attitudes and speech perception. The last part has reviewed existing literature on acoustic measures of accentedness in L2 English.

Overview of Language Assessment

In this section, the theoretical trends of language assessment are overviewed, followed by empirical account of language assessment research. According to Spolsky (1978), the history of language assessment theory can be divided into three chronological periods: *pre-scientific*, *psychometric-structuralist*, and *integrative-sociolinguistic*. The first "pre-scientific" period is prior to the early 1950's, in which language assessment itself was not a distinct discipline. Language testing followed whatever general principles of testing were available in the humanities or social sciences (Noor, 1995).

The second "psychometric-structuralist" period originated from the early 1950s' and lasted through the late 1960's. In this period, contrastive analysis became a thriving practice, as both structural linguistic and behavioral psychology combined to provide a scientific model for applied linguistics. As a result, language assessment focused on specific language elements such

as phonological, grammatical, and lexical contrasts between the target language (L1) and the second language (L2).

The third movement in language testing theory, the "integrative-sociolingistic" period, is dated from the late 1960's to the point at which Splosky (1978) was writing. The dissatisfaction with the structuralist and behaviorist approach to language teaching and assessment (e.g., no room for creativity) led to a wealth of linguistic research on communicative competence and on the contexts of language. Earlier in 1970s, the notion of communicative competence was expanded to include the importance of context beyond the sentences to appropriate language use, which includes both the discourse and sociolinguistic situation (Hymes, 1972). According to Bachman (1990), the single most important practical development in language testing was the realization that "a language testing score represents a complexity of multiple influence". Thus, Bachman (1990) insisted that a language test score cannot be interpreted simplistically as an indicator of the particular language ability we want to measure, in that a language test is also affected to some extent by a) the characteristics and content of the test tasks; b) the characteristics of the test taker, and c) the strategies the test taker employs in attempting to complete the test task. Therefore, since 1970s, there has been a need for language assessment and all language testing practitioners to reconsider the interpretation and uses made of language test scores (Lynch, 2001; Noor 1995).

Canale (1988) extended Spolsky's (1978) categorization of language testing periods by adding a fourth, called the *naturalistic-ethical tradition.* Canale claimed that the naturalistic-ethical trend reflects the social responsibility (ethical) aspect of testing along with concerns for naturalistic language use in tests. This trend insists that language tests and assessment measure students' competence by using naturalistic language, as well as assess competence by observing

students as they perform authentic language tasks. In addition, the word "ethical" reveals the responsibility of test users to ensure that language tests are "valuable experience and yield positive consequences for all involved" (Canale, 1988, p. 77). That is, an ethical approach to language testing makes clear the limitations of our tests to all stakeholders involved—not only test takers, but also their parents, their teachers, school administrations, and political decision makers (Hamp-Lyons, 2000). The naturalistic-ethical perspective focuses attention on several fundamental questions such as "what to test, how to test, and why to test." reflect the view that language assessment involves many complicated issues.

The ethical implication of assessment is being addressed by a growing number of scholars especially inside the professions of English as a foreign language (EFL) (Templer, 2004). Fulchers (1999) notes that the moral problems of the late 20th Century have finally caught up with applied linguistics and language testers. McNamara (2002) stresses critical analysis of industrialized language testing. Spolsky (1995), for example, raises probing questions about the institutional and policy origins of language testing, centering on the TOEFL. In the past half-decade, a more radical turn has emerged, as reflected in particular in the work of Elena Shohamy (2001) on 'critical language testing', which interrogates the very practice of language testing as a site of classification, social manipulation and control. The connections between social relationships, language, and power are central to critical pedagogies (Norton & Toohey, 2004). Given the importance of ethical implication in language testing, it has been argued that the concerns for ethical conduct must be grounded in valid test use (Bachman, 2000).

Currently, a new movement toward World Englishes in language testing seems to have grown up rapidly. In particular, the increasing use of international tests of English proficiency (e.g., TOEFL or TOEIC), has been condemned on the grounds that such tests are biased or unfair

to speakers of non-standards forms of English, because these tests privilege standard form (Hamp-Lyons & Davies, 2008). Jenkins (2006) argues that English now has a growing number of standard varieties, and therefore "there seems to be no good reason for speakers from the Outer or Expanding Circles to continue to defer to NSs of the Inner Circle" (p. 43). In addition, scholars (e.g., Taylor, 2006) make suggestions for tests of English as an International Language (rather than American and British Standard English), since this is becoming globally recognized and its descriptive codification is proceeding fast (Hamp-Lyons & Davies, 2008). As a result, the conventions for the assessment of English language will take different norms as English is a world language.

In parallel to those theory-based periods of language testing, from the empirical point of view, language testing and assessment have also witnessed the refinement of a rich variety of approaches and tools for research and development (Bachman, 2000). Furthermore, the concerns of language testers have evolved beyond traditional approaches. In recent years, testers have begun to focus on the uses, impact and consequences of tests and their role in educational, social, political and economic contexts (Shohamy, 2001). These trends seem to point to new areas of research (Bachman, 2000).

From the mid-1960s through the 1970s in particular, language assessment practice was informed by a notional view of language ability as consisting of four skills (listening, speaking, reading and writing) and discrete components (e.g., grammar, vocabulary, and pronunciation). Language assessment research was dominated by the hypothesis that language proficiency consisted of a single unitary trait which could be measured by a quantitative, statistical research methodology (Oller, 1976).

Most importantly, the influence of second language acquisition (SLA) research in the 1980s spurred language testers to investigate a wide variety of factors such as field independence/dependence (e.g., Chappell, 1988), academic discipline and background knowledge (Hale, 1988) and discourse domains, encouraged by 'communicative' approach, (Douglas and Selinker, 1985) on language assessment performance. In addition, language testers paid attention to examining the strategies involved in the process of test-taking itself (Cohen, 1987). Furthermore, language testers were challenged by Pienemann, Johnson, & Brindley's (1988) charge to explicitly take the language learner's developmental sequence into consideration in the design of language tests and in the interpretation of test scores. Therefore, by the end of the 1980s, language testing and assessment had merged into the mainstream of applied linguistics (Bachman, 2000).

In a similar manner, the period of the 1990s to present has seen a continuation of a trend that broadened the issues and concerns of language testing into the areas of applied linguistics (Bachman, 2000). However, in this phase the field has also witnessed expansions in research methodology. The methodological approaches employed in language testing research and practices have become increasingly diverse. Methods used in research on language proficiency testing now encompass *criterion-referenced measurement*, *Generalizability theory*, *item response theory,* and *structural equation modeling*. Additionally, *qualitative research approaches* are becoming increasingly common in language testing research, and are being used to investigate research questions such as the effects on test performance of test takers' characteristics and the processes and strategies they use to respond to assessment tasks (Banerjee & Luoma, 1997). See Lynch and Davidson (1997) more for an overview of criterion-referenced measurement; see Bachman (1997) for an overview of Generalizability Theory; see Kunnan

(1998) for an overview of structural equation modeling; and finally see Banerjee and Luoma (1997) for a review of qualitative approaches to traditional psychometric analysis.

However, due to the nature of this current study, the review gives brief account of the Rasch model in item response theory (IRT) in particular. IRT has become the dominant test development methodology for large-scale standardized language proficiency tests (Bachman, 2000). It is a measurement model that enables test developers to estimate the statistical properties of items separate from the abilities of test takers so that test items and test takers are not dependent upon a particular group of test takers or a particular form of a test. While several different IRT models are commonly used in educational measurement, the Rasch model , in its various forms, is the most widely used in language testing (e.g., McNamara, 1990). The Rasch multi-facet model has been applied to investigate the effects of multiple measurement facets, typically raters and tasks in language performance assessments (e.g., Eckes 2005; Engelhard, 2002, 2003; Kim 2006; Lynch & McNamara, 1998).

The Rasch model (or a multi-facet model) adjusts the raw ratings that candidates earn to account for variations in rater severity, skill difficulty, and difficulty of the practice area in which the ratings were earned. It is acknowledged as one of the most promising developments not only for investigating rater factors in performance-based language testing, but also for providing feedback to raters (or their supervisors) on their rating performance (Lumley & McNamara, 1995). Rasch measurement can be conducted using a computer program such as Linacre's (1996) FACETS. Using the FACETS method of analysis, scores are decomposed based on a number of facets in the performance setting. The facets normally include task difficulty, rater severity, item difficulty (Kondo-Brown, 2002). For example, bias analysis studies using a multifaceted Rasch model found extensive interaction effects that involved rater variation. That

is, rater's severity or leniency can be consistently biased toward specific task types (Wigglesworth, 1993), specific criteria (Wigglesworth, 1993), or a particular rating time (Lumbly & McNamara, 1995).

Overall, in the past several decades, language assessment research and practice have witnessed a rich variety of research approaches and tools. On the practical side, advances in the technology of test design and development, along with the availability and use of computer- and web-based test administration, scoring and analysis, have resulted in a greater range of assessment procedures than has ever before been available (Bachman 2000). At the same time, however, this progress has broadened the research questions that are to be investigated. Some of the critical research questions pertain to raters in L2 oral performance assessments. That is to say, who needs to agree with whom? Which raters agree with which other raters? Which raters agree with the test's rating criteria and the reasoning behind them? What is the contribution of rater characteristics to oral performance test scores? How objectively can rater judgment be justified? To what degree can the rater training neutralize the impact of biasing rater characteristics on oral assessment? Such issues related to rater proclivities merit the ongoing attention that they are now receiving not only in the field of language testing but also in the area of applied linguistics in general (Cohen & Reed, 2001).

Oral Language Assessments and Raters

Oral Language Assessments

Oral assessment in language learning has been the subject of intensive attention among second-language acquisition researchers (Iwashita, McNamara, & Elder, 2001). Oral proficiency tests in L2 learning contexts frequently serve as "gatekeeping tests" as people move around the

world in search of jobs and opportunities (Jenkins & Parra, 2003). L2 oral tests are increasingly prevalent, in assessments of all shapes and sizes, from classroom based assessment to standardized proficiency tests and everything in between (Bonk & Ockey, 2003). Such increasing interest in oral assessment is likely "a product of the increased interpretability of test scores, potential validity of the scores when linked to real world criteria, and positive washback effects of such assessment tools" (Bonk & Ockey, 2003, p. 90).

Accordingly, the assessment of spoken language has evolved from tests of oral grammar and pronunciation to interview and, more recently, semi-interview and multiple tasks (Cohen, 1994). Some well-known oral proficiency tests include American Council on the Teaching of Foreign Language's Oral Proficiency Interview (ACTFL OPI), the Language Assessment Scales-Oral (LAS-O), the Woodcock-Munoz Language Survey, the IDEA Proficiency Test, the Simulated Oral Proficiency Interview (SOPI), the Speaking Proficiency English Assessment Kit (SPEAK), the Test of Spoken English (TSE), and very lately the speaking test in iBT TOEFL. In recent years, more integrative assessment approaches have been developed such as multiple measures of speaking ability. These integrative oral assessment tasks include a verbal essay, giving an oral presentation, and reporting the contents of an article, making oral portfolios (Brown & Yule, 1983), and taking part in role play; see more examples in Cohen (1994).

Oral proficiency tests have been continuously challenged regarding their authenticity, validity, and reliability. For example, the OPI has been used by the United States government since World War II to assess the language skills of American personnel working abroad. Then, it entered into the academic world in 1982, when the ACTFL first published its guidelines. Yet, researchers have raised questions with regard to the definition of the different proficiency levels in the guidelines (Halleck, 1995) and illustrated problems of the proficiency interview (see more

in Cohen, 1994): (1) assessing achievement more than general English proficiency; (2) not being culturally sensitive; (3) interviewers' mostly controlling the topics; (4) not being rater friendly at rating processes, etc. The validity of tests has been also questioned through the International English Language Testing System's (IELTS) oral interview. Brown & Kathryn (2007) recently reported that differential behaviors by IELTS interviewers could affect the scores awarded to candidates; i.e., some interviewers tended to consistently present a difficult or easy challenge to candidates. In particular, the easier interviewers often shifted topic more frequently, with fewer turns per topic; they also asked more questions of a simple nature.

As seen in criticisms of the OPI and the IELTS, for example, oral language assessment faces certain drawbacks or weakness in the test itself. That is, assessing oral performances is more challenging than assessing other skills because of its highly subjective nature (emotion and identity expressed paralinguistically as well as linguistically), the problematic mechanism of rater reliability, validity of performance itself (McNamara, 1997), unclear rating scales, and ill-defined components of oral ability itself (Bozatli, 2006). Logistical problems such as time, finance, scoring difficulty, and administration are especially troublesome for oral assessment (Underhill, 1987). When it comes to the content of oral tests, Lantolf and Frawley (1985) argue that oral proficiency assessment typically does not give sufficient consideration to research in communicative competence. Of further concern, oral performance brings potential variability in tasks and rater judgment, as sources of measurement error (Bachman, Lynch, & Mason, 1995). The potential variability in rater judgments has been of particular concern for language assessment (e.g., Bachman et al. 1995; Barnwell, 1989; Cumming, 1990 for oral assessment; Pollitt and Hutchinson, 1987 for writing assessment).

It is useful to consider 'rating' as a factor that affects test scores in language testing in performance assessment. McNamara (1996) argues that "rating is a result of a host of factors interacting with each other" (p. 453). He interprets the rating as an end-product of an interaction among task, test-taker, testing performance, rating criteria, rater, and interlocutor. He presents this interaction as shown in Figure 2.1.

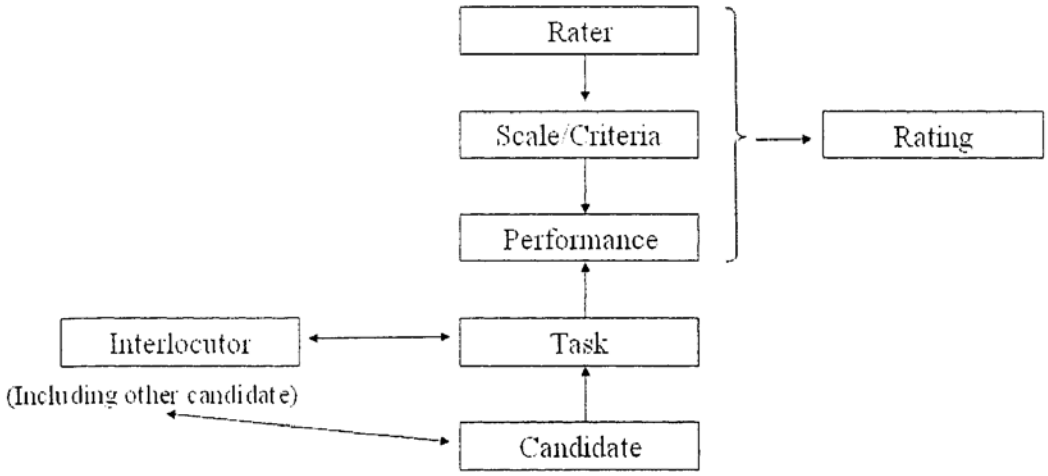


Figure 2.1

Proficiency and Its Relation to Performance (McNamara, 1996, p. 86)

In Figure 2.1, rater and rating scale/criteria are considered important variables influencing oral performances. In fact, this model depicts individuals in interaction as variables in the assessment process. However, this model does not detail in which way and to what extent rater characteristics affect the rating of L2 English oral performances. In fact, the variables such as raters' backgrounds and their attitudes are issues that are rather complex.

Another well known model of testing speaking ability was developed by Skehan's (1998) model which appears in Figure 2.2.

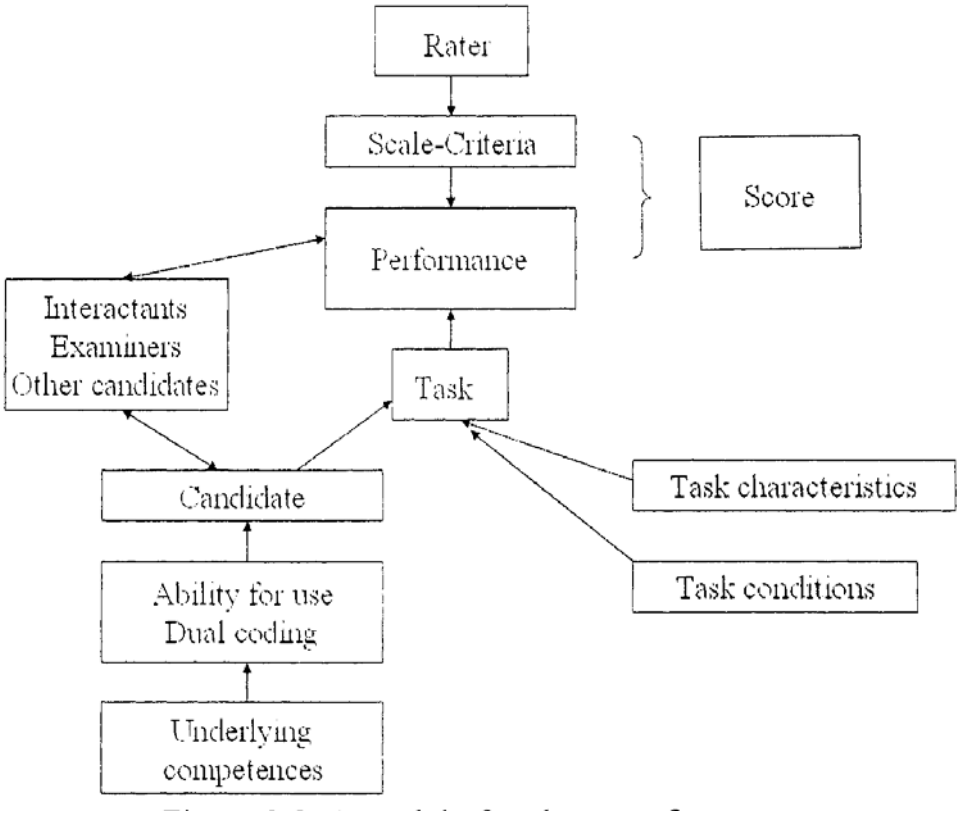


Figure 2.2

A Model of Oral Test Performance (from Skehan, 1998, p. 172)

Figure 2.2 elaborates on the roles of task and of candidate's competence, but it too fails to specify rater traits that can affect oral performance scores. This gap in influential models of oral performance assessment is striking, given the acknowledged impact of rater training and rater characteristics (Fulcher, 2003). In order to augment currently available models, the present study will investigate the effects of rater background characteristics on L2 oral performance assessment.

Raters

Performance test scores are closely linked to raters (McNamara, 1996). The rater is an additional source of measurement error. Nevertheless, research specifically pertaining to rater

effects (e.g., Bachman et al., 1995; Brown, 1995; Brown, Iwashita, & McNamara, 2005) and to the evaluation of spoken English proficiency (Shohamy, 1993) is still in some respects in exploratory stages (Boulet, van Zanten, Mckinley & Gary, 2001).

Previous research on rater effects on speech assessments has centered on two issues (Myford & Wolfe, 2000): (a) variability in rater rigor or severity and (b) differential salience and application of rating criteria among raters. Linacre (1989/1993) used the term 'severity' to refer both to the overall severity of the rater and to differences between raters in the way they interpret rating scales or criteria. Virtually all of the small handful of studies that have investigated rater severity on L2 oral performance assessments have found significant, meaningful differences among judges (Lumley & McNamara, 1995; Lynch & McNamara, 1988). McNamara and Adams (1991/1994), use of the term '*rater characteristics'* to cover both overall severity and more specific effects such as rater bias.

Conventional psychometric theory conceptualized rater characteristics in terms of the difference between the 'perfect' examiner and 'ordinary' examiners (Lumley & McNamara, 1995). Given the severe shortcomings of ordinary examiners, however, it seems rather obscure to define the concepts of 'perfectness' and 'ordinariness' themselves. Diverse rater groups, for example, native speaker raters as compared with non-native speaker raters, may differ in judging learners' L2 ability as a result of their backgrounds (Engber, 1987) and the set of criteria with which they operate in writing tests (Shohamy, Gordon, & Kraemer, 1992). In a fair test, to the contrary, raters must understand and apply the evaluation criteria in similar ways (Bejar, 1985; Brown et al., 2005; Myford & Wolfe, 2000; Pollit & Murray, 1995). If some raters give greater weight to, say, pronunciation and others focus on grammar, then it cannot be said that these raters are embodying the same test (Brindley, 1991). Therefore, it is important to know how

differently or systematically raters rate tests in a practical sense. In the following sections, the discussion will include possible variances due to different raters' characteristics including their training experiences.

Rater Characteristics

Effects of Rater Educational and Professional Experience

Expert raters, as compared with novice raters, may be less influenced by surface language features and more capable of examining language use, content, and rhetorical organization concurrently (Cumming, 1990). Status as a language teacher may confer expertise on a rater. Studies by Galloway (1980), Barnwell (1989), and Hadden (1991) all found that classroom teachers and non-teaching native speakers differ in their assessments of learners' L2 oral ability. Whereas Barnwell's (1989) results found that the non-teaching raters were relatively harsher than the teaching rater group, Galloway's (1980) and Hadden's (1991) findings indicated that teachers were more critical of students' grammatical abilities than were lay persons. On the other hand, Brown's (1995) showed that there was little evidence that raters with a teaching background would be more suitable than those with an industry background (or vice versa).

Galloway (1980) has also documented a difference between non-teaching native speakers residing in the learners' community and those living in the target language community. She observed that the group of non-teaching native speakers living in the USA, rather than in the learners' community, appeared to be more tolerant of all facets of the students' communicative competence. Similarly, Chalhoub-Deville's (1995) investigation of three rater groups (teachers of Arabic as a foreign language in the USA, nonteaching Arabic speakers residing in the USA for at least one year, and non-teaching Arabs living in Lebanon) indicated that teachers tended to

emphasize grammar in their assessment of students' proficiency, and non-teachers tended to be concerned with the more communicative aspects of the language.

In contrast, another set of studies found that linguistically naïve raters who had no professional expertise were quite sensitive to cross-dialectal and cross-linguistic differences in prosodic patterns, and that such listeners performed reliably, when judging foreign accents (Anderson-Hsieh & Koehler, 1988; Brennan & Brennan, 1981; Munro 1995). On that basis, Piske and colleagues (2001) suggested that a broad and diverse sample of raters should be recruited and not only one particular type of rater employed.

Thus, the research literature does *not* offer any consensus regarding the impact of rater professional expertise on the capacity to render accurate evaluations of speaking proficiency.

Effects of Rater Nationality and Native Language

Research is inconclusive regarding how raters' nationality and native language affect their ratings of examinee oral proficiency. Brown's (1995) results pertaining to the Japanese Test for Tour Guides showed that there is little evidence that native speakers are more suitable than non-native speakers.. However, the influence of raters' native speaker status did emerge in other studies, especially writing evaluation (Erdosy, 2004). For example, some studies (e.g., Fayer & Krasinski, 1987; Santos, 1988) have found NNS raters to be more severe than NSs, and explained their findings by referring to the considerable time and energy NNS assessors had invested in learning the target language themselves, which led them to attribute errors to a lack of commitment on the learners' part (Santos, 1988, p. 85).

Chalhoub-Deville and Wigglesworth (2005) inquired whether there was a shared perception of speaking proficiency among raters from different English speaking countries (Australia, Canada, the UK and the USA), when rating speech samples of international English

language students, and differences in stringency across rater nationality were indeed found. The UK raters were the harshest and the US raters were the most lenient. In a similar investigation of holistic ratings, Chalhoub-Deville (1995) argued that some rater groups (teachers of Arabic as a foreign language in the USA and non-teaching Arabs residing in the USA) are more or less lenient, compared with Arabs living in their home country (Lebanon), in judging learners' performance.

One factor that may contribute to greater tolerance of raters from particular language backgrounds for certain NNS accents is "the interlanguage speech intelligibility benefit" (Bent & Bradlow, 2003). A nonnative listener may be especially well equipped to interpret specific acoustic-phonetic features of a L2 that are matched with her or his own L1 (see Pickering, 2006). Major, Fitzmaurice, Bunta, & Balasubramanian's (2002) studies, for example, indicated that due to prosodic similarity in rhythms among Chinese, Japanese, and Spanish, (i.e, the lack of vowel reduction), Chinese and Japanese listeners understood Spanish accented English rather well. It is not clear, however, how this matched interlanguage benefit works in conjunction with other intelligibility factors (Pickering, 2006), nor how it affects the proficiency rating process.

Effects of Rater's Intercultural Contact and Exposure to NNS Varieties

Extensive experience in evaluating the speech of nonnative speakers may not necessarily be a prerequisite for rendering accurate judgments of nonnative speakers (Mattran, 1977). On the other hand, according to Thompson (1991), individuals unfamiliar with a particular World English variety generally perceive a higher degree of L2 foreign accent than do those who are familiar with that particular variety. Building on the research of Rubin (1992) and others, Powers, Schedl, Wilson-Leung, and Butler (1999) had all listeners complete a language background questionnaire at the outset of the experiment in which they indicated their degree of

familiarity with languages other than English. The scales pertained to participants' foreign-language study and travel, the nature and frequency of their contact with non-native speakers of English, and interactions with non-native speakers. Their findings showed that no variables were consistently related to judges' performance. Similarly, Derwing and Munro (1997) asked listeners to indicate on a scale of 1–5 the amount of contact they had had with people who speak with any foreign accent. Their results showed that listeners' self-reported exposure to various accents predicted their success at language identification, and also there was a correlation between exposure and intelligibility scores.

Indeed, these findings are generally consistent with evidence that amount of interaction with speakers of specific languages or World Englishes facilitates listening comprehension of those English varieties (Field, 2003; Gass & Varonis, 1984; Polio & Gass, 1998).

Effects of Rater Training

Rater training is presumably the solution to reduce variation due to rater effects. Through exposure to anchor point examples, through review of rating criteria, and by means of comparison with other raters, conventional rater training should be able to "calibrate" novice raters to a consistent standard and to recalibrate more experienced raters who may have drifted from that standard. Yet a body of research indicates that conventional training does not always reduce inter-rater differences (Brown, 1995; Myford & Wolfe, 2000). The benefits of training may be short term at best (Lumley & McNamara, 1995). Even with training, then, raters are not interchangeable. Reviewing the implication for rater training, McNamara (1996) recommended

> To accept that the most appropriate aim of rater training is to make raters internally consistent so as to make statistical modeling of their characteristics possible, but beyond

this to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some ways (p. 127).

Compensation for the nonconformity of individual raters can be achieved through statistical approaches such as Generalizability theory (e.g., Stansfield & Kenyon, 1991) and many-facet Rasch modeling (e.g., Linacre, 1993; McNamara & Adams, 1991; Lumley & McNamara, 1995). Generalizability theory provides a methodological approach to estimating the relative effects of variation in test tasks and rater judgments on test scores (Crocker & Algina, 1986; Shavelson & Webb, 1991). Many-facet Rasch measurement investigates rater fit and adjusts for rater severity (Linacre, 1993). Both G-theory and Rasch measurement manage the variability inherent when raters rate examinees on test items. However, according to Linacre (1993), Rasch test reliability is higher than G-theory reliability because Rasch error variance (by using adjusted scores) excludes item and judge variance whereas test reliability is estimated in a raw score metric in G-theory.

As useful as is this information regarding individual differences in rater severity and criteria utilization, and as valuable as are some theory or model-based solutions for statistically mitigating the effects of that variability, the psychometric foundation of oral proficiency tests would be strengthened were it possible to screen and select for raters who share characteristics that facilitated agreement both among themselves and also conformity to the intended evaluation criteria and standards. Thus, an examination of research on rater characteristics is warranted for these very pragmatic reasons, as well as for the purpose of advancing our conceptualization of the linkages between speech judgments and social perception.

Attitudes and Speech Perception

Attitudes and Perception

Another important factor that influences the rating process can be found in raters' attitudes toward non-native speakers' Englishes. There has been little research regarding raters' potential attitudes toward accented-English in language assessment. The term ' language attitude' is usually used to refer to beliefs about a specific language or to an orientation (positive or negative) towards a specific language that influences the individual's evaluation of that language and its speaker (Cluver, 2000). Gardener and Lambert (1972) early emphasized that the concepts of attitudes affect our judgments and perceptions of others; influence our speed and efficiency of learning; and help determine the groups we associate with.

Beginning with the ground-breaking work of Lambert and his colleagues (Lambert, Hodgson, Gardner, & Fillenbaum 1960), a number of matched-guise studies on language attitude have clearly demonstrated that people make moral, intellectual, and aesthetic judgments of others based on their language choice and accent alone. In matched-guise studies, listeners are asked to rate recorded speakers on a number of qualities, which may be divided into status-related qualities such as intelligence and ambition, and solidarity-related qualities such as friendliness and likeability. The speakers read a standard text; in the language-choice studies, the same bilingual speakers actually record both language versions of the text. The Linguistic Stereotype hypothesis (Lambert et al., 1960) holds that even short samples of nonprestige varieties of speech are sufficient to trigger among listeners a cascade of negative evaluations of speakers. Many of those evaluations are quite extraneous to language behaviors and touch upon physical characteristics like height and attractiveness, general intelligence, and civility.

These sorts of judgments may result in language-based discrimination. Students with "poor voices" are judged by teachers to be less intelligent than those with "good voices" (Seligman, Tucker, & Lambert, 1972). Australians with "broad" accents are rated by potential employers as unsuitable for high-status jobs (Seggie, Smith, & Hodgins, 1986). Speakers of stigmatized varieties of English are misinformed by landlords that there are no available apartments (Purnell, Idsardi, & Baugh, 1999). In this sense, non-native speakers may be denied raises or even fired by employers who claim they have poor language proficiency (Lippi-Green, 1997).

In the case of non-native speakers, issues of prestige and group membership are compounded if the speakers are members of stigmatized groups and have stigmatized accents which index them as such. Ryan and colleagues (Ryan, Carranza, & Moffie, 1977; Ryan & Sebastian, 1980) have looked at the case of US native English speakers' perception of Spanish-accented English, where non-native speakers are both outgroup and low prestige. Unsurprisingly, these non-native speakers were rated lower on measures of both status and solidarity.

Non-native speakers are very likely subject to linguistic stereotyping when their L2 language proficiency is being evaluated. Rating a speaker's language proficiency based on a language sample certainly seems much more reasonable than rating such traits as intelligence, competence, or kindness based on a language sample. However, even proficiency judgments may have more to do with listeners' attitudes about the speakers' ethnicity than with the speakers' actual intelligibility (Lippi-Green, 1997). Therefore, it is first helpful to take a closer look at the characteristics of listener's perception.

The term "perception" is often used in the speech perception literature to refer only to that portion of comprehension that listeners achieve through extraction of information from the

acoustic signal and perhaps visual articulatory cues (Lindemann, 2000). Large inter-listener differences are consistently found in the cross-language speech perception literature (e.g. Flege, Munro & Fox 1994; Jenkins, Strange, & Polka, 1995; Yamada, 1995). Jenkins et al. (1995) point out that such variability may be explained by differences in listeners' language experience. Yamada (1995) lists student motivation to learn a language as one of several possible sources of variability in their speech perceptions. Although these inter-listener differences are reported in studies of listeners' perceptions of a non-native language, we might also expect them to be applicable likewise in cases where a listener hears her native language, but is perceiving something akin to a non-native phonology used by a non-native speaker of that language.

There are numerous phenomena showing the susceptibility of speech perception to influence by information from non-auditory channels. Perhaps, the most well-known is the McGurk Effect (McGurk & MacDonald, 1976) which demonstrates an interaction between hearing and vision in speech perception. For example, in the McGurk effect, listeners who are presented with the auditory stimulus "ga-ga" while watching a (silent) video of a speaker saying "ba-ba" report hearing "da-da", thus integrating velar sound information from the auditory channel with bilabial information from the visual channel. This effect suggests that knowledge about a phoneme seems to have little effect on one's perception of it, and visual and auditory phonetic cues are perceptually integrated.

Other studies show an apparent influence of phonemic restoration on perception. Phonemic restoration is an auditory illusion in which listeners "hear" parts of words that are not really there. In earlier studies of the illusion (see review in Samuel, 1981), segments of words (phonemes) were replaced by an extraneous sound; listeners were asked whether any thing was missing and where the extraneous noise had occurred. Most listeners reported that the utterance

was intact and mislocalized the noise, suggesting that they had restored the missing phoneme. Similarly, Samuel (1981) found that listeners presented with a word with one phoneme replaced by white noise tended to hear the missing phoneme along with the extraneous (white) noise. That is, the listeners' identification of the word being presented results in their hearing the whole word, even though one phoneme is actually missing.

Research on the sociology of language has also demonstrated that perceived language variation triggers evaluative judgments about the speaker. For example, Strand's (1999) study and Strand and Johnson (1996) showed a relationship between social factors and listener's perception. They argue that expectations of gender proto-typicality and stereotypes can change listeners' perception of the physical signal. Strand (1999) found that listeners' stereotypes about gender, as activated by the faces and voices of speakers, altered the listeners' perception of the fricatives /s/ and /ʃ/. In general, the /s/ spectrum is concentrated at a higher frequency than the /ʃ/ spectrum, and men generally have lower frequency turbulence than women. Because male speakers typically use a less fronted, more grooved variant of /s/ (Naslund, 1993; cited in Strand, 1999), which results in a lower-frequency (and therefore more /ʃ/-like) variant, it is expected that listeners to accept more tokens of an /s/-/ʃ/ continuum as /s/ if they believe the voice to be that of a male speaker. What is especially interesting about Strand's findings is that the perceptual boundary between /s/ and /ʃ/ shifted depending on whether it was accompanied by a male or a female face producing the token on video — who might or might not be the same sex as the voice they were hearing. Strand argued that listener expectations (stereotypes) about how the speaker *should* sound, based on how they look, were affecting how they actually processed the speech.

In the review of the literature on the relation between stereotypes and the perception of language itself, von Hippel, Sekaquaptewa, and Vargas (1995) point out that expectations and stereotypes guide our understanding of the world. They argue:

> Any evidence that perceptual processes influence and are influenced by stereotypes and prejudice would have profound implications. People view their senses as documentary devices that faithfully translate the environment into understandable and manageable units . . . they *accept* what they see and hear. (p. 181)

Von Hippel et al. (1995) claim that those stereotypes can play a role at the most basic level of perceptual encoding of information. As a result, for instance, they declare that those who are prejudiced against groups of non-native speakers may *hear* a non-native speaker who is competent in English as unintelligible. Native speakers' lack of exposure to non-native speakers' English and lack of a shared common core of L1 phonological features with English might accentuate negative attitudes in native speakers (Jenkins, 2000).

When it comes to the relationship between perceived accent and intelligibility, Schmid and Yeni-Komshian's (1999) study provided interesting findings. Schmid and Yeni-Komshian examined how native English listeners perceive sentences produced by non-native speakers and reported that a strong foreign accent did not necessarily reduce the comprehension of speech produced by non-native speakers. Their findings showed that (a) listeners were more accurate and faster in detecting mispronunciations produced by native speakers than non-native speakers, and (b) were more accurate in detecting mispronunciations produced by non-native speakers with milder accents, as compared to heavier accents.

A highly influential study on the effect of attitude on perception was done by Munro and Derwing (1995). Munro and Derwing examined the interrelationships among accentedness,

perceived comprehensibility, and intelligibility in the speech of L2 learners. They asked NSs of English to listen to English speech produced by Mandarin NSs, to transcribe the utterances in standard orthography, and to rate them for degree of foreign-accentedness and comprehensibility on 9-point scale. They found that most listeners showed significant correlations between accentedness and the accuracy of their transcription, fewer listeners showed correlations between accentedness and perceived comprehensibility, and fewer listeners still showed a relationship between accentedness and intelligibility. Their findings suggest that even though the degree of foreign accent is correlated with perceived comprehensibility and intelligibility, a strong foreign accent may not necessarily reduce the comprehensibility or intelligibility of L2 speech. In other words, listeners' perception of NNSs' speech can be separate from their evaluations of the speaker's accent. It is possible, then, for either (or even both) of the two factors (comprehensibility and accuracy) to correlate with listener's attitude toward non-native speech. Therefore, a possible interpretation to Munro and Derwing's discussion would be that those who are prejudiced against groups of non-native speakers may *hear* a non-native speaker with accurate English pronunciation as having a heavy accent.

Rubin's (1992) study has shown listener expectations based on speaker ethnicity to have an effect on those comprehensibility and accuracy factors to some degree. In his study, participants listened to four minutes of a tape-recorded lecture produced by a native speaker of Standard American English. Instructor ethnicity was operationalized b projecting a photograph of either Caucasian woman or an Asian woman. Then, the results showed dramatic evidence that listeners reacted to factors extraneous to just language proficiency when judging NNSs' speech. Listeners who were shown a fabricated picture of an Asian delivering the lecture perceived more of a foreign accent and scored lower on a recall test than those who were shown a photo of a

Caucasian, even though what they heard was exactly identical. In other words, listening comprehension[2] appeared to be undermined simply by identifying (visually) the speaker as Asian. The potency of non-language factors in affecting listeners' reactions to NNSs is quite surprising even though the finding of this study may not generalize to more Western NNSs instructors such as instructors from France or Germany.

The sort of "reverse linguistic stereotype" effect to which Rubin's work points, that is, washback of general judgments about social groups to specific evaluations of individual speakers' language proficiency, is substantiated by other research. Nguyen (1993) has claimed that inherent rater biases against certain nationalities renders valid standardized testing of oral proficiency unattainable for speakers from those countries. More recently, Lindemann (2002, 2003) confirmed that generalized stereotypes affect perceived accent and perceived language proficiency, and in fact deter U.S. undergraduate students from effectively interacting with instructors whom they believe to be of particular (negatively stereotyped) NNS backgrounds.

Furthermore, Lindemann (2005) examined how native US English speakers construct social categories for people outside the US. Her close look at that one group's belief system provides insights that can be used in addressing linguistic discrimination, with information on how varieties and features of varieties are perceived. Two hundred eight US undergraduate students rated the English of students from 58 countries. Familiarity and socio-political beliefs about countries of speaker origin appeared to play a role in responses. Evaluation was often central to description; a category of stigmatized, often "broken", English was used to describe all

[2] Please note that in Rubin's (1992) study, a cloze test was used to measure listeners' comprehension ability. In a cloze test, listeners were given with the text of the lecture with some words (every-7th- word) missing after listeners had heard the lecture, and asked to fill in the missing words. In fact, a cloze test is more properly regarded as a measure of recall, not comprehension (Rubin, 1992). Therefore, the processes involved may be somewhat different from those involved in on-line perception, even though the recalling technique can be an important outcome of listening comprehension.

non-native speakers except Western Europeans. For example, Asian English was negatively evaluated overall, yet Korean and Japanese English tended to be rated even more negatively than Chinese English. While Mexican English was somewhat negatively evaluated, Russian English was considered as harsh and guttural.

It is possible that when listeners harbor certain stereotypes about speaker identity, they are rendered incapable of objectively assessing speaker pronunciation. Little is known about what individual differences predispose some listeners to be more prone to linguistic stereotyping than others, nor is there any standard attitude scale for measuring such stereotypes. While some scales are available for measuring generalized attitudes toward international teaching assistants (e.g., Bresnahan & Kim, 1993; Fox & Gay, 1994), these presume recent contact with such college instructors.

Attitudes toward International Teaching Assistant (ITA)

The social/linguistic stereotyping effect can be pernicious in that ratings of speaker accent can be distorted by perceptions (Nisbett & Wilson, 1977). This distortion can be particularly potent within a relation between undergraduate students and ITAs. It is often believed that the ITAs' lack of English proficiency hinders the undergraduates' ability to comprehend subject material (Smith, Strom, & Muthuswamy, 2005). In addition, undergraduates' ratings of ITAs' oral English proficiency correlates positively to their ratings of teaching proficiency (Davis, 1991; Inglis, 1993). If a student's discourse is difficult to decipher (e.g., instructor's pronunciation of words is not clear) students may perform less well in the classroom. Even though research shows that most non-native speakers possess sufficient language proficiency to accomplish their instructional goals, some people may still question the language skills of the non-native teacher (Llurda, 2004). In fact, NNS instructors are almost

universally regarded as being less competent teachers than their native-English speaking peers (Rubin & Smith, 1990).

Previous research (see summary in Rubin, 2002; see also Lindemann, 2002), documents that these student complaints are frequently more a function of students' stereotyped expectations than of instructors' objective language performance. Nonetheless, the negative attitudes and expectations held by many U.S. undergraduates can interfere materially with information uptake and exert a deleterious effect on learning. According to one survey, 40% of undergraduates at some point in their educations dropped or switched classes because the instructor was a NNS (Rubin & Smith, 1990). Lindemann (2003) demonstrated that U.S. students who have the most prejudiced expectations of internationals' English proficiency are least likely to engage vigorous questioning of those internationals. Reluctance to interact with one's instructor, no doubt does result in lowered learning outcomes. Students who harbor negative stereotypes of ITAs may "hear" interference where there is none.

The lack of intelligibility of ITAs accounted for 80 percent of the communication breakdown in Gallego's (1990) study of native speaker undergraduates' reactions to ITAs. Gallego found that undergraduates who listened to audiotapes of ITAs lecturing stopped the tapes and replayed them most often because of pronunciation problems, not vocabulary or grammar. When American undergraduates report problems with understanding their ITA, they usually cite pronunciation as the biggest difficulty (Schneider & Stevens, 1991). In another study, undergraduates' comprehension of audio-taped lectures did not differ if ITAs spoke with a comprehensible, foreign accent; their comprehension only lessened with an incomprehensible, foreign accent (Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002). When ITAs possess high

intelligibility, they are viewed more positively by their undergraduate students (Bresnahan et al., 2002).

However, changing a speech pattern or accent can be difficult and sometimes impossible (Derwing & Rossiter, 2003; Clarke & Garrett, 2004). Certainly, pronunciation remains an important part of ITA training as many international students who learn English as a Foreign Language have learned a decontextualized non-communicative version of English that is not readily understood by native-speaking undergraduates. Therefore, Bauer (1996) states that increasing the English language proficiency for teaching is probably a major focus of most ITA curricula.

In fact, ITA programs often target specific problem areas in pronunciation, such as the */l/* or */r/* sound, since some ITAs' native languages do not possess all the sounds that are required in American English pronunciation. Aside from specific phonological segmentals, differences in word stress can make speech difficult for native speakers to understand (Pickering & Wiltshire, 2000). Also, intonation is an important factor in ITA communication (Pickering, 2001, 2004). At the same time, fluency presents another challenge for the international teaching assistant since some ITAs speak too rapidly while others pause too frequently. Both extremes in speech rate interfere with the intelligibility of ITA speech. Research has addressed topics such as the acceptable speed for communicating within a university classroom (Pica, Barnes, & Finger, 1990), the interaction of speed and intelligibility in accented speech (Anderson-Hseih & Kohler, 1988), the lack of ITA fluency caused by pauses that fall in places other than between clauses or sentences (Wood, 2001), and excessive pausing (Grant, 2001) or irregular empty pauses (Round, 1987).

Therefore, based on previous studies assessing the relationship among students perceptions of English proficiency, ratings of teaching skills, and ITAs' actual pronunciation and accent, this study examines the contribution of suprasegmental characteristics in ITAs' speech to ratings of language proficiency and instructional competence by using ITAs' performances as speech samples and undergraduate students as raters.

Inter-group Contact as a Tool for Reducing Prejudice

This section reviews the effects of interactional inter-group contact on attitudes toward diverse and group relations, e.g., the "contact hypothesis" for reducing prejudice. It provides the support for the training method used with raters. Interactional contact between two groups has positive effects on group attitude, and improves group relations, and can reduce prejudice under certain conditions (Voci, 2003). According to the contact hypothesis (Allport, 1954), contact under these conditions: 1) equal status between the groups in the situation; 2) common goals; 3) no competition between the groups; and 4) authority sanction for the contact − will create a positive intergroup encounter, which will bring about an improvement in intergroup relations (Amichai-Hamburger & McKenna, 2006). In other words, more contact between individuals belonging to antagonistic social groups (defined by culture, language, skin color, nationality, etc.) tends to undermine the negative stereotypes they have of each other and to reduce their mutual antipathies, thus improving intergroup relations by making people more willing to deal with each other as equals (Forbes, 2004).

An early study surveyed social contacts and ethnic attitudes in four cities in different regions (Elmira in NY, Bakersfield in CA, Steubenville in OH, and Savannah in CA) of the United States (Williams, 1964). The study illustrated the kinds of statistical relations routinely observed when personal contact and prejudiced attitudes were correlated. Then, it was concluded

that in all four cities the individuals who reported having personal relations with one or more members of one or more racial or ethnic minorities were less likely to express prejudiced opinions than were those who had no such contacts. These results, though just correlational, suggest that prejudice can be reduced and inter-group relations improved by encouraging more contact across group boundaries (Miller, 2002; Pettigrew, 1998; Pettigrew & Tropp, 2000). That is, the more groups become interdependent in ways that require or encourage frequent communication across the linguistic boundaries, the less the prejudice between them (Forbes, 2004).

In terms of inter-group contact between ITAs and undergraduate students, studies have attempted to investigate the effect of interventions to increase undergraduates' ability to empathize and to take the perspective of their ITAs. Undergraduate students exposed to these interventions, versus those unexposed, rated ITAs with higher speaking competence (Yook, 1999). As undergraduates showed greater empathy for ITAs, their comprehension of course material also increased (Yook & Albert, 1999). These findings are consistent with the view that the factors influencing undergraduates' perceptions of ITAs may be anxiety, prejudice, and social stereotyping. Although scholars agree that undergraduates must learn more about intercultural communication (Inglis, 1993) and improve their listening skills (Rubin, 1992; Rubin & Smith, 1990), institutions rarely direct their efforts toward improving undergraduates' intercultural receptivity (Rubin, 1993).

Therefore, the current study also involves a socio-psychological intervention where ITAs and undergraduates have an opportunity to interact directly with each other to share their perspectives and participate in cooperative activities. The structure of the intervention was devised with conditions met for the contact hypothesis. Some ITA training programs have

involved undergraduates in a certain capacity (Civikly & Muchisky, 1991). Through the joint activities in the current study, it was expected that students would associate the positive feeling of success with each other.

Because of the inevitable contamination of accent perception with social stereotyping, it is necessary to adopt measurable parameters of accentedness through nonhuman instrumentation. Speech science has made progress toward identifying measurable features of pronunciation that affect comprehensibility. These more objectively measurable components of accent and comprehensibility are discussed in the next section.

Acoustic Measures of Accentedness

Experts are hardly unanimous regarding which specific linguistic aspects of NNS pronunciation affect intelligibility the most (Fayer & Krasinski, 1987). Until somewhat recently, most research examining characteristics of second language speech production had been concerned with phonemic segmental phenomena, that is, concerned with the "accuracy" of NNS' consonant and vowel formation (e.g., Cole, Jakimik, & Cooper, 1978). Currently, however, the consensus seems to be moving toward an appreciation of the role that differences in speaking rates, intonation patterns, and suprasegmentals (prosody) may play in intelligibility and listeners' assessments (Derwing & Munro, 1997, 2005; Munro, 1995). In other words, prosodic errors tend to contribute as much or more to perceived accentedness as do phonemic segmental errors (Anderson-Hsieh, Johnson, & Koehler, 1992; Anderson-Hsieh & Koehler, 1988; Munro & Derwing, 1995). Pickering (2004) argues:

> …linguistic competence is often perceived to be less crucial for functional competence than lexical or syntactic marking strategies... However, prosodic cues contribute

independently to the structure of the discourse, and they cannot be circumvented without a reduction in comprehensibility. (p. 39)

Speech rate (articulation rate and pausing) has been examined for its relation to communicative success. A slow speech rate is commonly cited as a facilitating characteristic of foreign language discourse (Derwing, 1990); i.e., accented speech heard at a reduced rate sounds less accented and more comprehensible than speech produced at a normal rate. Native listeners may prefer to hear accented speech at slower rates. Chaudron (1988) reviewed ten studies in which a slow rate of production was associated with benefits for listeners. Those benefits included increased time for processing, clearer boundary markers, and clearer pronunciation—relative to a rapid rate of speech. In contrast, Kelch (1985) argued that the common belief that "slower is better" indicates that a slow rate of speech production makes input cognitively rather than linguistically simpler. These studies, however, all looked at comprehension of non-native speakers of native speakers of English.

With respect to the comprehensibility of NNS speakers by native English speaking listeners, Anderson-Hsieh and Koehler (1988) likewise emphasized the importance of speaking rate for comprehension of heavily accented speech. Listening comprehension scores provided by native speakers of American English were significantly higher at regular speaking rates than at fast rates for native Chinese speakers as well as for native English speakers. This study further implied that speaking rate is critical for the comprehension of heavily accented speech.

Komos and Denis (2004) investigated the influential effect of speech rate on perceptions of NNSs' fluency. They explored which variables predict native and non-native speaking teachers' perception of fluency and distinguish fluent from non-fluent L2 learners. The two groups of ELLs were compared and their temporal and linguistic measures were correlated

with the fluency scores they received from three experienced native and three non-native speaker teacher judges. For all the native and non-native teachers, speech rate, including variables such as the mean length of utterance and phonation time ratio, was the best predictor of fluency scores. In short, prosodic deviance may affect comprehension more adversely than does segmental deviance.

Pauses are an especially important element related to speaking rate. In production and perception studies of pause boundaries in Dutch and in English, Swerts and Gerlykens (1994) found that longer pauses increase word boundary strength. Vaissiere (1983) suggested a universal tendency for pause-defined units in spoken discourse, with pauses between sentences being longer than pauses within sentences.

Analyses of nonnative speaker data show a qualitative difference in both placement and length of pauses which can materially affect the overall prosodic structure of the discourse (Pickering, 1999). Pausing is a distinctive feature of pronunciation that contributes to accentedness. In Pickering's (1999) study of two parallel lecture extracts, one given by NS teaching assistants and the other by Chinese international teaching assistants (ITA), she found that pauses in the NNS data were both longer and more irregular than those in the NS data. ITAs' pauses tended to break up conceptual units. Rounds (1987) found a prevalence of "empty pauses", regular moments of silence unrelated to boardwork and unrelated to dramatic effect. This erratic silence artificially increased the amount of silence in the discourse. These empty pauses were likely to be linked to negative perceptions of ITAs on the part of undergraduate students. In light of these differences in pause structure between NS and NNS TAs, the use of pauses to cue transaction boundaries warrants additional studies.

In context of pause, certain generalization shown has been made with types of semantic and grammatical content by transcribed recordings of NSs' English conversation (e.g., Pawley & Syder, 2000). Pawley and Syder suggest that pauses associated with selecting a single lexical unit within a phrase are often around 0.2 seconds, Pauses preceding the first clause in a multi-clause construction are seldom shorter than 0.5 and may even exceed 2 seconds. Chafe (1980) reports that pauses of less then a second (often less than half a second) typically mark the boundaries of separate successive "idea units" within a single episode or scene. He finds that pauses longer than a second, and especially pauses of more than two seconds, characteristically occur when the speaker is making a more radical conceptual shift, moving from one center of interest to another.

Pause-defined units in discourse fall into three major groups: (1) pauses of 0.8 seconds or longer constitute topic boundaries, called 'topic pauses' which clearly coincide with major semantic breaks; (2) those between 0.6 and 0.8 seconds, referred to as "substantial pauses," tend to coincide with single contours (i.e., the distinctive rising and falling patterns of pitch, tone, or stress) ; and (3) very short pauses between 0.1 and 0.5 seconds, identified as a 'sub-set' of the contour pauses, occurs in the context of incomplete syntactic structures (Brown, 1977; Brown & Yule, 1983) .

Nonstandard word stress can also undermine comprehensibility (Gallego, 1990). Lexical stress plays a central role in determining the profiles of words and phrases in accounts of metrical phonology (Hogg & McGully, 1987), and misplaced word stress appears to be more perceptually salient to NS listeners than are instances of mispronounced phonemes (Bond, 1999). Field (2005) tested both NS and NNS listeners in a psycholinguistic study in which lexical stress as well as vowel quality were manipulated on sets of disyllabic words. In his study, words were

recorded with normal acoustic cues and in conditions where stress was shifted leftward or rightward and in some cases, vowel quality altered. When tested with an intelligibility measure, while both groups of listeners were significantly handicapped by modified stress patterns particularly when the lexical stress was shifted to the right (i.e., to the second syllable), nonnative listeners demonstrated a lower success rate in identifying the words in the standard, unmanipulated group correctly.

Finally, intonation is likewise a key component of comprehensibility (Brazil, 1997). For example, intonation pattern characteristic of many East Asian speakers can cause U.S. listeners to lose concentration or to misunderstand the speaker's intent (Pickering, 2001). In particular, the choice of a rising, falling or level pitch on the focused word of a tone unit can affect both perceived information structure and social cues in L2 discourse. Pickering (2001) found critical differences when comparing ITAs with US instructors in the numbers of specific tone choices and the way these tones were used. Whereas NS teachers oriented their tonal structure toward a state of informational and social convergence with their student listeners, NNS teachers failed to exploit the English tonal system to increase comprehensibility and show involvement.

In addition to those violations of norms for English pitch movement (rising, falling, and steady), one of the major intonation features that affect NSs' comprehension on NNSs' speech is variation of pitch range. NNSs tend to manifest a narrow and compressed overall pitch range in comparison with NSs (Mennen, 1998; Pickering, 2004). NNSs tend to show a compressed pitch range and a lack of variety in pitch levels which can lead to a succession of mostly high or mostly low strings of syllables (Mennen, 1998; Pickering, 1999; Wennerstrom, 2000). For example, Wennerstrom (1994) measured the pitch of NS vs NNS in speaking tasks designed to elicit contrasts between high and low pitch in the areas of information structure and the boundary

structure. While the NSs manifested significant differences in their pitch to signal these contrasts in all the environments tested, groups of NNSs from Thai, Spanish, and Japanese backgrounds did not show as many differences. It was also reported that Finnish speakers of English used significantly narrower pitch ranges than native speakers of English (Hirvonen, 1967; Toivanen, 2001). This compressed pitch range is often combined with the assignment of equal stress to all syllables (Juffs, 1990). Very recently Kang, Rubin, and Pickering (under review) also found that this narrow pitch range factor exerted a negative effect on proficiency and comprehensibility ratings.

This contraction of pitch range affects NNSs' ability to indicate speech paragraph structure, normally perceived by the NS listener at least in part by an initial extra high pitch reset (Cutler, Dahan, & Donselaar, 1997; Nakajima & Allen, 1993; Swerts & Geluykens, 1994; Tench, 1996; Thompson, 2003; Wennerstrom, 2001; Yule, 1980). These high pitch reset levels are valuable cues to information structure in discourse. Many English language learners demonstrate no such paragraph-initial pitch changes, however (Wennerstrom, 1994, 1998).

As described so far, speech rate, pauses, stress, and intonation (particularly pitch range) appear to be important elements that can improve comprehensibility and intelligibility of non-native speakers of English. It is often believed that nonstandard word stress and intonation erodes the intelligibility of international teachers' speech (Anderson-Hsieh & Kohler, 1988; Gallejo, 1990). However a standard-like accent is not always sufficient for good communication; speakers who succeed in reducing the degree of "foreignness" in their accents based on objective measures are sometimes still heard as unintelligible (Llurda, 2000; Munro & Derwing, 1995). In other words, reduction of accent does not necessarily result in increased intelligibility *per se*.

Comprehensibility, in contrast to intelligibility, is often defined as the listener's ability to

understand the meaning of the word and utterance in its given context (Smith & Nelson, 1985). That is, comprehensibility betokens a higher level of understanding than intelligibility because it entails semantic processing in a context. "It refers to judgment on a rating scale of how difficult or easy an utterance is to understand" (Derwing & Munro, 1997, p.2). While speech comprehensibility can be a more objective characterization than intelligibility, its measurement nonetheless typically relies on expert ratings (Piske, MacKay, & Flege, 2001), and is hardly immune to the imposition of judgmental processes. Therefore, studies on comprehensibility have been mostly based on listener's perceptions of foreign accent (e.g., Munro & Derwing, 1995). The contribution to comprehensibility of exposure to accent, relative to other factors such as listener attitude, remains unknown.

New approaches to accent measurement are emerging. Whereas in the past comprehensibility was mainly measured with reference to listener dictation accuracy or expert ratings, elements of accent can now be detected by instrument and computer-assisted acoustical analysis. For example, Grover, Jamieson, and Dobrovolsky (1987) measured consistent differences in intonation contours signaling the continuation of an utterance in native English, French, and German sentences. When they measured fundamental frequency (F_0) slopes of lexical items preceding conjunctions, they observed that some native English learners of French tended to impose the intonational patterns of their L1 on their L2, and this resulted in a marked English accent.

In general, computer-assisted phonetic analysis (e.g., CSL: Computerized Speech Laboratory, or PRAAT: http://www.fon.hum.uva.nl/praat/) has assisted in characterizing different accents by examining patterns of fundamental frequencies or F_0 formants. In recent studies, it is becoming common to use these instrumentations rather than depending on

subjective judgments (see, for example, Ingram & Park, 1997; Levis & Pickering, 2004; Pickering, 2001, 2004; Schuetze- Coburn, Shapley, & Weber, 1991; Watt, 1997). Often the methodology must also incorporate discourse analysis to supplement the instrumental analyais, wherein an analyst identifies a pragmatic context in which a particular intonational contour would be expected. Following discourse analysis to identify an expected context for a particular intonation contour, computer-based analysis is used to confirm (or disconfirm) that the expected contour does indeed appear at that site in the speech stream (see Wennerstrom, 2001). Computer-assisted phonetic analysis likewise simplifies and renders more precise the task of ascertaining speech rate.

Overall, previous studies have typically described nonnative speech in terms of phonological characteristics (Shah, 2004)—when they have analyzed pronunciation at all--with only a few studies ascertaining the acoustic characteristics of those productions (e.g., Flege & Port, 1981; Magen, 1998). These acoustic studies were mainly studying single parameters of English, such as vowel duration differences in voicing contrast, acoustic vowel spaces, voice onset time of stop consonants, changes in fundamental frequency as related to intonation, or some measure of rate of speech (e.g., Flege, 1991; Flege & Port, 1981; Schmidt & Flege, 1996). Acoustic parameters combining rate, pauses, stress, and intonation of speech have not yet been well studied. In addition, most research utilizing acoustical analyses of speech are directed toward a linguistic sense of phonetics or phonology; they have not been applied to the assessment of oral performances.

Chapter Summary

One of the emerging questions in the field of language assessment pertains to rater characteristics. Rater variables that are likely potent include the effects of rater education and professional experience, rater nationality and native language, rater intercultural contact and exposure to NNS varieties, and rater formal training in linguistic courses. Besides rater background and demographic factors, raters' attitudes toward non-native speakers' English likely influences the rating process. Stereotypes about speaker identity prevent listeners from assessing speaker oral performances objectively. Therefore, it is warranted to adopt measurable parameters of accentedness through computer-assisted instrumentation. However, in studies in which instrumentation has been used to measure accentedness, research objectives have pertained exclusively to a linguistic sense of phonetics or phonology. Accordingly, applied linguists and language testers have yet to relate the acoustic characteristics of accented English to raters' assessments of L2 oral proficiency. Thus, the current study attempts to address the acoustic speech characteristics (speech rate, pause, stress, and intonation) of English-accented speakers as they influence native listeners' perception of accentedness and seeks to look for relationships between those acoustic measures of accentedness and impressionistic assessments of raters in oral performance.

CHAPTER THREE: METHODS

The present study investigated differences among raters of varying backgrounds in the assessment of oral performances. More specifically, this study has examined the effects of rater characteristics relative to the effects of objectively measured acoustic properties of accentedness of non-native speakers of English. Raters filled out a background questionnaire, completed a measure of linguistic stereotyping, and then rated 11 teaching presentations of international teaching assistants. Those samples of accented speech were subjected to acoustically measured suprasegmental analysis. Statistical procedures have compared the impact of rater characteristics with those of speaker accent on ratings of oral proficiency.

In this chapter, the research design of the present study has been described. The following issues are addressed: hypotheses, variables and their roles, instrumentation, sampling and data collection, and research design and data analysis.

Hypotheses

Recent research (Chalhoub-Deville & Wigglesworth, 2005; Kim, 2005) showed that there was a significant difference in ratings among raters from linguistically different backgrounds. A review of the literature (Bachman et al., 1995; Lumley& McNamara, 1995; Lynch & McNamara, 1998; Upshur & Turner, 1999) indicates that rating scores may be tied to particular characteristics of raters who score oral performances. However, the studies researched did not generally take account how each individual factor of raters' characteristics (e.g., rater education and professional experience, rater nationality and native language, rater intercultural contact and exposure to NNS varieties, and rater training) can affect the rating of oral assessments.

Furthermore, no previous study has examined relations between rater's assessment of oral performances and acoustic measures of accentedness in English-accented speeches in ratings of L2 oral performances. Therefore, this study tested the supposition that raters' background characteristics—including their rater training status—along with attitudes toward L1-accented English, influence their rating of oral performances to a degree comparable to the influence on ratings of acoustic properties of the speech samples analysis.

The following hypotheses were developed for this study:

H1: Oral performance ratings are inversely proportional to measured propensity to linguistic stereotyping.

H2: Rater background characteristics account for significant variance in ratings of oral performance

H2a: Oral performance ratings conducted by native speakers of English differ from ratings conducted by NNS raters (nondirectional hypothesis).

H2b: Oral performance ratings are directly proportional to the amount of self-reported contact by raters with NNS friends and acquaintances.

H2c: Oral performance ratings are directly proportional to the amount of rater formal training in language and linguistics.

H2d: Oral performance ratings are directly proportional to the amount of rater's experience in teaching/tutoring English as a second language or foreign language.

H3: The following suprasegmental properties of speaker's vocal productions account for significant variance in ratings of oral performance

H3a: The fluency factor directly predicts rated oral performance.

H3b: The level of the irregular boundary markers inversely predicts rated oral proficiency.

H3c: The incidence of the hesitation marking inversely predicts rated oral performance.

H3d: The acoustic parameters of speech rate, pauses, stress, and intonation contribute individually unique and statistically significant variance in predicting oral performance ratings.

H4: In ratings of NNS oral performance, the cluster of rater background characteristics, rater linguistic stereotyping, and the cluster of measured speaker acoustical properties all contribute unique and statistically significant variance in predicting oral performance ratings.

H6. Raters who received a socio-cultural *sensitization* intervention (training) are more lenient in oral proficiency ratings as compared with (a) prior to the intervention and (b) raters who did not receive such intervention.

Variables and Their Roles

The study included several sets of dependent and independent variables. The following table provides a summary of the major variables, their role in the research design, and their sources. The table can serve as a preview for the remaining parts of the methodological details in this chapter.

Table 3.1
Dependent/Independent Variables and Their Roles

Type of Variable	Variable Name	Data Collection Phase	Source (Instruments)
Dependent Variables	Rated speaker comprehensibility	Rating of ITA speech samples	Comprehensibility scale
	Rated accent standardness	Rating of ITA speech samples	Part I in CSEI**: items 10, 26, 35, 43
	Rated instructional competence	Rating of ITA speech samples	Part I in CSEI: items 11, 15, 20, 28, 33, 39, 45
	Rated English language proficiency	Rating of ITA speech samples	Part II in CSEI
	Rated superiority	Ratings of ITA speech samples	Part I in CSEI
	Rated social attractiveness	Ratings of ITA speech samples	Part I in CSEI
	Rater leniency (speaking ability) scores	Multi-faceted Rasch modeling	Subjected on oral proficiency ratings
Independent variables: *Speech sample supra-segmental*	Suprasegmental cluster 1: Acoustic fluency	Identified from the results of Hierarchical Cluster Analysis	PRAAT*** analysis for 12 speech samples
	Suprasegmental cluster 2: Irregular boundary	Identified from the results of Hierarchical Cluster Analysis	PRAAT analysis for 12 speech samples
	Suprasegmental cluster 3: Hesitation markers	Identified from the results of Hierarchical Cluster Analysis	PRAAT analysis for 12 speech samples
Independent variables: *Rater characteristics*	Rater native/non-native English language speaker status	Responding to the background questionnaire	Background Questionnaire items: 4-7
	Rater exposure to non-native English speaking friends and acquaintances in U.S	Responding to the background questionnaire	Background Questionnaire items: 19-25
	Rater linguistic sophistication (Formal training in language studies/Linguistics)	Responding to the background questionnaire	Background Questionnaire items: 14-16
	Rater amount of experience in English language teaching and tutoring	Responding to the background questionnaire	Background Questionnaire items: 17-18
	Rater reactions to courses taught by ITA (Self-report: grades hurt due to NNS' instructors)	Responding to the background questionnaire	Background Questionnaire items:21-23
	Index of rater reverse linguistic stereotyping-*superiority*	residuals from regression on the physical attractiveness	SEI* items administered during initial RLS

		item	measurement procedure
	Index of rater reverse linguistic stereotyping-*social attractiveness*	residuals from regression on the physical attractiveness item	SEI items administered during initial RLS measurement procedure

* Speech Evaluation Instrument (Zahn & Hopper, 1985)
** Comprehensive Speech Evaluation Instrument
***Software program for acoustical analysis of the speech stream

Instrumentation

Background Questionnaire

Building on the research of Rubin (1992) and others (e.g., Powers, Schedl, Wilson-Leung and Butler, 1999), a questionnaire was developed to obtain background information hypothesized to affect participants' rating of English accented speech performances (see Appendix A). The initial part of the background questionnaire included questions asking general information about participants such as sex, age, linguistic and ethnic background, educational background, and current and professional background. Next, more information was requested about raters' experiences with and general attitudes about nonnative speakers. Specific questions concerned (1) the extent of exposure to non-native English speaking friends and acquaintances; (2) length of sojourns in non-Anglophone nations; (3) formal training in language studies/ experience in English language teaching; and (4) reactions to the courses taught by international teaching assistants (ITAs). To ascertain the degree of contact participants have with nonnative speakers of English, raters were asked to respond to four 5-point Likert-type scale items (i.e., friends or social acquaintances, colleagues or business acquaintances, teachers/teaching assistants, and others). Items were scored 1= very infrequent/several times a year or less to 5=

very frequent/daily or almost daily. This scale item was adapted from Powers et al. (1999) and Derwing and Munro (1997).

Comprehensibility Scale

The speaker's comprehensibility measure (Appendix B) utilized five 7 point scales' *semantic differential scales* (e.g., see Kerlinger, 1975). The instrument was initially designed for an ETS contracted project and was used for this study as well. It was expanded upon Derwing & Munro's (1995, 1997) single item. These five items were described by incorporating conventional conceptualization of comprehensibility and intelligibility (Munro & Derwing, 1995; Derwing & Munro, 1997). That is, comprehensibility in this scale refers to "judgment on a rating scale of how difficult or easy an utterance is to understand" (Munro & Derwing, 1995, p. 2)..

Each comprehensibility item posed polar opposite descriptions at either end of seven equal-appearing intervals (e.g., "hard to understand : : : : : : : easy to understand"). The five items were: easy/hard to understand, incomprehensible/highly comprehensible, little effort/lots of effort to understand, unclear/clear, simple/difficult to grasp the meaning. Note that as Derwing & Munro (1997) point out, the term comprehensibility has been used in many different ways (see Gass and Varonis, 1984 where it was used in a sense of intelligibility; Smith (1992) regards it as being at a higher level of understanding than intelligibility) and that conceptualization is adapted here.

Cronbach's alpha for the five-item comprehensibility measure was .94. Accordingly, the sum of these five items was utilized as a composite measure for subsequent analysis. The composite comprehensibility and proficiency measures correlated at $r = .83$ ($p < .01$).

Reverse Linguistic Stereotyping Measure (RLSM)

Research on language attitudes has utilized a variety of dependent variables to gauge the dimensions of perception whereby listeners judge speakers (Edwards, 1982). Rubin (1992; Rubin & Smith 1990) built on the matched-guise technique originally developed by Lambert and Tucker (1972) as a means of ascertaining the degree of linguistic stereotyping manifested by listeners. The RLSM utilized Zahn & Hopper's (1985) Speech Evaluation Instrument (SEI).

For the measure of reverse linguistic stereotyping, first listeners heard two different sections of a 4 minute audio tape simulating a portion of a college lecture on astronomy, previously used in similar language and attitude research (e.g., Rubin, 1992). They listened to the lecture segment once with a Caucasian face projected, and once with an Asian face (counterbalanced for order). The lecturer for both listening passages was the same speaker of Standard American English, originally from a small town in Michigan, who was a teacher of speech communication and was acknowledged by peers to have a particularly clear speaking voice. However, the speaker was identified via fabricated photographs and dossiers as either an international teaching assistant or a US teaching assistant. That is, as subjects listened to different segments of the lecture, one of the two slide photographs representing the instructor was projected as a screen each time. Instructor ethnicity was operationalized by projecting a photograph of either a Caucasian or an Asian man.

To avoid confounding ethnicity with physical attractiveness, both models were similarly dressed, were of similar size and hair style (i.e., dark-haired), and were photographed in the same setting and pose (standing in front of a whiteboard). There was also a short distracter played between the two target listening tasks. All listeners heard a distinctly East Asian speaker of intermediate intelligibility. The lecture topic was held constant by selecting one article on a

science (i.e., topic, *constellations* originally appearing in the *New York Times*), which pertained to the description of galaxies and clusters (Appendix C-3). After hearing the lectures, participants completed a Speech Evaluation Instrument (SEI) in Zahn & Hopper (1985). They also completed a partial cloze test of listening comprehension which is not analyzed in this dissertation.

The reverse linguistic stereotyping procedure ultimately yielded two dependent variables that index distinct dimensions of linguistic stereotyping: *superiority-RLS* and *social attractiveness-RLS.* Superiority and social attractiveness, in turn, are two of three subscales that derive from the SEI. The third subscale that factored out in the original development of the SEI is *dynamism*. The items included in each subscale — as specified in the original instrument development by Zahn and Hopper (1985) — are shown in table 3.2. Internal consistency reliability (Cronbach's alpha) was calculated for each of the three subscales separately for (1) the Euro-American guise and (2) for the East-Asian guise ratings. All reliabilities were considered to be acceptable (>.80) as shown in Table 3.2 ; therefore, the total sum of each component item comprised composite measures of each of the three subscales, and were used as predictors in final regression analyses.

Table 3.2
Speech Evaluation Instrument (SEI) Items and Internal Consistency Reliabilities

Reliability (Cronbach's alpha)	Superiority	Social Attractiveness	Dynamism
	advantage/disadvantaged	kind/unkind	lazy/energetic
	poor/rich	unfriendly/friendly	unsure/confident
	unclear/clear	cold/warm	passive/active
	complete/incomplete	unappealing/appealing	talkative/shy
	white/blue collar	unlikable/likable	strong/weak
	intelligent/unintelligent	bad/good	enthusiastic/hesitant
	fluent/not fluent	sour/sweet	
	disorganized/organized	hostile/good natured	
	uneducated/education	nice/awful	
	illiterate/literate	considerate/inconsiderate	
	lower/upper class	honest/dishonest	
	experienced/inexperienced		
Euro-American guise	.82	.91	.80
East Asian guise	.81	.93	.80

RLSM was indexed by subtracting speech evaluations accorded to the Euro-American guise from those accorded to the East Asian guise (since the actual speaker in either event was the identical NS). A manipulation check indicated that the East-Asian guise was perceived to be a "person of color" more so than was the Euro-American guise [M_{Asian}=5.95, M_{Euro} =3.69; t_{68} = 6.72, $p < .000$]. In pre-testing, the two photographs were not judged to be significantly different in terms of physical attractiveness, however in actual data collection the Asian guise was

perceived as significantly more attractive than the Euro-American. In fact, physical attractiveness can have strong influence on social judgments (Riniolo, Johnson, Sherman, & Misso, 2006). Therefore, any differences between the photographs in rated physical attractiveness of these photographs were compensated for statistically by using rated physical attractiveness as a covariate. Specifically, composite variables for *superiority*-RLS and for s*ocial attractiveness*-RLS were created by the following steps: (1) obtained unstandarized residuals from regression of SEI composite constituent items on the values of the physical attractiveness item → (2) subtracted the SEI 2 residuals (Asian guise) from the SEI 1 (American guise).

Speech Evaluation Instrument (SEI)

The SEI (Appendix C-1) developed by Zahn and Hopper (1985) has been used in dozens of studies of language attitudes (e.g., Cargile, 2002; Dailey-O'Cain, 2000; Luhman, 1990; Rubin 1992; Rubin & Smith, 1990). It typically factor analyzes into three dimensions: (1) *superiority*, (2) *social attractiveness*, and (3) *dynamism*. Interspersed among the SEI items were additional semantic differential items checking the ethnicity manipulations of the photographs ("Caucasian/European ethnicity: : : : : : : [3]Oriental/Asian ethnicity") and measuring impressions of accent ("Speaks with American accent: : : : : : : Speakers with foreign accent") and of teaching competence ("Poor teacher: : : : : : : Effective teacher"), as well as several unrelated filler scales. In all, the instrument contained 35 semantic differential items.

Cloze Test

[3] Due to the software (Facilitate) limitation used for the online rating, the instrumental format of semantic deferential measures was slightly modified as follow: Caucasian/European ethnicity ___/___/___/___/___/___/___ Oriental/Asian ethnicity.

A cloze test of listening comprehension was adopted from Rubin (1992) (See Appendix C-2). To measure listening comprehension, a cloze test of the speech texts was constructed. Subjects were presented with a written transcript of the lectures with approximately every 7th word deleted, saving the first sentences, which were kept intact. Only exact recall was scored as correct. Nevertheless, due to the primary interest in the contribution of raters' background characteristics and acoustic features of accented English to oral proficiency ratings, the results of the cloze test were excluded for the further analysis.

Dependent Variables

Previous studies emphasized the assessment of sociolinguistic skills and knowledge of instructional communication that the ITA needs to master in order to be a competent instructor in an American university setting (Fox & Gay, 1994; Sarkisian & Maurer, 1998). In contrast, the present study focused on both language assessment (Elder, 2001; Jenkins & Parra, 2003) and instructional assessment (Douglas, 2001; Pae, 2001) of ITAs.

Composite Speech Evaluation Instrument (CSEI)

The composite speech evaluation instrument was developed by the researcher (Appendix D), extending the SEI of Zahn and Hopper (1985). Those earlier scales measured general impressions of speakers. The CSEI in this study were administered after participants heard each of the ITA speech samples. Because the ITA speech samples were selected from instructional encounters, and because those speech samples were subjected to intensive acoustical analysis, the CSEI was designed to include more specific items pertaining to linguistic competence and to instructional competence. This instrument functioned as the primary tool to measure participants' ratings of ITAs' instructional skills and to measure perceptions of the English language abilities of those speakers.

Using sets of semantic differential items, the CSEI consisted of the following four segments: (1) ratings of English language proficiency, (2) ratings of instructional competence, (3) ratings of accent standardness and (4) the SEI (Zahn & Hopper, 1985). The four rating instruments were randomly interspersed in a single instrument tool—the CSEI (see Appendix D).

English proficiency rating scales. The English language proficiency rating scale consisted of 8 semantic differential items assessing instructor's pronunciation/ accent, grammar, vocabulary, overall communication skills, etc. It was developed by the researcher based on Kim (2005). The instrument asked raters to evaluate ITAs' English language proficiency on 7-point scales for the following items: pronunciation/accent, grammar, vocabulary, overall communication ability, the use of efficient expression, and the effective word choice. Finally it inquired about the rater's own opinion about what constitutes English proficiency as emphasized by Reed and Cohen (2001). The coefficient alpha (internal consistency reliability) values of the subscales were .92 and .93 for the pre-test and post-test administrations, respectively. Accordingly, ratings on the eight English proficiency items were summed into a single scale measure.

Instructional competence scales. The instructional competence rating scale was composed of 9 semantic differential items. It was adopted from a rating scale of ITAs' proficiency and instructional quality in Sarwark, Smith, MacCallum, and Cascallar (1995). Examples of the questions were "effective teacher : : : : : : :ineffective teacher" or "approachable : : : : : : : unapproachable". The internal consistency reliability coefficient of this scale was .92 for both pre test and posttest administrations. Accordingly, ratings on the nine instructional competence items were summed into a single scale measure.

Accent standardness ratings. The accent standardness rating scale (e.g., Speak with foreign accent : : : : : : : Speak with American accent) was composed of four semantic differential-type items. It was an extended version of Munro and Derwing (1995)'s single item. Its internal consistency was marginally acceptable: .70 for the pre-test and .71 for the post-test administrations. . Accordingly, ratings on the four accent rating items were summed into a single scale measure. Correlations among the four dependent variable measures were computed and turned out to be reasonably correlated, with Pearson *r* values of .67 and above, all at the $p <$ .01 level.

Speech evaluation scales. The subscale structure found by Zahn and Hopper (1985) was adopted in the present study. The Superiority subscale (e.g., uneducated : : : : : : : educated or intelligent : : : : : : : : unintelligent) consisted of twelve items Internal consistency (coefficient alpha) was .80. Accordingly the twelve Superiority items were summed into a single measure. The Social Attractiveness subscale (e.g., kind : : : : : : : unkind or warm : : : : : : : cold) was composed of 11 items. Its internal consistency was also .80 and above. Finally, the Dynamism subscale (e.g., lazy : : : : : : : energetic or passive : : : : : : : active) was made up of 6 items with coefficient alpha of .80) . Therefore, the total scores of these measures were used for the final analysis.

Rater Leniency Scores

One dependent variable used in portions of this study was rater leniency-severity score. In order to obtain severity scores for each of 70 raters across the different performance dimensions were elicited from Multifacet Rasch analysis. A computer software program, Linacre's (2005) FACETS for Windows version 3.58.0 was utilized for the analysis. Through FACETS, speaking scores were analyzed based on a number of facets in the performance setting.

Parameters for the pre-test and post-test Rasch model were speakers (n=11) and raters (n=63) because only 63 raters completed the post-test speech rating. That is, two facets were arranged as rater in the first column and speaker in the second column. Oral proficiency ratings was the only dependent variable (among the seven response measures) subjected to this Rasch analysis. The oral proficiency rating scores was selected because that particular variable was considered to be the major dependent variable of this study. Rater ability scores, *Fair-M average scores*[4], for the leniency or severity of the raters as well as speaker ability scores were obtained both for the pre-test and the post- test. The rater leniency scores were used for the final analysis. Data in this study were structured in the opposite manner from conventional Rasch models, i.e., from the higher-rater-assigned scores to lower-rater-assigned scores. Therefore, the high Fair-M average scores stand for raters' leniency ability rather than severity ability. Changing the directionality in this manner eased the interpretation of results, to make it more intuitive.

Qualitative Data Collection

Additionally, two open-ended questions were appended to the rating procedures for the purpose of this study. The first open-ended question asked participants to comment in their own words about each ITAs' English proficiency. The second open-ended question inquired what nationality the participant presumed the ITA to be (as in Lindemann, 2005). Nationality attributions are not analyzed in this dissertation report.

Brief online interviews were also conducted as a qualitative supplement to the main quantitative data collection. These interviews were conducted after participants completed both

[4] The "fair average" transforms the Rasch measure back into an expected average raw response value. This value is in a standardized environment in which all other elements interacting with this element have a zero measure or the mean measure of all elements in their facet. This is "fair" to all elements in the facet, e.g., this adjusts raw ratings for severe and lenient raters. This enables a "fair" comparison to be made in the raw score metric, in the same way that the measure does on the linear latent variable. *Fair-M* uses the facet element means as the baseline. *Fair-Z* uses the facet local origins (zero points) as the baseline (Linacre's FACETS manual, 2005)

phases of the rating tasks. Raters were asked to respond to the brief online interview (MSN Messenger or Gmail Chat Window). Because of the low participation rate, however, that is, only two persons who agreed to the interview, the researcher emailed 68 participants an open-ended questionnaires separately prepared for trained raters and untrained raters respectively (Appendix E) . When raters' responses required clarification, several iterations of correspondence took place between the researcher and the specific raters.

The open-ended questionnaire interview procedure in this study was considered to be a form of intensive debriefing (Hocking, Stacks, McDermott, 2003). Hocking et al. argue that debriefing should not only establish or maintain a favorable relationship between participants and researchers, but also provide researchers with "feedback relevant to the efficacy of the research procedures" (p. 64). The questions in the open-ended questionnaires or the discussions during the interviews were constructed in such a way so as to gain insights into the rating process of the speech performance samples , to examine raters' priority in speech assessment of accented Englishes, and − for trained raters in particular − to elicit the raters' impression on one-hour intercultural intervention with ITA. (See regular debriefing statement in Appendix I). The following questions were asked:

- Could you tell us your overall impression about the online rating? Any difference between Phase I and Phase II?
- **How was your impression about the informal meeting, one hour intercultural intervention with international teaching assistants? (for trained raters only)**
- Please tell us some interesting cases of your ratings, if any. What and why you provided that rating?
- Please explain the way you approached and performed the assessment of the speech samples online. How did you feel about this online speech rating?
- What components (English language proficiency, instructional completence, instruction quality, and accent standardness) do you believe were the most important in your rating of the speech samples? Why?

    - English language proficiency (e.g., 1=low proficiency, 7=high proficiency)
    - Instruction competence (e.g.,1= effective teacher, 7=poor teacher)
    - Comprehensibility (e.g., 1= easy to understand, 7= difficult to understand)
    - Accent standardness (e.g., 1= speak with foreign accent, 7= speak with American accent)

Interview data were used only as supportive evidence to elaborate and help explain the quantitative data results (see discussion of this mixed methods model in Creswell and Clark, 2007).

Sampling and Data Collection

Participants

International teaching assistants (ITAs) were not the actual participants in this study. Rather, they served as the sources of the speech performance samples which were the objects of partifipant judgments. The need for rating ITA competence via performance tests of their instructional communication is well established (Briggs, 1994). Bailey (1984) argued that ITAs

must be able to effectively use English to present their expertise in the subject matter in an organized fashion. Noor (1995) asserted that there is an urgent need to assess ITAs' oral English proficiency and instructional ability due to the increasing number of international and non-native speakers of English in the teaching force. As Williams (2006) suggested, efficient evaluation of ITAs can give the ITA needed feeback for improving verbal and instructional skills.

Native English speaking (NS) undergraduates served as raters (participants), because they are the intended audience for ITA English discourse in the classroom. Therefore they were the natural evaluators in this situation. Besides, it has been argued that ITAs teaching in the U.S. must be trained to accept and understand student ratings, since such practices may not be common in the ITAs' educational background but will be routine in U.S. (Bauer, 1996). Once the ITA learns the value of student ratings, undergraduates should be incorporated into the training process so that the ITA becomes comfortable interacting with undergraduates (Williams, 2006). Moreover, through training experiences, the undergraduates themselves can develop some intercultural understanding to equip them for having an ITA in content area classes (Damron, 2003). Thus, the involvement of undergraduates in the ITA training process mutually benefits both the ITA and the students as the two seek to gain a better understanding of the other's frame of reference.

The design of the proposed study revolved around two phases of data collection. In the first phase—pretesting—raters listened to and evaluated eleven samples of ITA classroom discourse. In the second phase—posttesting which took place about 7-8 weeks after pretessting —the identical raters evaluated the same eleven speech samples a second time. At posttest, however, a subset of 29 raters had participated in a social-psychological intervention, in which undergraduates interacted in a structured but pleasant activity for an hour with ITAs.

In Phase I, 76 undergraduate participants were selected such that they could be expected to collectively vary across the predictor variables: (a) varying English language speaker status (native/non-native), (b) composite index of exposure to non-native English speaking friends and acquaintances, and (c) formal training in language studies/experience in English language teaching. Participants were recruited to represent a continuum of variation across those background dimensions. They were recruited by advertising in the campus newspaper and in world languages classes on campus. (See Appendix H for the recruitment letter). The attempt was to identify a sample that spans a range of second language proficiency from beginner to advanced learner. No participants with previous experience in standardized assessment rating activities were selected. Participants were remunerated at a rate of $8/hour.

Although 76 participants began the study, usable sets of pretest and posttest data were obtained from just 70. The distribution of rater background characteristics of interest appears in table 3.3.

Table 3.3
Selected Participant Background Characteristics

		N=70 (14 male/ 56 female)	Weeks taught/tutored		Linguistics/ TESOL classes		% Weekly contact with NNSs	
			M	*SD*	*M*	*SD*	*M*	*SD*
Native speaker status	NS	48	13.68	34.10	1.15	2.63	6.04	11.21
	NNS	22	12	19.75	0.84	1.64	49.54	29.68
ESL/FL teacher or tutor	Yes	14	52	55.12	3.55	4.37	19.74	29.02
	No	56	0	0	0.50	1.06	17.90	24.07

Those 6 dropouts (5 females and 2 male) were all NSs who have not taken any linguistic classes at all. Most of them had no ESL/EFL teaching experience except one who had taught English as a second language for 24 weeks. Their contact with NNSs was about 6 % per week.

Speech Performance Samples

To obtain the speech samples which the participants rated, 11 ITAs[5] gave permission to utilize their oral performance presentations for this study. To avoid confounding of acoustic patterns with speaker gender, and to help assure easily measurable fundamental frequency patterns, all speech samples were obtained from male ITAs . The samples were excerpted from the ones given in conjunction with required classes in instructional oral English for ITAs who failed to achieve a score exceeding 45 on the SPEAK test. The SPEAK test is comprised of retired versions of the Test of Spoken English, originated by the Educational Testing service but scored locally, The simulated instructional presentations described a concept from each presenter's major course of study. They lasted approximately 20 minutes, however only a 4-5 minute segment was selected from each for acoustical analysis.

Selecting lecture segments for acoustical analysis and subsequent rating is similar to standard elicitation procedures in pausological research (e.g., Riazantseva, 2001). That is, segments consisted of continuous speech with no interruptions of others. An uninterrupted, narrative segment of discourse is necessary because computer analysis of speech phenomena becomes very difficult in an interactive task, where it might happen that the two speakers talk simultaneously (Kormos & Denes, 2004). Therefore, the segments were selected to include the necessary continuous linguistic environments to observe proportion of native-like pauses, speech

[5] Originally 12 ITAs' speech performances were acoustically analyzed, but one ITA withdrew his consent form, and therefore his speech file was excluded from the subsequent rating procedures.

rates, stress, and intonation (i.e., pitch range). That is, the linguistic context required a 4-5 minute segment of ITAs' narrative lecture did not have any interruption of students' questions, responses, and laughter. In addition, in choosing ITAs' speech samples, the TAST (TOEFL® Academic Speaking Test) scores of the ITAs were collected to independently identify ITAs' communication skills.

Three male US NS TAs were also audio-recorded using wireless microphones while lecturing in their own classrooms. The US TAs had been chosen as TA mentors, and were therefore relatively experienced "model" TAs. These NS recordings provided a baseline of standard native speaker performances for acoustic analysis. For each model US TA, a chunk of 3-4 minutes of monologic lecture-like explanation from within the lesson, was subjected to acoustical analysis. These baseline samples were used to examine in a descriptive manner the differences between native and nonnative speaker realizations in terms of pause structures and speech rates.

Suprasegmental Measures: Acoustic Analysis of Speech Rates, Pause Structures, Stress, and Intonation (i.e., Pitch Range)

Suprasegmental profiles were created using selected quantifiable acoustic measures of speech rate, pausing, word stress, and pitch range. As these acoustic parameters are gradient in nature, measurements were taken of the range of baseline native speaker realizations of these significant features and the degree of difference between native and nonnative speaker realizations were calculated (see Table 3.4). The specific acoustic indices reported here were chosen based on previous work investigating phonetic cues that indicate discourse structure in native speaker models and and that revealed suprasegmental correlates of perceived comprehensibility (e.g., Pickering, 1999; Riggenbach, 1991; Towell, Hawkins, & Bazergui,

1996). The PRAAT computer program (Boersma & Weenink's PRAAT for Window, 2007), assisted in the analysis of each of the acoustic indices. For each speaker, 4-5 minutes of continuous speech were identified and demarcated with the program cursor. The discourse segment was selected from the medial portion of the lesson. Segments that were noisy or in which speakers exhibited a creaky voice or vocal fry had to be excluded, as the sound extraction function of the PRAAT is unable to read this data. All the speech samples were transcribed in order to gain precise acoustic measures.

Fluency Measures

According to Kormos and Denes' (2004) findings, for both native and non-native teachers, speech rate, mean length of utterance, and phonation time ratio are the best predictors of perceived oral fluency. They explored variables that could predict native and non-native speaking teachers' perception of fluency and distinguish fluent from non-fluent L2 learners. They investigated speech samples collected from 2-3 minutes of 16 Hungarian L2 learners' speeches. Six experienced teachers rated these samples for overall fluency.

The present study utilized the Kormos and Denes (2004) predictors, and added others. Rate of speech was calculated according to the method recommended by Riggenbach (1991) and Kormos and Denes (2004). The total number of syllables produced in a given speech sample was divided by the amount of total time required to produce the speech sample, (including pause time) expressed in seconds. In calculating articulation rate, the total number of syllables produced in a given speech sample was divided by the amount of time taken to produce them in seconds. Unlike in the calculation of speech rate, pause time was excluded for this variable. Articulation rate was expressed as the mean number of syllables produced per minute over the total amount of

phonation time. Following Riggenbach (1991), for articulation rate all semantic units were counted, ‘‘including filled pauses and partial words (using the criterion that partial words contain not just an initial consonant but also a vowel and thus are recognizable as words)’’ (p. 428).

The mean length of runs was calculated as an average number of syllables produced in utterances between pauses of 0.1 seconds and above. Towell et al. (1996) point out that there has been an ongoing debate among researchers about the cut-off point of pause length. Since the analysis of this study has adopted the pause unit model of Brown and Yule (1983), pause length of 0.1 was the cut-off point for the analysis. Finally, phonation-time ratio was calculated as “a percentage proportion of the time taken to produce the speech sample’’ (Towell et al., 1996, p. 91).

Pausing Measures

For the measures of pause structure, the pause unit model of Brown (1977) and Brown & Yule (1983) was adopted. First, in analyzing pauses, only pauses over 0.1 seconds were considered. Pauses shorter than 0.1 seconds were considered micro-pauses and were not regarded as hesitation phenomena (Riggenbach, 1991). The total number of pauses was divided by the total amount of time spent speaking (expressed in seconds). The mean length of pauses was calculated by dividing the total length of pauses above 0.1 seconds by the total number of pauses above 0.1 seconds. The total number of filled pauses—such as *uhm, er, mm*—was divided by the total amount of time expressed in seconds.

In addition, the proportion of irregular topic boundary pauses in utterances was calculated by dividing the number of irregular topic pauses by the total number of topic boundary pauses. The *regular* topic boundary pauses are defined as pauses of 0.8 seconds or longer, which clearly coincide with major semantic breaks (Brown, 1977; Brown & Yule, 1983). Accordingly, this

study analyzed the location of these topic pauses in TAs' speech performances by looking at the semantic contexts and units of speech, determined the status of regularity or irregularity, and calculated the proportion of the irregular topic pauses in each of 11 ITAs' speech; i.e., the number of the irregular topic pauses were divided by the total number of topic boundaries. The places of these irregular topic pause production were analyzed and discussed qualitatively in the discussion section where other empty pauses or silences were illustrated with comparison of Rounds (1987).

Word Stress Measures

Stress and pitch measures were all assessed using a combination of auditory and instrumental analysis. While auditory analysis is concerned with the hearing and comprehension of phonetic sounds of words of a language, instrumental analysis is involved in the analysis of signals by computers using computer software such as PRAAT. That is, underlying all the stress and pitch measures was an initial identification of prominent and non-prominent syllables using both auditory and instrumental analysis of acoustic characteristics of pitch and amplitude. Prominent syllables were stressed syllables which showed characteristics of longer duration, higher pitch, and grater amplitude than unstressed (non-prominent) syllables. Every clause was analyzed for pitch using the PRAAT pitch function. Pitch range was set from 75 Hz to 300 Hz in PRAAT for the optimized intonation of male speakers. Each measure was then produced.

Levis and Pickering (2004) demonstrated that when the same utterances are produced as part of a short spoken text as opposed to isolated utterances, there is a consistent difference in word stress pattern. Therefore, when determining prominent syllables or words, discourse contexts were considered. Pace and space for stressed measures directly followed Vanderplank's (1993) approach. Pace was measured by calculating the average number of prominent syllables

per run. The space measure was obtained by calculating the proportion of prominent words to the total number of words.

Intonation Measures (i.e., Overall Pitch Range)

In the case of overall pitch ranges, the measures were conducted through the calculation from pitch on prominent (stressed) syllables with point of F_0 maximum and minimum for initial 100 words selected among 4-5 mins utterances. (However, these 100-word selections were, in fact, mostly drawn from the middle parts of the original lectures; therefore they are not susceptible to interpretation as reflecting only utterance-initial data). The prominent segments were identified by the features of prominence, or fundamental frequency (F_0) peaks which distinguish prominent syllable from the surrounding content. The level of the pitch was described following Brazil (1997).

An example unit is shown in figure 1, with prominent syllables represented in CAPS. The pitch on all prominent syllables was measured and given in Hz.

Excerpt 3.1

(.10) //todAY I'm not GOing to // (.47) // TEll you about the mAp of the UNIted
154.9 147.2 142. 145.5 124.48
StATes// (.22)
111.3

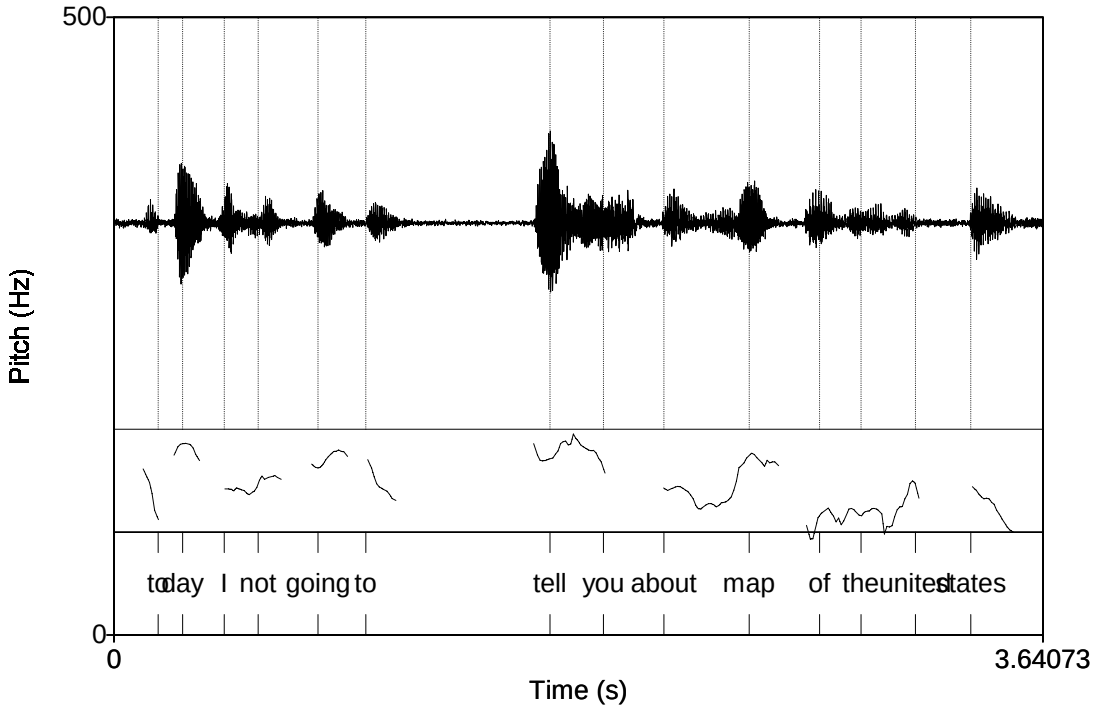


Figure 3.1
An Example of the Transcription of Shown for Pitch Ranges in PRAAT

(Note: Due to the contraction of the spectrogram itself for the limited space, the pitch contour and phonological segments may not appear to be exactly paralleled)

The following is the summary of 12 acoustic measures completed for this study.

Rate Measures

1. Syllables per second: This is a measure of the mean number of syllables produced per second for the 4 minutes sample.
2. Articulation rate: This is a measure of the mean number of syllables produced per minute over the total amount of time talking and excluding pause time.
3. Mean Length of Runs: This is a measure of the average number of syllables produced in utterances between pauses of 0.1 secs and above.

4. Phonation time Ratio: This measure expresses the percentage of time spent speaking as a proportion of the total time taken to produce the speech sample.

Pause Measures

1. Number of silent pauses: This measure reflects the number of silent pauses per 4-5 minute speech.
2. Mean Length of pauses: This measure reflects the total length of pauses of 0.1 or greater by the total number of pauses of 0.1 or greater.
3. Number of filled pauses: This measure reflects the number of filled pauses (but does not include repetitions, restarts or repairs) per 4-5 minute speech.
4. Mean length of filled pauses: This measure reflects the average of length of filled pauses occurring per 4-5 minute speech.
5. Proportion of irregular topic boundary: This measure reflects the proportion of irregular topic boundary pauses (between pauses of 0.8 secs and above) in entire utterances.

Stress Measures

1. Pace: This measure reflects the average number of prominent syllables per run (Vanderplank, 1993).
2. Space: This measure reflects the proportion of prominent words to the total number of words (Vanderplank, 1993).

Pitch Measure

1. Overall Pitch Range: This measure reflects the pitch range of the sample based on the point of F_0 minima and maxima appearing on prominent syllables per sample.

Table 3.4 shows the eleven ITAs' background information with performance scores on each of the linguistic/acoustical measures. Although the ITAs were by no means homogenous in their speech characteristics, it can be readily seen from this table that ITAs were considerably less fluent, spoke with less intonational variation, made more frequent or longer silent pauses, and had more frequent use of hesitation markers, compared to the US TAs. ITAs also spoke noticeably slower than US TAs and they paused for considerable amount of time without any dramatic effects or changes of semantic units.

Table 3.4
ITA Speech Performance Samples and their Suprasegmental Profiles

Partici-pants	Nation-ality	TAST score1	Speech Rate	Articu-lation Rate	Mean Length of Runs	Phona-tion Time Ratio	Silent Pause per min	Mean Length Silent Pause	Filled Pause per min	Mean Length Filled Pause	Pace	Space	Ratio Ire-gular topic pause	Overall Pitch Range
ITA1	Saudi Arabia	21.25	2.99	4.27	5.90	69.90	30.83	0.58	6.10	0.31	30.83	0.24	0.4	89.7
ITA2	Nepal	18.75	2.69	4.23	5.19	63.68	31.70	0.69	3.10	0.26	30.54	0.27	0.44	82.9
ITA3	Saudi Arabia	18.75	2.26	4.10	4.50	55.07	30.43	0.89	3.53	0.45	31.09	0.29	0.65	74
ITA4	South Korea	SPEAK 40	2.75	4.45	4.69	61.10	35.40	0.66	7.67	0.27	29.49	0.25	0.5	79.8
ITA5	Japan	16.25	2.14	3.84	3.86	55.79	33.83	0.78	5.38	0.38	28.14	0.30	0.56	71.8
ITA6	Russia	23.75	2.67	4.30	5.80	62.05	28.20	0.80	3.98	0.21	26.21	0.26	0.33	100.2
ITA7	India	25	3.05	4.69	4.30	64.91	43.29	0.49	7.62	0.38	28.60	0.24	0.3	113.7
ITA8	China	20	2.84	4.41	4.88	64.33	35.46	0.60	9.14	0.35	32.40	0.26	0.43	82.9
ITA9	Japan	16	2.25	3.61	3.76	62.40	36.50	0.61	9.20	0.48	26.38	0.35	0.67	62.5
ITA10	China	14.75	2.20	3.81	4.23	57.80	31.83	0.79	15.91	0.38	26.53	0.32	0.8	40.8
ITA11	China	12.25	1.23	3.11	4.12	39.74	18.34	1.97	3.72	0.33	16.61	0.38	0.7	43.4
US TA1	US	NA	4.82	6.20	11.11	77.81	22.75	0.52	2.99	0.34	35.07	0.18	0.18	130.07
US TA2	US	NA	4.04	4.94	8.61	81.30	28.40	0.38	1.44	0.39	37.97	0.25	0.25	194.42
US TA3	US	NA	4.39	5.09	12.02	86.15	22.22	0.37	1.44	0.14	46.19	0.26	0.13	134.7

[1]maximumTAST score is 30. ITA4 had no TAST score, only SPEAK, for which the maximum score is 60.

Hierarchical Cluster Analysis of the Suprasegmental Variables

Because a set of 12 linguistic predictors would be far too cumbersome to use in any sort of follow-up analysis, it was necessary to adopt a data-reduction strategy to combine those variables into some smaller number. Therefore, hierarchical cluster analysis (HCA) was conducted as a preliminarily analysis in order to reduce the number of acoustic variables. HCA is a method for finding relatively homogeneous clusters of variables based on measured characteristics. It is most useful when you want to cluster a small number of objects (SPSS 15, HCA tutorial). It starts with each variable in a separate cluster and then combines the cluster variables sequentially, reducing the number of clusters at each step. Proximity scores correspond to similar distributions of the variables. The function of HCA is thus analogous to factor analysis.

Precedent for HCA in applied linguistics research is applications of cluster analysis on student activities and behaviors in classroom observation sessions (e.g., Gayle, 1980; Ross, 2001) where the ungainly matrix of observation data is reduced into smaller subset of interpretable summaries of the similarities/ dissimilarities among the observed classes. The method adopted here utilized the *Between-groups Linkage* (i.e., SPSS "cluster method"), to optimize homogeneity with clusters of variables. All the acoustic measures were first scaled to standardized scores. As described in the *Results* chapter, the twelve acoustic measures were found to merge into three logical clusters: (a) the fluency factor, (b) irregular boundary, and (c) hesitation markers.

Rater Training: Intercultural Contact Intervention

The literature on effects of rater training (e.g., Weigle, 1998) generally compares the ratings of judges who had received rating with those of untrained raters. The literature contains

few, if any, true experiments in which the pre- and post-training ratings of one randomly selected group of raters is compared with two sets of ratings of a randomly selected group of raters who had received NO intervening training. The findings of the study could help fill that gap.

Training in the present study, however, departed entirely from typical indoctrination about the scoring rubric and then calibration with anchor-point speech samples. Because so much of the rationale for this study rested on linguistic stereotyping effects, the training was more like a social-psychological inoculation against linguistic stereotyping. Prior studies in prejudice reduction have found that interactional contact between two groups can have positive effects on intergroup attitudes and can improve intergroup relations (Voci, 2003). According to the contact hypothesis (Allport, 1954), contact under certain conditions, such as equal status and institutional support, creates a positive intergroup encounter, which, in turn, brings about an improvement in intergroup relations (Amichai-Hamburger & McKenna, 2006). Besides, true acquaintance can lessen prejudice (Allport, 1954).

Therefore, the training intervention for this study took place as an informal interaction. Based on suggestions of prejudice reduction (Allport, 1954; Sherif, 1966), it consisted of a one hour informal meeting between ITAs and a random sample of 29 raters in this study with the following conditions for prejudice-reducing contact; (1) equal status for members of the different groups ensured; (2) the groups working for common goals; (3) the contact task with cooperative interdependence between the group members; (4) sufficient time given with other stress relieved; (5) a high potential for interpersonal acquaintance between members provided; and (6) participants seen as being typical of their groups. The circumstances were pleasant with refreshments. A selected group of five-six advanced English, gregarious ITAs from the total of

eleven ITAs whose speech was being rated were asked to converse with the raters, trying to collaboratively solve mystery puzzles together and sharing their cultural backgrounds and academic commitments with the undergraduates. Raters were, however, not informed that the groups of ITAs were the same ITAs as they had rated at Phase I.

In order to have each of the 6 criteria above met in the training intervention here, first of all, the equal status aspect was considered given that both ITAs and undergraduate raters were treated as inter-group members who had equal status in the eyes of the researcher. They pursued a common goal, for example, solving a mystery puzzle and exchanging non-verbal communication skills. Moreover, the inter-group contact task, 'solving the mystery puzzle' highly required group cooperation; i.e, group members were required to share several pieces of their information to solve the mystery. As Sherif (1966) argues, realistic group conflict could be derived from in-groups and out-groups pursuing different goals, but this inter-group cooperation, working for the common goal, could lead to more favorable perceptions. In addition, the contact time in the intervention was considered to be sufficient in reducing group members' stress. Each inter-group member was served with refreshments and introduced themselves to other group members before they started the puzzle solving task. Consequently, not only through the activities that each group member was assigned to, but also through the introduction of each member he/she had to make, group members received a great potential for interpersonal acquaintance between members. Finally, participants in the intervention meeting were considered to be reasonably typical groups of ITAs and US undergraduates. Throughout this prejudice reduction intervention, it was hoped that the trained raters could return to the speech rating task with a stronger attention to ITA speech characteristics rather than with an overwhelming negative disposition toward ITA speech.

A group of 5-6 advanced English, gregarious ITAs met as a group with 6-7 undergraduate raters to be "trained" via this intervention. 5 meetings were arranged for this intervenion. Most of the ITAs attended the meetings more than once in which case the activity tasks were altered to avoid the familarity effect. To be precise, at a meeting, one large group of 5-6 ITAs and 6-7 undergraduates were divided into 3 different small groups consisted of 1-2 ITAs and 2-3 undergraduates. After a short introduction of group members, the reseracher outlined the agenda of the meeting and, after that, each group was given a collaborative task, i.e., resolving the mystery puzzle (either *Robbery* or *Murderer*, see Appendice G-1 and G-2), and asked to report answers to the following questions: *who committed the crime?*, *how and why did they do it?*. The small group members were exchanged for the subsequent group activities so that each one of the ITAs could meet all of the undergraduate raters at the meeting. The ITAs and the undergruaduate raters were asked to converse with each other, talking about their cultural backgrounds and of their academic commitment. More specifically, the ITAs and undergraduate raters discussed differences of cultures, non-verbal communication (e.g., hand gestures), or proverbs that turned out to be influential to their lives (see Appendix F). Once individual group discussion were over, participants were asked to briefly report their discussion findings in class. Finally, the reseracher ended the session with appreciation of their participation in the intervention.

Procedures

Initial measures were obtained in face-to-face meetings with varying numbers of participants. After providing informed consent, undergraduate student raters participating in this study completed the 'rater background questionnaire' (See Appendix A). In addition, participants were administered the measure of "*reverse linguistic stereotyping*" . That is, participants heard

two different sections of about 3-minute audio taped lecture of Standard American English. They listened to the passages with two different faces associated with each. In other words, as subjects listened to the college lecture, a slide photograph representing the instructor were projected as a screen, with a photograph of either a Caucasian or an Asian man. Immediately after hearing the lecture, participants were asked to complete a Speech Evaluation Instrument (SEI; Zahn & Hopper, 1985) and a partial cloze test of listening comprehension. The background questionnaire and the measure of lingustic stereotyping were administered several weeks prior to their actual rating sessions for the ITA speech samples so as to avoid contamination or engendering a social desirability mindset for the ratings.

After about 6 weeks had elapsed, participants rated the eleven 4-5 minute ITA speech samples for each of the dependent variables. Participants completed a CSEI, but did not receive any specific training for the use of (analytic) rating scales. All rating data were collected online. Following the initial meeting, participants used computers of their own choosing, so long as they had sound capability and high speech Internet access. They conducted the ratings at times they selected for their own convenience. The online rating portal allowed them to complete the rating over multiple occasions, but most participants reported that they completed the rating task at one sitting. The online mode of administration was utilized in this study not only because it was convenient for participants and efficient for experimenters, but also because it mimics the rating procedure used in high stakes assessments of NNS English speaking proficiency, that is, the iBT TOEFL®.

Once signed on, raters first listened to a practice speech sample. The practice speech sample exposed raters to each of the measurement scales, as well as familiarized them with the operation of the online rating environment. The purpose of the trial sample was to provide raters

with proper expectations at the beginning of the task (Derwing, Rossiter, Munro, & Thomson, 2004).

Raters were instructed to rate each of the speech samples based on a series of 7-point Likert scales with the following subscales incorporated in CSEI: (1) a measure of English language proficiency (2) a measure of instructional competence, (3) a measure of accent standardness, and (4) the SEI—which yielded subscales for *superiority, social attractiveness*, and *dynamism*. After completing all of the rating scales, participants encountered text boxes where they responded to the two open-ended questions.

A week after the 1-hour psycho-social intervention with ITAs, the 70 undergraduate raters (29 trained raters + 41 untrained raters) were intivited to return to the online website to listen to the same 11 samples used in Phase I and complete once again the CSEI for each ITA. All the 29 "trained" raters completed the second ratings, but only 34 out of 41 provided responses for the second ratings. Therefore, the total number used for the analysis of training effects was 63 raters.

Research Design and Analysis

The quantatative reserch design of the study revolved around two phases of data collection. In Phase I, raters with no training intervention rated 11 speaking performances. In Phase II, a subset of raters (29) were given social-psychological rater intervention (training), in which undergraduates interacted for an hour informally with ITAs. They rated the same 11 speech samples. In order to best understand the outcome of the results from the data collected, the researcher employed several statistical analyses: two sets of regressions for rater characteristics vs acoustic characterisitcs, hierarchical cluster analysis (HCA), Rasch multi-facet

analysis, and two-way repeated measures of ANOVA (time x training). Finally, an alternative analysis method, mixed random coefficent modeling (MRCM; Littell, Milliken, Stroup, Walfinger, and Schabenberger, 2007), was performed in an effort to intergrate two sets of regressions; that is, to integrate the rater characteristic predictors and the suprasegmental predictor regressions into a single coherent model. The Statistical Package for Social Science (SPSS), the SAS statistical analysis software package, and FACET (Linacre, 2005) were utilized for this statistical analysis.

The data sets were doubly repeated measure collections, since all 70 raters rated the same 11 speech samples, and this rating procedure took place once in both Phase I and again in Phase II.

Multiple Linear Regression Analyses

The primary analysis used in the study was multiple linear regression, which is a statistical procedure in which scores on one or more variables (i.e., independent variables) are used to predict scores on another variable (i.e., dependent variable). All regressions run in this study were "classical" regressions rather than hierarchical or step-wise. That is, all predictors were entered in a single step in order to estimate the unique contributions of each to the various dependent variables.

In order to examine the effects of the two different sets of predictor variables, the analyses required separate multiple regression runs for (1) rater background characteristics as predictors and (2) speech acoustic indicators as predictors. In these analyses, the researcher used the scores of 70 raters in Phase I only. Variance due to rater background characteristics was ignored in the regression model of testing for the predictive power of the acoustic indicators. In the regression model of background variables, speaker acoustic characteristics were ignored in

the similar manner. The criterion variables for the former rater background regression were all averaged across the 11 speech samples for each of the 70 raters (i.e., *N*=70). Similarly, the latter acoustic regression had a total sample number of 11 by using the averaged scores across the 70 speech samples.

The equation model for a multiple regression of the acoustic characteristics on the various rater responses takes the form: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$, where Y is the dependent variable, the b's are the regression coefficients for the corresponding x (independent) terms, b_0 is the constant or intercept, and ε is the error term reflected in the residuals. The regressions of rater background variables on the oral performance ratings were based on the following model: $Y = b_0 + b_1T_1 + b_2T_2 + b_3T_3 + b_1T_4 + b_1T_5 + b_1T_6 + b_1T_7 + \varepsilon$, where T's are independent terms with intercept b_0 and the error term ε.

Separate regressions were run for each criterion or reponse variable. The response (dependent) variables included (1) *comprehensibility ratings*, (2) *oral proficiency ratings*, (3) *instructional competition ratings*, (4) *accent standardness ratings*, (5) *rater leniency scores* derived from multi-faceted Rasch modeling, and two reverse measures: (6) *superiority* and (7) s*ocial attractiveness*. It has been argued that interpreting *dynamism* in language and attitude studies is problematic. Whereas *superiority* and *social attractiveness* are clearly scales that measure prestige versus non-prestige social judgments, nonstandard speakers are not infrequently judged more dynamic than standard speakers (Rubin & Williams, in preparation; Williams, 2006). Because of its ambiguous evaluative valence, *dynamism* was eliminated from further analysis.

Explanatory (independent) variables included the following clusters of suprasegmental measures driven from the resutls for Hierarchical Cluster Analysis: (a) the fluency factor, (b) the

irregular boundary, and (c) the hesitation markers. Predictor variables also include the following rater background and dispositional measures: (d) native/non-native English language speaker status (e) composite index of exposure to non-native English speaking friends and acquaintances; (f) linguistic sophistication (formal training in language studies and linguistics); (g) experience in English language teaching and tutoring; (h) reaction to courses tuaght by international instructors; scores on a reverse linguistic stereotyping task; (i) *superriority-RLS*, and (j) *social attractiveness*-RLS ─ each served as predictors in MLM analysis.

Degree of exposure to NNSs was indexed by raters' self-reports of the number of hours spent with NNSs during a typical week. Linguistic sophistication was derived by summing (a) the number of college classes in linguistics, applied linguistics, or TESL methods (b) years of foreign language study. The amount of teaching/tutoring experience was determined by summing raters' teaching/tutoring experience in weeks. The reaction to ITAs' courses was the sum of number of the courses of which grades were considered to be hurt because courses were taught by ITAs

Mixed Random Coeffient Modeling (MRCM)

The regression analyses above have permitted only indirect comparison of the amount of variance accounted for by the two sets of predictors—i.e., speech sample acoustic characteristics vs. rater background characteristics. Using multiple linear regression, it is not possible to test the effects of the two sets of predictors simultaneously. Therefore, alternatively random coefficient modeling (MRCM) was considered for further analysis. The researcher believed that the MRCM approach could provide an account to examine the effects of rater characteristics relative to the effects of objectively measured acoustic properties of accentedness in a conjoined technique.

Therefore the data were alternatively analyzed through a mixed random coefficient model to see conjoint contribution of raters' background and acoustic features of accentedness of English to performance ratings. In the data for this study speakers are crossed with raters, and both speakers and raters have attributes that are presumed to influence ratings. Therefore it is desirable to examine the behavior of each outcome as a function of both speaker and rater predictors. As a result, MRCM was conducted by using PROC MIXED (Littell et al., 2007) which is the mixed model analysis component of the SAS statistics system. The purpose of MRCM is to be able to make better predictions, as well as to describe accurately the relationship that is present, by introducing additional parameters, the variance and covariance components across variables. These random coefficient models are compromises between modeling each context separately with its own model and modeling all contexts simultaneous with the same model (Kreft & De Leeuw, 2006).

By the same token, there exists score homogeneity, because measurements are crossed among raters and speakers, and are also correlated within the same individuals (within the same raters and/or within the same speakers). The intra-class correlation is a measure of the degree of dependence of individuals (Kreft & De Leeuw, 2006). These dependencies violate the assumption of independent observations in the traditional linear regression model. The effect of this violation is to increase in the probability of a Type I error and the existence of intra-class correlation makes the test of significance in traditional linear models too liberal (Barcikowski, 1981). The mixed random coefficient analysis solves some of these problems of dependencies among observations.

In ordinary regression models the parameter estimates that specify the regression line are intercept and slopes. Traditionally these coefficients are assumed to be fixed, and the values

are estimated from the data. In a mixed modeling framework, the coefficients of the first-level regression model are treated as random as well (Littell et al., 2007). To specify the model to be fit, it is necessary to write the model out in order to interpret each of the parameters. Especially due to the complexity of the data structure of this study, having several predictors at each level, it was not obvious how to parameterize the model so that the output could be used directly to answer the research questions. The model equations and notations followed Singer's guidelines (1998). Separate models were written for each of the two levels and then they were combined together to yield the single level representation required for PROC MIXED.

As to the equation model, the following approach expresses the outcome $Y_{ij,}$ using a pair of linked models: one at the speaker level and another at the rater-level. At the speaker level, a speaker's outcome as the sum of an intercept for the rater (β_{0j}) and a random error (r_{ij}) associated with the i^{th} speaker in the j^{th} rater. At the rater level, the rater intercepts were expressed as the sum of an overall mean (γ_{i0}) and a series of random deviation from the mean (u_{ij}).

In this formulation, there are **seven** background variables and **three** acoustic features included.

X_i = speaker, i=1, 2, 3

T_j = rater, j= 1, 2, 3, ….7

Speaker-related equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \beta_{3j}X_{3j} + r_{ij}$$

where $r_{ij} \sim N(0, \sigma^2)$.

Rater-related equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_{1i} + \gamma_{02}T_{2i} + \gamma_{03}T_{3i} + \gamma_{04}T_{4i} + \gamma_{05}T_{5i} + \gamma_{06}T_{6i} + \gamma_{07}T_{7i} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_{1i} + \gamma_{12}T_{2i} + \gamma_{13}T_{3i} + \gamma_{14}T_{4i} + \gamma_{15}T_{5i} + \gamma_{16}T_{6i} + \gamma_{17}T_{7i} + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}T_{1i} + \gamma_{22}T_{2i} + \gamma_{23}T_{3i} + \gamma_{24}T_{4i} + \gamma_{25}T_{5i} + \gamma_{26}T_{6i} + \gamma_{27}T_{7i} + u_{2j}$$
$$\beta_{3j} = \gamma_{30} + \gamma_{31}T_{1i} + \gamma_{32}T_{2i} + \gamma_{33}T_{3i} + \gamma_{34}T_{4i} + \gamma_{35}T_{5i} + \gamma_{36}T_{6i} + \gamma_{37}T_{7i} + u_{3j}$$
$$\beta_{4j} = \gamma_{40} + \gamma_{41}T_{1i} + \gamma_{42}T_{2i} + \gamma_{43}T_{3i} + \gamma_{44}T_{4i} + \gamma_{45}T_{5i} + \gamma_{46}T_{6i} + \gamma_{47}T_{7i} + u_{4j}$$
$$\beta_{5j} = \gamma_{50} + \gamma_{51}T_{1i} + \gamma_{52}T_{2i} + \gamma_{53}T_{3i} + \gamma_{54}T_{4i} + \gamma_{55}T_{5i} + \gamma_{56}T_{6i} + \gamma_{67}T_{7i} + u_{5j}$$

Where

$$u = \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} \sim N(0, G), \text{ and } G = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} \\ & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & & \sigma_2^2 & \sigma_{23} \\ & \textit{symmetric} & & \sigma_3^2 \end{pmatrix}$$

Substituting the rater-related equations into the speaker-related equation yields:

Composite Model: $Y_{ij} = \gamma_{00} + \gamma_{01}T_{1i} + \gamma_{02}T_{2i} + \gamma_{03}T_{3i} + \gamma_{04}T_{4i} + \gamma_{05}T_{5i} + \gamma_{06}T_{6i} + \gamma_{07}T_{7i} + \gamma_{10}X_{1j} + \gamma_{11}T_{1i}X_{1j} + \gamma_{12}T_{2i}X_{1j} + \gamma_{13}T_{3i}X_{1j} + \gamma_{14}T_{4i}X_{1j} + \gamma_{15}T_{5i}X_{1j} + \gamma_{16}T_{6i}X_{1j} + \gamma_{17}T_{7i}X_{1j} + \gamma_{20}X_{2j} + \gamma_{21}T_{1i}X_{2j} + \gamma_{22}T_{2i}X_{2j} + \gamma_{23}T_{3i}X_{2j} + \gamma_{24}T_{4i}X_{2j} + \gamma_{25}T_{5i}X_{2j} + \gamma_{26}T_{6i}X_{2j} + \gamma_{27}T_{7i}X_{2j} + \gamma_{30}X_{3j} + \gamma_{31}T_{1i}X_{3j} + \gamma_{32}T_{2i}X_{3j} + \gamma_{33}T_{3i}X_{3jn} + \gamma_{34}T_{4i}X_{3j} + \gamma_{35}T_{5i}X_{3j} + \gamma_{36}T_{6i}X_{3j} + \gamma_{37}T_{7i}X_{3j} + u_{0j} + u_{1j}X_1 + u_{2j}X_{2j} + u_{3j}X_{3j} + r_{ij}$.

Where $r_{ij} \sim N(0, \sigma^2)$ and $u \sim N(0, G)$

Repeated Measures ANOVAs

The researcher conducted 2 (trained group or nontrained group x 2 (Phase I or Phase II) mixed factorial ANOVAs with the second of these factors constituting a repeated measure. A separate such ANOVA was run for each of the seven dependent variables. The purpose of these ANOVAs was to ascertain whether raters showed difference between before and after the socio-

psychological interventions with ITAs, and to examine to what extent a course of training affects rating of L2 oral performance.

Open-ended Online Interview Responses

The researcher collected the responses of the online interviews and open ended questionnaires from the 18 participants (7 trained raters, 9 untrained raters, 2 trained raters for the online interview). The online conversations from two participants were printed out. The interview data about raters' responses to their rating process for the evaluation of the ITA oral presentation speech samples have been used to explicate the quantitative results to uncover additional variables pertinent to the rating process (Rao & Woolcock, 2003). In other words, the interview data has been only utilized to add depth and to comment upon the psychological processes inferred from the main quantitative findings.

CHAPTER FOUR: RESULTS

Introduction

Variance due to rater background and disposition counts as measurement error, because second language learners' English language oral proficiency is often judged by human raters, who vary in their assessments due to these factors that are irrelevant to the learners' language performance. Recently language assessment researchers have probed the nature of language proficiency to better understand its complexity as well as to develop more useful measures of it. Accordingly, this study has compared the impact of selected linguistic and nonlinguistic factors that may affect ratings of speech samples of NNSs' oral performances. More specifically, it explored attitudinal and acoustically measured suprasegmental factors (rate, pause, stress, and intonation) in the accented speech of international teaching assistants (ITAs). It further sought to determine the degree to which rater training (a brief socio-psychological intervention function) could reduce the impact of rater characteristics on speech ratings.

As introduced in Chapter I, the general research questions of this study are as follow:

1. What is the relative impact of rater background characteristics on ratings of L2 oral performance?
2. What is the impact of objectively measured suprasegmental characteristics of accented English on ratings of L2 oral performance?
3. To what extent does a course of training (a brief socio-psychological intervention function) affect ratings of L2 oral performance?

This chapter presents the results of the data analysis performed in this study. The first section presents the effects of rater attitudes and background variables on L2 oral performances

ratings, including examination of difference between NS and NNS raters in oral proficiency rating.

The second section of this chapter reports the effects of acoustically measured suprasegmental factors on L2 oral performance ratings, and it also provides details about the data reduction process (hierarchical cluster analysis) by which the large number of suprasegmental measurements were melded into a small number of multivariate clusters. A comparison between international TA and US TA acoustic features is also presented.

Third, an alternative integrative analysis is presented using a mixed random coefficient model that permitted examination of the simultaneous effects of rater background and dispositional variables on the one hand and of speech sample suprasegmental characteristics on the other.

The fourth section of this chapter provides analysis of training effects computed by a series of 2 (time) x 2 (training) repeated measures ANOVA.

Finally, interview data are presented to support the quantitative results of the present study.

Effects of Rater Attitudes and Background Variables on L2 Performance Ratings

The first set of research questions pertained to the impact of rater attitudinal and background factors on their ratings of NNSs' oral performances. Although a number of additional variables were collected on the rater background questionnaire, the analysis of the present study only include the following seven predictor variables: (1) English native language status (native or nonnative speaker of English), (2) exposure to NNSs, (3) linguistic sophistication, (4) amount of ESL/EFL teaching/tutoring experience, (5) prior negative

experiences in courses taught by ITAs, (6) and (7) two indices of reverse linguistic stereotyping. As foreshadowed in Chapter 2, these were the rater traits that have the strongest precedent in previous research literature as predictors of rating biases.

Regression of Rater Background and Attitudinal Factors on Ratings of NNSs' Oral Performance

To answer the first research question, "what is the relative impact of rater background characteristics on ratings of L2 oral performance?", a multiple regression was performed. The seven rater background and attitudinal factors—(1) native speaker status, (2) time spent with NNSs, (3) linguistic sophistication, (4) teaching experience, (5) reactions to ITA courses (i.e., grades hurt by ITAs), (6) superiority-RLS, and (7) social attractiveness-RLS—each served as predictors in separate regressions for each of six dependent variables.

The seven dependent variables that served as criterion variables for the regression of the rater background characteristics were ratings of (1) oral proficiency, (2) instructional competence, (3) comprehensibility, (4) accent standardness, (5) superiority-RLS, (6) social attractiveness-RLS and (7) rater leniency/ speaker ability scores as derived from Rasch modeling of oral proficiency scores. Each of these dependent variables (except for rater leniency, which was a single score assigned to each rater and therefore did not require averaging) was averaged across the 11 speech performances for each of the 70 raters (i.e., N=70). Each of the seven resulting dependent variables was found to be normally distributed with skewness[6] close to zero (-.19 of minimum and .86 of maximum of distribution statistics in all seven cases) and kurtosis[7]

[6] Skewness is a measure of symmetry, or more precisely, the lack of symmetry. The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right (from *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 04.07/08).

[7] Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. That is, data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with

around 3 (2.26 of minimum and 3.96 of maximum of kurtosis distribution statistics in all seven cases). The normality tests was also conducted through the Kolmogorov-Smirnov test, which is the principle goodness of fit test for normal and uniform data sets (Stuart, Ord, & Arnold, 1999). The results of the test showed that the differences between the distribution of the data set and a normal one were not significant ($p > .05$ in all seven cases), and therefore the null hypotheses (H_0) were not rejected.

Each of the seven multiple regression models was a "classic" regression in the sense that all independent variables were entered as a block at one time. In that way, the results show the unique contribution of each of the predictors, controlling for the simultaneous contributions of all the other predictors. According to Stockburger (1998), it is safest to use this simultaneous method, when there are relatively low numbers of cases.

The following indices are important in interpreting the output of multiple regression analyses. Associated with multiple regression is R^2, the square of the multiple correlation, which indicates the percent of variance in the dependent variable explained collectively by all of the independent variables. The β's are the regression coefficients, representing the amount the dependent variable y changes when the corresponding independent changes 1 unit. The constant, where the regression line intercepts the y axis, represents the value the dependent y will take when all the independent variables are 0. The standardized version of the β coefficients are the beta weights, and the ratio of the beta coefficients is the ratio of the relative predictive power of the independent variables.

---

low kurtosis tend to have a flat top near the mean rather than a sharp peak. The kurtosis for a standard normal distribution is three. Positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution (from *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, retrieved on 04.07/08).

Part correlations, known as squared semi-partial correlation coefficients, were used to assess each variable' unique contribution. They represent the amount by which R^2 is reduced if a particular independent variable is removed from the regression equation. That is, they express the unique contribution of the independent variable as a proportion of the total variance of the dependent variable (Cohen, 1988). Partial correlation coefficients express the unique contribution of the independent variable as a proportion of R^2. In this study, the researcher used R^2 directly as effect-size estimates (Cohen, 1988).

Table 4.1 shows the zero-order correlations among the seven background and attitudinal predictor variables. The results support the fact that the seven predictor variables are relatively independent of one another; the collinear relationships among them are not very strong. The strongest correlation among these seven variables is between the NNS status variable and exposure to NNS variable with a correlation coefficient of .35 (low to moderate).

Table 4.1
Correlations Among Rater Background and Attitudinal Variables

	Linguistic Sophistication	Teaching experience	Time spent with NNS	Grade hurt by ITAs	Superiority -RLS	Social attractive-ness -RLS
NNS status	.29(*)	.03	.35(**)	-.06	.14	.16
Linguistic sophistication		.25(*)	-.01	-.08	.18	.30 (*)
Teaching experience			.07	.08	-.07	.23
Time spent with NNS				-.16	.15	.24
Reaction to ITA course					-.06	-.17
Superiority-RLS						.17

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Rater Background and Attitudinal Factors as Predictors of Oral Proficiency Ratings

The regression model of the seven background and attitudinal factors against ratings of oral proficiency was statistically significant ($F_{7, 59} = 3.40$, $p < .005$). Approximately 30% of the variances in oral proficiency ratings was explained by the 7 predictor variables selected for this model ($R^2 = .32$ and adjusted $R^2 = .23$). Table 4.2 shows the regression and correlation coefficients for the seven background and attitudinal predictors.

Table 4.2
Multiple Regression of Rater Background and Attitudinal Factors on Oral Proficiency Ratings

	Unstandard--ized Coefficients β	Standarized Coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial Correlation	Part Correlation
(Constant)	42.27		15.42	.000			
NNS status	-5.08	-.32	-2.78	.008	-.30	-.35	-.32
Linguistic sophistication	-.05	-.04	-.33	.743	.04	-.05	-.04
Teaching experience	.30	.33	2.64	.011	.29	.35	.31
Time spent with NNS	-.10	-.27	-2.01	.051	-.24	-.27	-.23
Grade hurt by ITAs	-3.36	-.34	-2.79	.007	-.25	-.37	-.32
Superiority-RLS	.07	.06	.45	.654	.02	.06	.05
Social attractiveness -RLS	.09	.08	.60	.551	.13	.08	.07

Note. Model $R^2 = .32$, F (7, 59) = 3.40, $p = .005$; Adjusted $R^2 = .23$

Three of the seven predictor variables (viz., NNS status, language teaching experience, and grades hurt by ITA) contributed significantly to the prediction of variances in the oral proficiency ratings. The variable time spent with NNSs just barely missed the critical value associated with alpha < .05 (i.e., t (59) = -2.01, $p = .051$) with a part correlation of -.23. NNS status revealed a negative regression coefficient (*beta* = -.32). It means that the NNSs (the NS coded as 0 and the NNS coded as 1) were more stringent than the NSs in ratings of oral

proficiency. Oral performance ratings were shown to be inversely proportional to the number of self-reported occasions when raters' grades were hurt by ITA instruction (*beta* = -.34). In addition, raters' experience in teaching languages did predict oral proficiency ratings (part correlation about 30%). The amount of raters' teaching/tutoring experience in languages was associated with leniency in terms of evaluating NNSs' oral proficiency (*beta* = -.33). None of the other rater trait variables exerted statistically significant effects on this dimension of ITA evaluation.

Rater Background and Attitudinal Factors as Predictors of L2 Instructional Competence Ratings

The regression model of the seven background and attitudinal factors on ratings of ITAs' instructional competence was statistically significant ($F_{7, 59} = 3.60$, $p < .005$). About 30% of the variances in the ratings of ITAs' instructional capability was explained by the 7 predictor variables selected for this model ($R^2 = .34$ and adjusted $R^2 = .24$). Table 4.3 below presents the regression coefficients and associated correlations for the seven background and attitudinal predictors.

Table 4.3
Multiple Regression of Rater Background and Attitudinal Factors on Instructional Competence Ratings

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	40.37		16.82	.000			
NNS status	-.15	-.01	-.08	.937	-.10	-.01	-.01
Linguistic sophistication	-.16	-.16	-1.21	.233	.02	-.16	-.14
Teaching experience	.42	.52	4.19	.000	.46	.51	.48
Time spent with NNS	-.07	-.20	-1.51	.139	-.08	-.21	-.17
Grade hurt by ITAs	-2.61	-.30	-2.47	.017	-.21	-.33	-.28
Superiority-RLS	.04	.04	.32	.752	-.02	.05	.04
Social attractiveness -RLS	.09	.09	.68	.503	.18	.09	.08

Note. Model R^2 = .34, F (7, 59) = 3.60, p =.003; Adjusted R^2=.24

Two of the seven predictor variables namely, raters' teaching/tutoring experience and raters' negative reactions to previous courses taught by ITAs – significantly affected judgments of ITAs' instructional competence. Rater's teaching experience was strongly and positively related to the ratings of ITAs' instructional ability (β =.52, t (59) = 4.19, p < .001) with a part correlation of .48. This result indicates that the more teaching experience the raters have, the more positive are their ratings of ITAs' teaching performance. Expectedly, raters' negative past experience in the ITA's courses negatively affected their judgment of ITAs' instruction (β =-.30, t (59)=-2.47, p < .05) with part correlation of -.28. None of the other rater trait variables exerted statistically significant effects on this dimension of ITA evaluation.

Rater Background and Attitudinal Factors as Predictors of Comprehensibility Ratings

The regression model of the seven background and attitudinal factors on the ratings of comprehensibility in NNS's oral performances was statistically significant ($F_{7, 59} = 2.44$, $p < .05$). Less than 25% of the variances in L2 comprehensibility ratings was explained by the 7 predictor variables selected for this model ($R^2 = .25$ and adjusted $R^2 = .15$). The regression coefficients and associated correlations for the seven background and attitudinal predictors are presented in Table 4.4 below.

Table 4.4
Multiple Regression of Rater Background and Attitudinal Factors on Comprehensibility Ratings

	Unstandard-ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	21.36		13.30	.000			
NNS status	-.34	-.04	-.27	.786	-.09	-.04	-.03
Linguistic sophistication	-.04	-.07	-.46	.645	.09	-.07	-.06
Teaching experience	.20	.39	3.00	.004	.34	.39	.37
Time spent with NNS	-.03	-.17	-1.18	.244	-.06	-.16	-.14
Grade hurt by ITAs	-1.92	-.35	-2.72	.009	-.29	-.36	-.33
Superiority-RLS	.06	.09	.71	.483	.07	.10	.09
Social attractiveness -RLS	.03	.04	.30	.769	.16	.04	.04

Note. Model $R^2 = .25$, F (7, 59) = 2.44, $p = .031$; Adjusted $R^2 = .15$

The regression results for the comprehensibility ratings are similar to those of the judgments of ITAs' instructional competence. Both raters' teaching experience and negative reactions to previous classes with ITAs were statistically significant predictors of comprehensibility ratings. The standardized regression coefficients, the beta weights, were .39 (*p*

<.01) and -.35 (p <.01) respectively. The corresponding part correlations of these two variables were .37 and -.33. It means that listener language teaching experience was directly proportional to perceived speaker comprehensibility. On the other hand, the comprehensibility scores were inversely proportional to listeners' self- reported experience of poor instruction in courses taught by ITAs. None of the other rater trait variables exerted statistically significant effects on this dimension of ITA evaluation.

Rater Background and Attitudinal Factors as Predictors of Accent Standardness Ratings

The seven background and attitudinal factors were regressed on the ratings of "foreign" accentedness in ITAs' speech, and the regression model was statistically significant ($F_{7, 59} = 3.33$, p <.01; $R^2 = 32$ and adjusted R^2=.22). The 7 predictor variables selected for this model collectively explained about 20-30% of variance in accent standardness ratings. Table 4.5 illustrates the regression coefficients and associated correlations for the seven background and attitudinal predictors.

Table 4.5
Multiple Regression of Rater Background and Attitudinal Factors on Accent Standardness Ratings

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	13.00		12.91	.000			
NNS status	.16	.03	.21	.837	-.05	.03	.02
Linguistic sophistication	-.05	-.12	-.87	.389	-.02	-.12	-.10
Teaching experience	.16	.49	3.89	.000	.42	.48	.46
Time spent with NNS	-.02	-.15	-1.10	.279	-.08	-.15	-.13
Grade hurt by ITAs	-1.11	-.31	-2.50	.016	-.20	-.33	-.29
Superiority-RLS	-.10	-.27	-2.16	.035	-.25	-.28	-.27
Social attractiveness -RLS	-.02	-.04	-.32	.754	.05	-.04	-.04

Note. Model R^2 = .32, F (7, 59) =3.33, p =.006; Adjusted R^2=.22

Rater teaching experience as well as negative reaction to ITA course likewise came out to be significant predictors of perceptions of NNSs' accent standardness (β= .49, t (59) = 3.89, p ,<.000 and β = -.31, t (59) = -2.50, p < .05 respectively) with part correlations of .46 and -.29. A high score on the accent standardness ratings means that the speech was perceived as closer to the US standard, and a low score means that it diverged from the US standard accent. Inspection of beta weights indicate that teaching experience was positively related to perceived accent standardness of the ITA being rated in this instance, while negative learning experiences in ITA classes was inversely related to perceived standardness ratings of the target ITA. In addition to those two predictors, RLS-*superiority* showed a statistically significant contribution to variance in NNS standardness ratings (β=.-.27, t (59) = -2.16, p <.05) with a part correlation of -.27. The negative coefficient indicates that the more negatively raters tended to stereotype NNSs on this

dimension, the less tolerant they became in listening to the NNSs' accented speech. To be descriptive, if naïve raters manifested linguistic stereotyping of NNSs by judging the Euro-American guise superior to the East-Asian guise (e.g., the Euro-American guise, more intelligent, more educated, more fluent, more literate, more organized, etc, than the East-Asian guise), they also were likely to find NNSs speech more accented and foreign-sounding in scoring of oral performances. None of the other rater trait variables exerted statistically significant effects on this dimension of ITA evaluation.

Rater Background and Attitudinal Factors as Predictors of Superiority Ratings

The regression model of the seven background and attitudinal factors against the ratings of superiority (i.e., the superiority SEI subscale, which reflects perceived intellectual/social/speaking competence and status) was statistically significant ($F_{7, 59} = 3.62$, $p < .005$). Approximately 30% of the variances in superiority ratings was collectively explained by the 7 predictor variables selected for this model ($R^2 = .34$ and adjusted $R^2 = .24$). The regression and correlation coefficients for the seven background and attitudinal predictors are provided in Table 4.6.

Table 4.6
Multiple Regression of Rater Background and Attitudinal Factors on Superiority Ratings

	Unstandard-ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	61.49		21.97	.000			
NNS status	-4.79	-.30	-2.47	.017	-.33	-.32	-.28
Linguistic sophistication	-.44	-.27	-2.24	.029	-.15	-.29	-.26
Teaching experience	.43	.46	3.67	.001	.38	.46	.42
Time spent with NNS	-.08	-.26	-1.69	.097	-.22	-.24	-.21
Grade hurt by ITAs	-2.24	-.22	-1.82	.075	-.11	-.25	-.21
Superiority-RLS	-.07	-.05	-.44	.662	-.04	-.06	-.05
Social attractiveness-RLS	-.07	-.05	-.42	.676	-.08	-.06	-.05

Note. Model R^2 = .34, F (7, 59) =3.62, p =.003; Adjusted R^2=.24

The results of the regression analysis reveal that dummy-coded native speaker status was moderately related to the superiority judgments. The negative coefficient (β= -.30, t (59) = -2.47, p <.05) with a part correlation of -.28 indicates that NSs were associated with higher scores in judged superiority, whereas NNSs tended to give lower superiority scores. Teaching experience (β =.46, t (59) = 3.67, p <.01) exerted a statistically significant impact on superiority ratings with the observed beta weight indicating a directly proportional relation between the two variables. In addition, the linguistic sophistication factor (number of Linguistics courses taken + years of foreign language studied) showed a negative relation with perceptions of speaker superiority. The standardized regression coefficient, the beta weight, of this variable was -.27 (p <.05) with a part correlation of -.26. It indicates that the more language/linguistic background raters have acquired by taking formal linguistic courses or learning foreign languages, the more

negatively they judged NNSs' superiority. None of the other rater trait variables exerted statistically significant effects on this dimension of ITA evaluation.

Rater Background and Attitudinal Factors as Predictors of Social Attractiveness Ratings

The regression model of the seven background and attitudinal factors on the ratings of ITAs' social attractiveness (i.e., kind, friendly, sweet, warm, considerate, and so forth) was statistically significant ($F_{7, 59} = 3.52$, $p < .005$). About 30% of the variance in evaluations of ITAs' social attractiveness was explained by the 7 predictor variables selected for this model ($R^2 = .33$ and adjusted $R^2 = .24$). Table 4.7 below shows the regression coefficients and associated correlations for the seven background and attitudinal predictors.

Table 4.7
Multiple Regression of Rater Background and Attitudinal Factors on Social Attractiveness Ratings

	Unstandard-ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	52.86		21.27	.000			
NNS status	-.48	-.03	-.25	.805	-.15	-.04	-.03
Linguistic sophistication	-.23	-.22	-1.63	.110	-.11	-.23	-.19
Teaching experience	.46	.55	4.40	.000	.41	.53	.51
Time spent with NNS	-.06	-.17	-1.25	.218	-.12	-.17	-.14
Grade hurt by ITAs	-2.57	-.29	-2.35	.023	-.15	-.32	-.27
Superiority-RLS	-.08	-.07	-.60	.549	-.03	-.09	-.07
Social attractiveness -RLS	-.27	-.25	-2.01	.050	-.09	-.27	-.23

Note. Model $R^2 = .33$, F (7, 59) =3.52, p =.004; Adjusted R^2=.24

Factors of both raters' teaching experience and negative reaction to ITAs' courses were statistically significant predictors of ITAs' social attractiveness. The standardized regression

coefficients for each of these variables were .55 (p <.001) and -.29 (p < .05) with part correlations of .51 and -.27 respectively. This result points out that raters' previous language teaching experience was directly proportional to perceptions of ITA social attractiveness. In contrast, raters who held negative attitudes toward ITAs from their previous experience by having poor grades exerted the opposite assessment. They rated ITAs to be cold, unkind, unlikable, awful, and the like.

Moreover, social attractiveness-RLS showed negative regression coefficients (β=.-.25, p =.05) for rated social attractiveness. That is, the degree to which raters held negative stereotypes of NNSs on this dimension indeed resulted in more negative judgments of ITAs.

Rater Background and Attitudinal Factors as Predictors of Rater Leniency Scores

Rater leniency scores, *Fair-M average scores* derived as outcomes from Rasch analysis, also served as a dependent variable. These leniency scores were obtained from ratings of oral proficiency. The regression model of the seven background and attitudinal factors on scores of rater leniency scores was statistically significant ($F_{7, 57} = 2.37$, p <.05). Less than 30% of the variance in rater leniency scores was collectively explained by the 7 predictor variables selected for this model (R^2 = .31 and adjusted R^2=.18). Table 4.8 presents the regression coefficients and associated correlations for the seven background and attitudinal predictors.

Table 4.8
Multiple Regression of Rater Background and Attitudinal Factors on Rater Leniency

	Unstandard-ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	41.24		14.62	.000			
NNS status	-4.55	-.30	-2.13	.038	-.28	-.29	-.25
Linguistic sophistication	-.03	-.03	-.22	.829	.05	-.03	-.03
Teaching experience	.21	.23	1.77	.083	.20	.24	.22
Time spent with NNS	-.11	-.30	-2.11	.040	-.27	-.29	-.26
Grade hurt by ITAs	-2.69	-.28	-2.17	.035	-.21	-.29	-.27
Superiority-RLS	.07	.06	.46	.646	.03	.07	.06
Social attractiveness -RLS	.12	.11	.77	.445	.12	.11	.09

Note. Model R^2 = .31, F (7, 57) =2.37, p =.024; Adjusted R^2=.18

The results appear to be similar to those in the ratings of oral proficiency ratings shown in Table 4.1 due to the fact that these rater ability scores were derived from the oral proficiency ratings subjected. Factors of NNS status, amount of contact with NNSs, and negative reaction to ITAs' courses turned out to be significant in predicting the rater's leniency-severity ability. None of the other rater trait variables exerted a statistically significant effect on rater leniency. The standardized regression coefficients, the beta weights, of these variables were -.30 (p <.05), -.30 (p < .05), and -.28 (p < .05) respectively. The corresponding part correlations were -.25, -.26, and -.27 in that order. These results indicate that NSs were more lenient than NNSs, and the degree of exposure to NNSs and NNSs' past negative experience in ITAs' courses were inversely proportional to the ratings of ITAs' oral proficiency. None of the other rater trait variables showed statistically significant effects on this dimension of ITA evaluation.

Effects of Acoustic Suprasegmental Measures of Accentedness on L2 Performance Ratings

The second set of research questions that motivated this study sought to estimate the impact of objectively measured acoustic indices of suprasegmental characteristics of accented English on ratings of L2 oral performance. From each of the 11 ITAs' speech samples, an acoustic profile of 12 suprasegmental measures was generated. Those 12 measures included 4 measures of speech rate, 5 measures of pauses, 2 measures of stress, and 1 measure of intonation (i.e., pitch range). After preliminary clustering analysis to reduce the 12 acoustic measures to a more manageable set of predictors for multiple regression analysis—as discussed in Chapter 3—the study considered the following three acoustic measures as predictors of speech evaluations:

Preliminary Analysis of Acoustic Measures of Accentedness

Correlations Among Twelve Suprasegmental Measures

Prior to any variable reduction analysis such as hierarchical cluster analysis (HCA), zero-order correlations among 12 acoustic measures were calculated and the results are shown in Table 4.9. The results indicated that speech rate, articulation rate, phonation time ratio, pace, space, proportion of irregular topic boundary, and overall pitch range were strongly correlated with each other. The remaining temporal measures such as frequency or length of pauses, especially, measures of filled pauses, did not correlated strongly with other suprasegmental variables.

Table 4.9
Correlations Among Twelve Suprasegmental Measures

Variables	Art. rate	Mean length run	Phon. time ratio	# of silent pause	Mean length silent pause	# of filled pause	Mean length filled pause	pace	space	Irregular topic boundary	overall pitch range
Speech rate	.94**	.57	.94**	.72*	-.88**	.09	-.28	.82**	-.93**	-.77**	.82**
Art. rate		.52	.78**	.68*	.52	-.002	-.33	.79**	-.96**	-.79**	.85**
Mean length run			.52	-.16	-.24	-.32	-.73*	.36	-.66*	-.64*	.53
Phonation time ratio				.71*	-.93**	.20	-.15	.81**	-.28**	-.63*	.67*
# of silent pause					-.86**	.38	.29	.68*	-.55	-.39	.54
Mean length pause						-.31	-.06	-.87**	.72*	.48	-.60
# of filled pause							.33	.08	.10	.39	-.36
Mean length filled pause								.03	.45	.55	.56
Pace									-.78**	-.45	.56
Space										.822**	-.85**
Overall pitch range											-.95**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

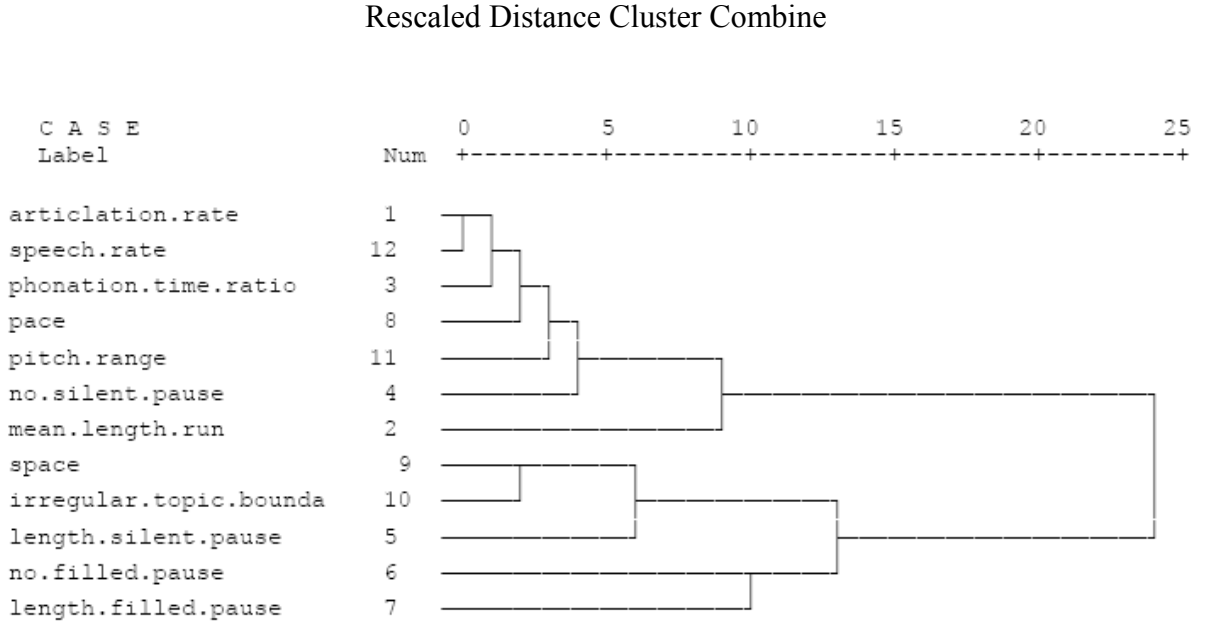
Hierarchical Cluster Analysis of Acoustic Suprasegmental Variables.

To examine the effect of acoustic properties on variance in ratings of ITAs' oral performances, acoustically measured suprasegmental variables had to be entered as predictors for each of the dependent variables. Twelve such measures were obtained via computer-assisted (PRAAT) analysis: four speech rate variables, five pause measures, two stress measures, and one pitch range measure. However, these 12 predictor acoustic variables would prove unreasonable

for the regression analysis because the total *N* for this acoustic regression was only 11. Therefore, hierarchical cluster analysis (HCA) was conducted to reduce the number of acoustic variables by creating a smaller number of justifiable composite variables.

HCA allows us to specify the distance or similarity of the measures to be used in clustering. HCA results are conventionally represented graphically as dendrograms, where each step in the clustering process is shown as a node of the tree. Distance parameters are calculated and then the variables are rank ordered according to those distances. Selecting the number of clusters that best represents the entire data structure is analogous to using a scree criterion for determining the number of factors to retain in a factor analysis. The dendrogram in Figure 4.1 visualizes clusters of acoustic markers.

Figure 4.1
Dendrogram Showing Clusters of 12 Acoustic Suprasegmental Variables Using Average Linkage (Between Groups) Procedure



The resulting clusters of acoustic features were characterized as (1) *acoustic fluency*, which was comprised of (1a) pace [number of prominient sylables per run], (1b) mean length of runs, (1c) phonation time ratio, (1d) articulation rate, (1e) syllables per second, (1f)pitch range, and (1g) number of silent pauses; (2) *irregular boundary*, which was comprised of (2a) space [proportion of prominent words], (2b) mean length of silent pauses, and (2c) proportion of irregular topic boundaries [intraclausal pauses of 0.8 or above]; and (3) *hesitation markers* , which were comprised of (3a) number of filled pauses and (3b) mean length of filled pauses.

More specifically, the *fluency factor* cluster includes all the measures typically considered as indicators of speech rate (e.g., syllable per second, articulation rate, proportion of time in phonation). Moreover, it is further defined by one of the indicators pertaining to prominent stressed syllables, i.e., the number of prominent syllable per run. These stressed syllables are crucial to listeners' perceptions of NNSs' speech (Field, 2005) and likely to interrupt fluent runs if not approprately produced. For example, too much prominent stress can mean low fluency. In addtion, overall pitch range (difference between the F_0 maxima and minima) is a component of this factor. Studies have reported that fluent units of discourse tend to display wide tonal variation whereas low fluency includes narrow and compressed pitch range, which leads to a succession of mostly high or mostly low strings of syllables (e.g., Mennen, 1998; Pickering, 1999; Wennerstrom, 2000). Finally, the number of silent pauses is closely linked to the fluency in the discourse. That is, even though pausing in general detracts from fluency, regular silent pausing at clause boundaries can contribute to fluency. The number of silent pauses is considered as important for recognizing junctures between idea units.

The *irregular boundary* cluster encompasses proportion of prominent word stress and length of silent pauses which includes iregular topic boundary units and mean length of overall silent pauses. Speakers' pauses are closely related to determinations of topical unit boundaries. If a speaker pauses relatively long and frequently, he/she may not stress words at grammatical unit boundaries. In other words, irregular stress can result in unsymetrical intonational and phrasial boundaries.

Hesitation marking is determed by the number and length of the filled pauses. Filled pauses are especially salient disfluencies that most listeners find distracting.

Intercorrelations among three clustered variables were esitmated. Table 4.10 shows the zero-order correlations among the three suprasegmental cluster variables.

Table 4.10
Correlations Among Three Acoustic Suprasegmental Clusters

	Acoustic fluency	Irregular boundary	Hesitation marker
Acoustic fluency	1	.44 (p =.173)	-.15 (p =.661)
Irregular boundary		1	.51 (p =.109)
Hesitation marker			1

None of the three acoustic clusters turned out to be significantly correlated among each other. In other words, those three clusters were reasonably independent predictors of one another .

Independent-Sample T-Test

Prior to the primary multiple regression analyses, the eleven ITAs were compared with the three US TAs by means of independent-sample t-tests (following Komos & Denes, 2004). These *t*-tests determined whether the two group mean scores on each of of the three

suprasegmental variables differed, (provided—as assumed by the *t*-test statistic—that the underlying distributions of the three dependent variables were normal and that the two groups exhibited equal variances).

Despite the small sample sizes, the *t*-test results reported in Table 4.11 reveal statistically significant differences between ITAs and U.S. instructors in two of the three clustered acoustic variables investigated: acoustic fluency and irregular topic boundary. This indicates that U.S. teaching assistants—in comparison to ITAs—were higher in measures of the acoustic fluency factor; i.e., US TAs spoke faster, their silent pauses took up a smaller proportion of their speaking time, they produced longer stretches of discourse between pauses, and they exhibited more variation in intonation with wide pitch ranges. At the same time, the measured scores of the irregular topic boundary factor were lower in US TAs than ITAs. In other words, compared to US TAs, ITAs produced longer and more frequent empty/silent pauses which did not coincide with any semantic breaks or dramatic effects. Also, ITAs did not necessarily assign (lexical and phrasal) stress at grammatical unit boundaries. Thus, acoustic properties such as the acoustic fluency factor (e.g., speech rate, the use of stressed words, or variation in pitch) were characteristic of exemplary USTAs, more so than ITAs.

Table 4.11
t-tests Comparing Speaker Language Status Groups on Three Clusters of Suprasegmentals

Cluster Variable	Group	*N*	Mean	*SD*	*t*	*P*
Acoustic fluency	ITA	11	-2.32	3.01	-6.04	.000
	US TA	3	8.50	.40		
Irregular boundary	ITA	11	.79	2.44	2.53	.027
	US TA	3	-2.90	.69		
Hesitation marker	ITA	11	.40	1.50	1.89	.083

	US TA	3	-1.46	1.54

Furthermore, prior to conducting regression analyses, zero-order correlations among the seven dependent variables and three independent acoustic variables were calculated. These correlations are reported in Table 4.12

Table 4.12
Correlations Among Seven Dependent Variables and Three Suprasegmental Clusters (N=11)

Dependent Variable	Acoustic fluency	Irregular boundary	Hesitation marker
Oral proficiency	.64*	-.49	-.41
Instructional competence	.69*	-.38	-.54
Comprehensibility	.53	-.39	-.44
Accent standardness	.62*	-.50	-.41
Superiority	.65*	-.52	-.36
Social attractiveness	.42	-.25	-.32
Speaker ability	.69*	-.59	-.26

Note. *$p < .05$

Regression of Acoustic Measures of Accentedness on Ratings of L2 Oral Performance

Linear Regressions

The three acoustic suprasegmental factors—acoustic fluency, irregular boundary, and hesitation markers—each served as predictors in separate regressions. Six of the same dependent variables previously used in the regression of the rater background characteristics (i.e., ratings of oral proficiency, instructional competence, comprehensibility, accent standardness, superiority, and social attractiveness) were all averaged across the 70 raters for each of the 11 speaker performances (i.e., *N*=11). The sample size was rather small for the

multiple regression analysis. Nevertheless because the scores were averaged scores across 70 raters, they appeared to be fairly normally distributed, with an observed skewness of (-1.14, .31, and .50 of distribution statistics for acoustic fluency, irregular boundary, and hesitation marking respectively) and kurtosis a little less than 3 (i.e., 2.69 for acoustic fluency, -1.93 for irregular boundary, and -2.36 for hesitation marking). Kolmogorov-Smirnov normality test revealed that the differences between the distribution of the data set and a normal one for each of the acoustic variables were not significant ($p > .05$).

Acoustically Measured Suprasegmental Factors as Predictors of Oral Proficiency Ratings

The regression model of the three acoustic factors against oral proficiency ratings was statistically significant ($F_{3, 10} = 7.5$, $p < .05$. Approximately 60-70% of the variance in oral proficiency ratings was collectively explained by the three acoustic variables selected for this model ($R^2 = .76$ and adjusted $R^2 = .66$). The regression and correlation coefficients for the three acoustic predictors are provided in Table 4.13.

Table 4.13
Multiple Regression of Acoustically Measured Suprasegmental Factors on Oral Proficiency Ratings

	Unstandardiz --ed coefficients β	Standarized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	34.22		42.14	.000			
Acoustic fluency	.55	.75	3.18	.015	.50	.77	.59
Irregular boundary	-2.39	-.64	-2.34	.052	-.41	-.66	-.43
Hesitation marker	-.53	-.20	-.83	.436	-.64	-.30	-.15

Note. Model $R^2 = .76$, F (3, 10) =7.46, p =.014; Adjusted R^2=.66

Acoustic fluency (β =.75, t (10) = 3.18, $p < .05$) is positively and strongly related to oral proficiency ratings. The part correlation of this variable is .59. Neither of the other two

predictors achieved statistical significance for this dependent variable (though irregular boundary factor was close).

Acoustically Measured Suprasegmentals as Predictors of Speaker Instructional Competence Ratings

The regression model of the three acoustic factors against instructional competence ratings was statistically significant ($F_{3, 10} = 10.08$, $p < .01$). Over 70% of the variance in oral proficiency ratings was explained collectively by the three acoustic variables selected for this model ($R^2 = .81$ and adjusted $R^2 = .73$). The regression and correlation coefficients for the three acoustic predictors are shown in Table 4.14.

4.14
Multiple Regression of Acoustically Measured Suprasegmental Factors on Instructional Competence Ratings

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	37.38		53.40	.000			
Acoustic fluency	.52	.74	3.55	.009	.54	.80	.58
Irregular boundary	-.98	-.36	-1.94	.094	-.35	-.59	-.25
Hesitation marker	-.97	-.52	-3.06	.018	-.80	-.76	-.39

Note. Model $R^2 = .81$, F (3, 10) =10.08, p =.006; Adjusted R^2=.73

Acoustic fluency shows a strong and positive relation with ratings of ITAs' instructional competence (β =.74, t (10) = 3.55, $p < .01$) with a part correlation of .58. Hesitation marking was inversely associated with the scores. The negative coefficient , β = -.52, for hesitation marking indicates that raters found ITAs' frequent use of filled pauses to be a sign of incompetent of teaching.

Acoustically Measured Suprasegmental Factors as Predictors of Speaker Comprehensibility Ratings

The regression model of the three acoustic factors against comprehensibility ratings was not statistically significant, given the low sample size ($F_{3, 10} = 3.91$, $p > .05$). However, R^2 for this model was .63 and adjusted R^2 was .47.

Acoustically Measured Suprasegmental Factors as Predictors of Speaker Accent Standardness Ratings

The regression model of the three acoustic factors against accent standardness ratings was statistically significant ($F_{3, 10} = 7.43$, $p < .05$). Approximately 70% of the variance in oral proficiency ratings was collectively explained by the three acoustic variables selected for this model ($R^2 = .76$ and adjusted $R^2 = .66$). The regression and correlation coefficients for the three acoustic predictors are shown in Table 4.15.

Table 4.15
Multiple Regression of Acoustically Measured Suprasegmental Factors on Accent Standardness Ratings

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	12.46		37.68	.000			
Acoustic fluency	.23	.77	3.27	.014	.50	.78	.60
Irregular boundary	-1.02	-.66	-2.44	.045	-.41	-.69	-.45
Hesitation marker	-.18	-.17	-.68	.519	-.62	-.25	-.13

Note. Model $R^2 = .76$, F (3, 10) =7.43, p =.014; Adjusted R^2=66

As the table above indicates, acoustic fluency (β =.77, t (10) = 3.27, $p < .05$) was positively and significantly related to listeners' perceptions of accent standardness, whereas the irregular boundary (β = -.66, t (10) = -2.44, $p < .05$) was inversely related to the accent

standardness judgment. Raters consider long or awkward silent pauses to be foreign and un-standard. Hesitation marking was not associated with the accent standardness ratings.

*Acoustically Measured Suprasegmental Factors as Predictors of Speaker Superiority Rating*s

The regression model of the three acoustic factors against superiority ratings was statistically significant ($F_{3, 10} = 6.52$, $p < .05$). Between 60%-70% of the variance in oral proficiency ratings was collectively explained by the three acoustic variables selected for this model ($R^2 = .74$ and adjusted $R^2 = .62$). The regression and correlation coefficients for the three acoustic predictors are shown in Table 4.16.

Table 4.16
Multiple Regression of Acoustically Measured Suprasegmental Factors on Superiority Ratings

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	53.04		73.77	.000			
Acoustic fluency	.44	.71	2.88	.024	.52	.74	.56
Irregular boundary	-1.70	-.54	-1.88	.103	-.36	-.58	-.36
Hesitation marker	-.60	-.27	-1.05	.327	-.65	-.37	-.20

Note. Model $R^2 = .74$, F (3, 10) =6.52, p =.019; Adjusted R^2=62

Acoustic fluency (β =.71, t (10) = 2.88, $p < .05$) was strongly and positively related to perceived speaker superiority. But neither irregular boundary nor hesitation marking was significantly related to superiority ratings.

Acoustically Measured Suprasegmental Factors as Predictors of Social Attractiveness Ratings

The regression model of the three acoustic factors against social attractiveness ratings was not statistically significant ($F_{3, 10} = 1.66$, $p > .05$). R^2 for this model was .42 and adjusted R^2 was .16.

Acoustically Measured Suprasegmental Factors as Predictors of Speaker Ability Scores

The regression model of the three acoustic factors against speaker ability scores derived from Facet analysis on oral proficiency ratings was statistically significant ($F_{3, 10} = 8.28$, $p < .05$). Approximately 60-70% of the variances in oral proficiency ratings was explained by the three acoustic variables selected for this model ($R^2 = .78$ and adjusted $R^2 = .69$). The regression and correlation coefficients for the three acoustic predictors are shown in Table 4.17.

Table 4.17
Multiple Regression of Acoustically Measured Suprasegmental Factors on Speaker Ability Scores

	Unstandard--ized coefficients β	Standardized coefficients β	*t*-value	sig. *p*	Zero-order correlation	Partial correlation	Part correlation
(Constant)	34.22		33.36	.000			
Acoustic fluency	.66	.69	3.03	.019	.59	.75	.54
Irregular boundary	-1.75	-.35	-1.36	.217	-.26	-.46	-.24
Hesitation marker	-1.41	-.41	-1.73	.127	-.69	-.55	-.31

Note. Model $R^2 = .78$, F (3, 10) =8.28, p =.011; Adjusted R^2=69

Results in this model were the same as the one in the regression of oral proficiency ratings because the speaker ability was calculated based on the oral proficiency scores. Acoustic fluency (β =.66, t (10) = 3.03, $p < .05$) indicated a strong and positive relation with the proficiency rating whereas the other two acoustic variables showed a weak and negative relation.

Alternative Integrative Analysis

The study tried to investigate the *simultaneous* contributions of both rater background characteristics and acoustic features to oral proficiency rating. In other words, the study attempted to test the hypothesis that listeners' background characteristics can influence their

perceptions of foreign speakers' oral proficiency to a degree comparable to the influence on ratings of acoustic properties of NNS speech. The strongest demonstration of this pattern would need to find a way to incorporate rater characteristics into the same statistical model as speaker suprasegmental measures.

Intra-Class Correlation Among Raters

One procedure for examining how much measurement error variance can be attributed to raters calculated the intra-class correlation among raters for each of the six dependent variable (See table 4.18). The intra-class correlation is a measure of the degree of dependence among individuals (Kreft & De Leeuw, 2006). The intra-class correlation is computed from the sums of squares. The analysis in this study was based on Shrout & Fleiss' (1979) suggestion, obtaining the intra-class correlation coefficients through the SPSS *reliability* procedure tool. The analysis employed a two-way random effects model where rater effects were random and measures effects were fixed, by selecting the consistency option in SPSS.

Table 4.18
*Intra-class Correlation Coefficients for Six Dependent Variables (70 raters for ratings of 11 speech sample*s)

Dependent variables	Intra class correlation		
	Single rater	Multiple raters	Coefficient Alpha for inter-coder reliability
Oral proficiency	.28	.96	.96
Instructional competence	.30	.97	.97
Comprehensibility	.28	.96	.96
Accent standardness	.31	.97	.97
Superiority	.25	.96	.96
Social attractiveness	.22	.95	.95

The reliability coefficients for a single judge's rating appear to be very low, ranging from .22 to .31 for all six dependent variables. In contrast, the reliability of the average rating of the 70 judges shows high reliability, with values of .95 and above. Based on the low reliability of a single judge's ratings, we can conclude that considerable measurement error (unreliability) resides in each rater's judgments.

The Mixed Random Coefficient Model

Intra-class correlations are informative about rater error variance in the aggregate, but still don't permit conclusions about the specific sources of rater error, such as differences in amount of contact with NNSs or differences in reverse linguistic stereotyping. And intra-class correlation cannot directly compare error variance due to raters with presumably true score variance due to suprasegmental aspects of accent. Therefore, instead of conducting indirect comparisons of the amount of variance accounted for by the two sets of predictors—as was done earlier in this chapter—a mixed random coefficient model (MRCM) was employed as a tool that is flexible enough to examine the effects of rater characteristics relative to the effects of objectively measured acoustic properties of accentedness in a conjoint technique.

The MRCM incorporated seven background predictors: NNS status, linguistic sophistication, amount of teaching experience, time spent with NNS, negative past experience in ITAs' courses, and the two measures of reverse linguistic stereotyping—superiority and social attractiveness. The very same MRCM also incorporated the three acoustically measured suprasegmental variables: acoustic fluency, irregular boundary, and hesitation markers. Consequently, ten predictive variables were entered for this MRCM analysis. The model applied to this analysis was based on Littell, Milliken, Stroup, Walfinger, and Schabenberger's[8] (2007)

[8] The researcher received consultations of the mixed model analysis directly from authors of the book, *SAS for Mixed Models*. (Littell, et al., 2007).

SAS for Mixed Models and also guided by Singer[9] (1998). SAS PROC MIXED statements was used for the analysis, as suggested by Ramon Littell, the first author of the primary book about MRCM.

In the data structure of this study exists group homogeneity, which increases the intra-class correlation[10]. Besides, measurements are crossed in raters and speakers, and are also correlated within the same individuals. Because there were two crossed factors, rather than one nested in the other, the data structure of this study has "covariates" on both the speakers and raters (the attributes).Therefore, the speakers were considered to be fixed and the raters to be random, which would make the interaction random. PROC MIXED statements are as follow:

```
 proc mixed method=ml covtest data;
class speaker rater;
model response = speaker speaker_attributes rater_attributes/solution ddfm=bw;
random rater rater*speaker;
run;
```

Mixed Model analysis was conducted for each of 6 dependent variables. The rater leniency-severity variable was excluded from this analysis since the model did not converge by having too many likelihood cases. The information about the goodness of fit of multiple models selected for each of the 6 dependent variable analysis was provided in Appendix J.

Effects of Rater and Speaker Predictors on Oral Proficiency Ratings

A random coefficients regression model was used for oral proficiency ratings as the dependent variable with ten predictor variables. Table 4.19 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables. The values in the estimate column show regression coefficients, for example, raters that differ by 1 point in teaching experience differ by .17 (round up) points in oral proficiency

[9] The Mixed Random Coefficient Model used in this study was also confirmed by Judith Singer through personal email correspondences (March 14, 2008).

rating. Its standard error of 0.05 (round up) yields an observed *t*-statistic of 3.15 ($p < .005$), which indicates that there is a statistically significant relationship between raters' teaching experience background and oral proficiency scores—strong and positive relationship.

Table 4.19
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Oral Proficiency Ratings

Solution for Fixed Effects

Effect	Estimate	Standard error	DF	*t*-value	Pr > /t/
Intercept	42.21	1.26	646	33.53	<.001
NNS status	-3.57	0.98	646	-3.66	0.001
Linguistic sophistication	0.002	0.08	646	0.03	0.975
Teaching experience	0.17	0.05	646	3.15	0.002
Time spent with NNS	-0.09	0.02	646	-4.21	<.001
Grade hurt by ITAs	-3.44	0.55	646	-6.29	<.001
Superiority_RLS	0.10	0.07	646	1.44	0.151
Social attractiveness_RLS	0.06	0.06	646	1.05	0.292
Acoustic fluency	0.49	0.09	646	5.63	<.001
Irregular boundary	-1.58	0.51	646	-3.07	0.002
Hesitation marker	-1.17	0.32	646	-3.64	0.001

The results of this mixed random coefficient regression appear to be relatively consistent with those of the linear regression discussed earlier in Tables 4.2 and 4.13 for the prediction of rater's background characteristics and acoustic factors on oral proficiency ratings. Teaching experience shows a strong/positive relationship whereas NNS status and grade hurt by ITA variables indicate a strong/negative relationship. Also, the acoustic fluency factor is significantly and positively related to the proficiency scores. However, due to the efficient operation of error

variance generated from the crossing nature of the raters and speakers, the mixed model makes the test of significance more powerful. Consequently, several other predictor variables turned out to be statistically significant in this model with the increased power. Time spent with NNS exerts a negative relationship. In addition, all the remaining two acoustic variables, irregular boundary and hesitation markers, came out to be statistically significant with negative associations. It indicates that frequent uses of long pauses or filled pauses affect rater's language proficiency judgment negatively. Overall, all the speaker-related predictors showed strong effects in predicting oral proficiency ratings.

Effects of Rater and Speaker Predictors on Instructional Competence Ratings

A random coefficients regression model was run for instructional competence ratings as the dependent variable with ten predictor variables. Table 4.20 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables.

Table 4.20
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Instructional Competence Ratings

Solution for Fixed Effects

Effect	Estimate	Standard error	DF	*t*-value	Pr > /*t*/
Intercept	40.58	1.15	646	35.39	<.001
NNS status	-0.40	0.88	646	-0.46	0.646
Linguistic sophistication	-0.12	0.07	646	-1.76	0.078
Teaching experience	0.25	0.05	646	5.26	<.001
Time spent with NNS	-0.06	0.02	646	-2.86	0.004
Grade hurt by ITAs	-2.58	0.50	646	-5.21	<.001
Superiority_RLS	0.08	0.11	646	0.77	0.441
Social attractiveness_RLS	0.19	0.12	646	1.66	0.097
Acoustic fluency	0.49	0.08	646	6.35	<.001
Irregular boundary	-1.70	0.46	646	-3.69	0.001
Hesitation marker	-1.17	0.29	646	-4.04	<.001

Patterns obtained in this mixed random coefficient regression appear to be moderately consistent with those of the linear regression discussed earlier in Tables 4.3 and 4.14 for the prediction of rater's background characteristics and acoustic factors on instructional competence ratings. Teaching experience, negative past experience in ITA courses, and acoustic fluency are particularly strongly related to the dependent variable, which would be expected results from the earlier linear regression analysis. In addition, time spent with NNS (negative, β= -.06; t (646) =-2.86; p=.004) appears to be statistically significant. The remaining two acoustic factors are now statistically significant in predicting instructional ability scores.

Effects of Rater and Speaker Predictors on Comprehensibility Ratings

A random coefficients regression model was computed for comprehensibility ratings as the dependent variable with ten predictor variables. Table 4.19 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables.

Table 4.21
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Comprehensibility Ratings

Solution for Fixed Effects

Effect	Estimate	Standard error	DF	*t*-value	Pr > /t/
Intercept	19.60	0.88	646	22.29	<.001
NNS status	0.39	0.65	646	0.59	0.555
Linguistic sophistication	-0.03	0.05	646	-0.49	0.624
Teaching experience	0.09	0.04	646	2.54	0.011
Time spent with NNS	-0.04	0.02	646	-2.66	0.008
Grade hurt by ITAs	-1.34	0.39	646	-3.47	0.001
Superiority_RLS	-0.03	0.04	646	-0.58	0.561
Social attractiveness_RLS	0.06	0.07	646	0.84	0.399
Acoustic fluency	0.31	0.06	646	5.20	<.001
Irregular boundary	-1.23	0.35	646	-3.47	0.001
Hesitation marker	-0.73	0.22	646	-3.28	0.001

Patterns obtained in this mixed random coefficient regression appear to be consistent with those of the linear regression discussed earlier in Tables 4.4 for the prediction of rater's background characteristics on comprehensibility ratings. Earlier, a linear regression model of

acoustic predictor variables on comprehensibility ratings was not statistically significant. In addition to the two significant background predictors: teaching experience and negative reaction to ITAs' course variables, acoustic fluency is strongly related to the comprehensibility scores. The directionality of those relationships emerges the same as the earlier linear regression model. However, this mixed model analysis gave other variables statistical significance such as time spent with NNS and remaining two acoustic predictors.

Effects of Rater and Speaker Predictors on Accent Standardness Ratings

A random coefficients regression model was computed for accent standardness ratings as the dependent variable with ten predictor variables. Table 4.22 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables.

Table 4.22
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Accent Standardness Ratings

Solution for Fixed Effects

Effect	Estimate	Standard error	DF	*t*-value	Pr > /t/
Intercept	13.38	0.49	646	27.51	<.001
NNS status	-0.34	0.38	646	-0.90	0.366
Linguistic sophistication	0.01	0.029	646	0.32	0.752
Teaching experience	0.05	0.02	646	2.27	0.023
Time spent with NNS	-0.01	0.01	646	-1.40	0.161
Grade hurt by ITAs	-1.13	0.21	646	-5.34	<.001
Superiority_RLS	-0.08	0.03	646	-2.97	0.003
Social attractiveness_RLS	-0.03	0.02	646	-1.49	0.138
Acoustic fluency	0.22	0.03	646	6.96	<.001
Irregular boundary	-0.89	0.19	646	-4.76	<.001
Hesitation marker	-0.33	0.12	646	-2.82	0.005

Patterns obtained in this mixed random coefficient regression appear to be rather similar to those of the linear regression discussed earlier in Tables 4.5 and 4.15 for the prediction of rater's background characteristics and acoustic factors on perceived accent standardness ratings. Teaching experience, negative reaction to ITAs' course, and superiority-RLS among rater related predictors were strongly related to the accent standardness scores. These three predictors showed strong effects in the regular regression model. The two remaining measures of acoustic predictors turned out to be statistically significant in addition to the acoustic fluency. Directionality of all these predictor variables is identical in both regression models.

Effects of Rater and Speaker Predictors on Superiority Ratings

A random coefficients regression model was computed for superiority ratings as the dependent variable with ten predictor variables. Table 4.23 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables.

Table 4.23
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Superiority Ratings

Solution for Fixed Effects

Effect	Estimate	Standard error	DF	*t*-value	Pr > /t/
Intercept	61.44	1.16	646	52.97	<.001
NNS status	-0.081	0.02	646	-4.04	<.001
Linguistic sophistication	-0.30	0.07	646	-4.29	<.001
Teaching experience	0.27	0.05	646	5.56	<.001
Time spent with NNS	-2.85	0.90	646	-3.17	0.002
Grade hurt by ITAs	2.09	0.50	646	-4.14	<.001
Superiority_RLS	-0.12	0.06	646	-1.85	0.065
Social attractiveness_RLS	-0.14	0.05	646	-2.62	0.009
Acoustic fluency	0.38	0.07	646	5.16	<.001
Irregular boundary	-1.21	0.44	646	-2.72	0.007
Hesitation marker	-1.07	0.28	646	-3.86	0.001

As seen in Table 4.23, most of the predictor variables except for superiority-RLS variable turned out to be strongly related to the dependent variable of superiority ratings. Predictors that show negative relationships with the superiority ratings were NNS status, linguistic

sophistication, time spent with NNS, negative reaction to ITAs' course, two measures of linguistic stereotyping, and two suprasegmental variables (irregular boundary and hesitation markers). On the other hand, teaching experience and acoustic fluency exerted a strong and positive effect on prediction of the dependent variable.

Effects of Rater and Speaker Predictors on Social Attractiveness Ratings

A random coefficients regression model was computed for social attractiveness ratings as the dependent variable with ten predictor variables. Table 4.24 presents parameter estimates for the fixed effects, which shows the relationships between the dependent variable and predictor variables.

Table 4.24
Parameter Estimates of Mixed Models for Rater and Speaker Characteristics as Predictors of Social Attractiveness Ratings

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	*t*-value	Pr > /*t*/
Intercept	54.22	1.06	646	51.18	<.001
NNS status	-0.45	0.82	646	-0.55	0.581
Linguistic sophistication	-0.28	0.06	646	-4.45	<.001
Teaching experience	0.30	0.04	646	6.88	<.001
Time spent with NNS	-0.07	0.02	646	-3.59	<.001
Grade hurt by ITAs	-2.53	0.46	646	-5.50	<.001
Superiority_RLS	-0.06	0.06	646	-0.93	0.352
Social attractiveness_RLS	-0.15	0.05	646	-2.71	0.007
Acoustic fluency	0.28	0.07	646	3.94	<.001
Irregular boundary	-1.04	0.43	646	-2.42	0.016
Hesitation marker	-0.08	0.27	646	-0.30	0.765

The regression model of acoustic factors as predictors of social attractiveness ratings was not statistically significant. As seen Table 4.24, not all the three acoustic factors are strongly related to the dependent variable. The irregular boundary factor as well as the acoustic fluency factor was significant, but hesitation markers did not show any statistical significance. However, this mixed model analysis made two more rater background variables significant compared to the results from the regular regression model. Teaching experience (positive coefficient, β=-.31; t (646) =6.88; p<.001) and grade hurt by ITA (negative, β=-2.53; t (646) =-5.50; p<.001) predictors are still significant. Linguistic sophistication and time spent with NNS variables were added to the list of statistically significant predictors. Directionality of relationships among variables did not change between these two different regression models.

Training effect

The third research question of this study was "To what extent does a course of training (a brief socio-psychological intervention function) affect rating of L2 oral performance?" A gregarious group of five-six ITAs from the total 11 were asked to converse with the raters. As detailed in Chapter 3, this positive cross-cultural contact experience was designed to enable the trained raters to return to the rating tasks with a less biased attentiveness to ITA speech characteristics. One week after this socialization with ITAs, the 70 undergraduate raters (29 trained raters and 41 untrained raters) were invited to listen to the same 11 samples used in Phase I and to complete the same battery of ratings they had previously done. All the 29 trained raters completed the second ratings, but only 34 out of 41 untrained participants provided responses for the second round of ratings. Therefore, the total sample size for the analysis of training effects involved 63 raters. The time lag between Phase I and Phase II was about 6 weeks.

In order to examine differences between oral performance ratings before training and ratings after training, 2 (time of rating) x 2 (training group) mixed factorial ANOVAs, with repeated measures on the first factor, were computed separately for each of the seven dependent variables (viz. ratings of oral proficiency, instructional competence, comprehensibility, accent standardness, superiority, social attractiveness, and rater leniency). Rater leniency was calculated for oral proficiency scores via Rasch modeling, as was the case for Phase I ratings.

Table 4.25 provides cell means for the seven dependent variables broken down by group (trained vs. untrained) and by time (time1 vs. time2).

Table 4.25
Descriptive Statistics of Seven Dependent Variables for Time x Training Status

	Trained (*n*=29)				Untrained (*n*=34)			
Dependent Variables	Time 1 (pre-test)		Time 2 (post-test)		Time 1 (pre-test)		Time 2 (post-test)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Oral proficiency	32.80	5.04	35.35	7.53	35.62	5.92	35.88	6.82
Instructional competence	35.38	4.25	37.68	6.90	38.81	5.57	38.57	6.01
Comprehensibility	18.03	3.19	20.00	3.78	19.69	3.73	20.00	4.26
Accent standardness	11.99	1.96	12.52	2.78	12.45	1.71	12.98	2.65
Superiority	51.42	4.09	53.64	7.76	54.74	5.7	54.75	7.58
Social attractiveness	48.44	4.76	50.50	7.64	50.09	6.04	49.85	6.12
Rater leniency scores	33.67	6.08	35.18	6.57	34.82	6.06	35.50	7.22

Effects of training would be revealed in the interaction between time of testing and training status. If training did exert a meaningful effect, the increases from Phase I to Phase II would be greater for the trained group than for the untrained group. Inspection of Table 4.25,

reveals that trained raters' mean scores apparently increased at post-test for all the seven dependent variables, compared to the pre-test scores. The 2 x 2 repeated measures ANOVAs were run to determine whether that apparent pattern was statistically significant for any of the seven dependent variables.

No statistically significant main or interaction effects were identified in ANOVAs of accent standardness, superiority, social attractiveness, and severity ratings.

Effects of Training on Instructional Competence Ratings

The summary of the ANOVA results for perceived instructional competence is shown in Table 4.26. While no main effect achieved statistical significance, the time x training group interaction effect was statistically significant ($p < .05$).

Table 4.26
Time x training status ANOVA of Instructional Competence Ratings

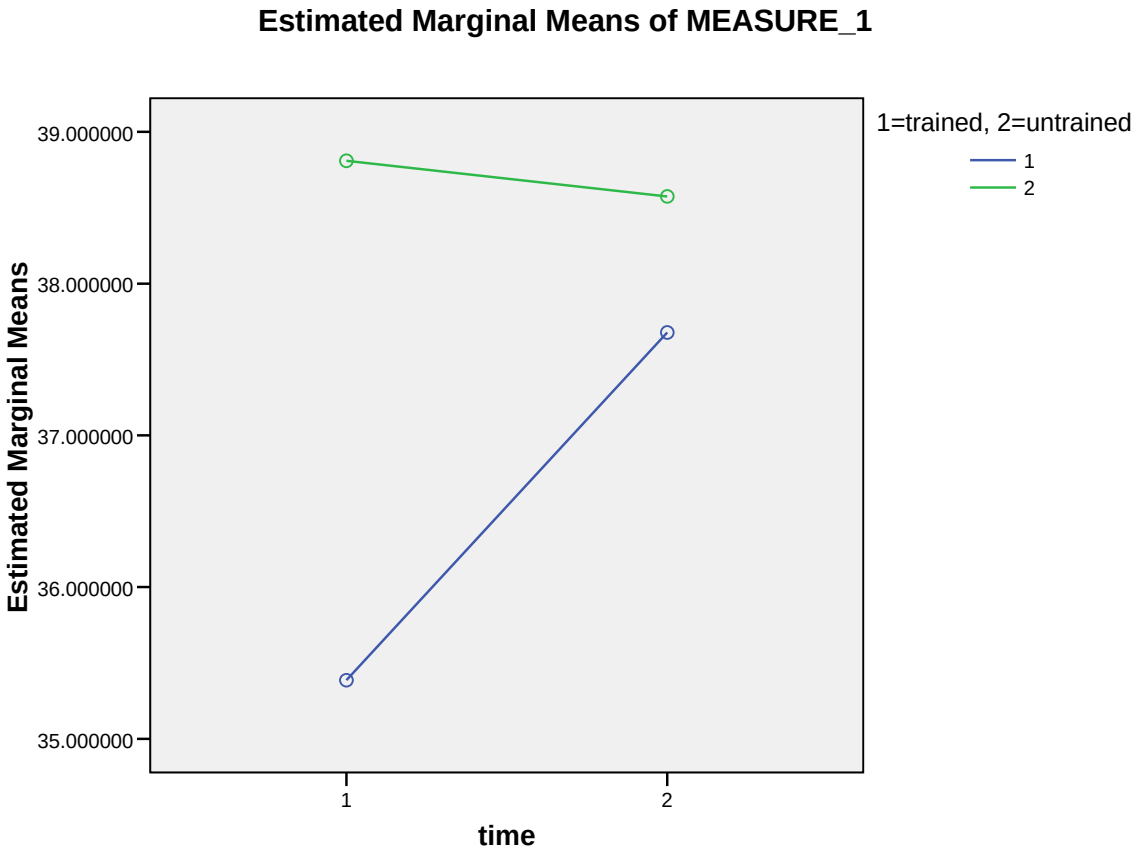
Source	*df*	Mean Square	*F*	*P*	Partial Eta^2
Training	1	145.13	2.67	.107	.04
Error $_{between}$	61	54.28			
Time	1	32.89	2.69	.106	.04
Time x Training	1	49.63	4.05	.049*	.06
Error $_{within}$	61	12.25			

Tukey's HSD procedure was conducted to analyze the interaction between time and training. All pair-wise contrasts were examined. Statistically significant ($p < 0.05$) were the following pair-wise contrasts (see table 4.25 for the mean values): the ITA ratings on this variable of the time2-trained group exceeded their ratings at time 1. Time1-untrained group ratings exceeded the time1-trained group ratings, and the time 2-untrained group even exceeded

the time1-trained group. Although the group that received no intervention started out higher than the untrained group, and in some cases remained higher than the trained group, on this variable the trained group's instructional competence ratings of ITAs rose from Phase I to Phase 2, whereas that was not so for the untrained raters.

The plot in Figure 4.2 illustrates the interaction between training group status and time of rating on ratings of ITAs' instructional competence. As shown in Figure 4.2, the rating scores of the untrained group remains fairly level from Time 1 to Time 2, whereas the trained group mean scores increases drastically from pre-training to post-training.

Figure 4.2
Interaction Between Time of rating and Training status on Instructional Competence Ratings



Effects of Training on Comprehensibility Ratings

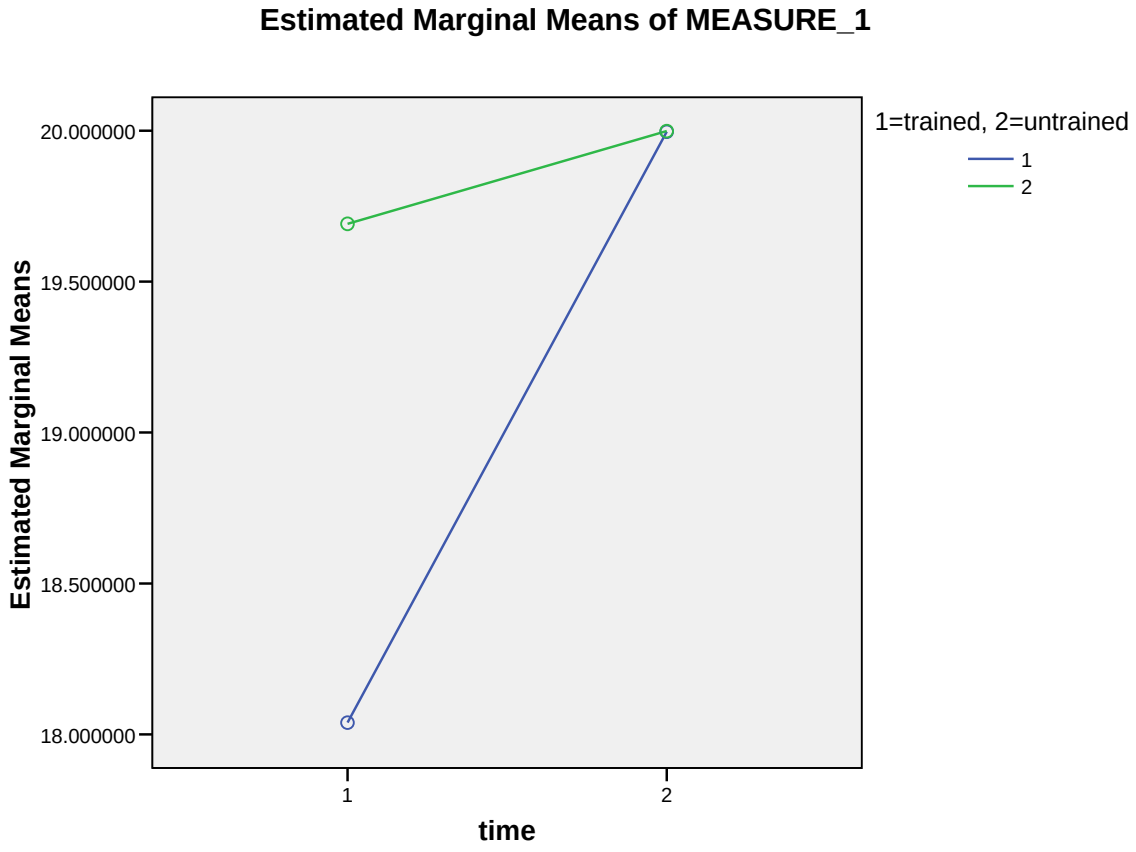
The 2 (training group) x 2 (time of rating) ANOVA for ITA comprehensibility ratings is summarized in Table 4.27. A time main effect emerged such that post test scores (M_{post}=20) exceeded pretest scores (*M*pre=18.86)

Table 4.27
Time x training status ANOVA of Comprehensibility Ratings

Source	*df*	Mean Square	*F*	*P*	Partial Eta^2
Training	1	21.29	.89	.349	.01
Error $_{between}$	61	23.89			
Time	1	39.91	8.16	.006**	.12
Time x Training	1	21.18	4.33	.042*	.06
Error $_{within}$	61	4.89			

Of more importance was the statistically significant interaction between training group status and time of rating. Tukey's HSD procedure was also applied to examine all pair-wise contrasts. The pattern obtained was the same as the case of instructional competence ratings. ITA ratings of the trained group at Phase II exceeded their ratings at Phase I. The mean ratings of the untrained group at Phase I exceeded the mean ratings of the training group at Phase I. And the mean rating of the nontraining group at Phase II exceeded the mean of the training group at Phase I. The important part of this patter is that members of trained group found ITA speech more comprehensible following the psychosocial intervention, whereas comprehensibility ratings of the nontraining group remained relatively steady between Phases I and II. . In addition, difference between the trained and the untrained at time 1 disappeared at time 2. Figure 4.3 illustrates the plot of interaction effect between time of rating and training group status on comprehensibility ratings.

Figure 4.3
Interaction Effect Between Time and Training on Comprehensibility Ratings



Open-ended Questionnaire Items and Interview Results

The primary purpose of the online interview and the open-ended questionnaire items was to better understand the otherwise hidden rationale for rater responses to NNSs' oral performance. This sort of qualitative supplement to quantitative findings can sometimes reveal *the reasons why* participants made the decisions they did. It has the potential to uncover additional variables that could have potentially influenced the rating process (Rao & Woolcock, 2003).

These qualitative data were collected by means of two different approaches: (1) open-ended comments provided online immediately after completing the rating tasks and (2) overall feedback elicited by direct email or instant messenger contact initiated by the researcher. The first approach gathered participants' comments about ITAs' English proficiency and instructional ability right after raters completed online rating. Text boxes at the bottom of each online rating tool were provided for this purpose. The responses to the online interview/ e-mail questionnaire were collected 2-3 weeks after the entire rating procedure had been completed. Comments elicited by both techniques were combined for purposes of analysis. Comments about three broad topics were solicited via these questions: (1) statements about ratings related to raters' own background (2) statements about perceptions of the characteristics of the speech samples, and (3) impressions of the socio-psychological intervention meeting.

To elicit open-ended comments following the rating task, the researcher included the following item in the rating tool: "Please jot down a sentence or two or some words or phrases that describe this teacher's level of oral proficiency in English and your reaction to it. How will this teacher do in a typical undergraduate classroom?" In other words, raters encountered this open-ended question eleven times, after they completed the rating scales for each of the eleven ITA speech samples.

The interview questionnaire queried about raters' process of evaluating the speech performance samples, their impression about the online rating, interesting cases of their ratings, or their prioritization of assessment components in speech rating of accented Englishes. For trained raters, questions included raters' impression of the one-hour intercultural intervention with ITA. Only 18 raters (7 untrained NSs, 2 untrained NNSs, 5 trained NSs, and 4 trained NNSs) out of 70 raters responded to the researcher's request for online open-ended

questionnaires. Two of them participated in online instant messenger tool, Gmail Chat Window. The other 16 responded to open-ended questionnaires sent by the researcher electronically (See Appendix E). Table 4.28provides codes necessary for identifying the sources of the quotations given below.

Table 4.28
ID Codes for Raters

ID Code	Description
NST	Native speaker who received training (socio-psychological intervention)
NNST	Non-native speaker who received training (socio-psychological intervention)
NSNT	Native speaker who did not receive training (socio-psychological intervention)
NNSNT	Non-native speaker who did not receive training (socio-psychological intervention)

Statements About Ratings Related to Raters' Own Background

Raters expressed various opinions about which of their own traits and experiences influenced their own rating process. Raters indicated that their evaluations could be influenced by their personal experience and native language status when they evaluated ITAs' speech samples.

The Familiarity Factor

First, a considerable amount of raters' comments was related to their exposure to non-native speakers or ITAs, or familiarity with NNSs' accents.

> I immediately recognized one of the TA's samples and so I think that created a bias opinion on my part. I think to the average ear his English may be somewhat hard to

understand but because I had him for a semester, it was much easier for me and I had no trouble in understanding him. (NST)

It was interesting to find and realize that some accents are equally as "thick" but vary completely in my understanding of the content, depending on what the speaker's native language was. I wonder if some native speakers are simply harder to understand in English because their native language differs so much for English (not necessarily their proficiency or how long they have been speaking English) or I'm not used to that accent? (NSNT)

I think the results and findings of the study would be very interesting. It is an important issue (language in teaching) that seems to affect even students' grades or class attendance!! I was frankly astonished by my inability to distinguish particularly between different Asian accents (Chinese, Japanese, Korean, Philipino, etc.). It was almost embarrassed- to be educated and exposed to different accents and ethnicities and have not a clue about how to distinguish between them. (NSNT)

Some instructors had unfamiliar foreign accent, and that really interfered me with understanding. I don't think they would be helpful in the class. (NSNT)

Negative Experiences with ITAs

In addition to exposure and familiarity with specific NNS speech patterns, raters spontaneously reported that their previous experience in courses taught by ITAs influenced the

oral proficiency ratings they assigned. The quotations below express raters' negative reactions to ITAs' courses.

> The most important thing to me (in this rating) was to be able to understand what the professor/T.A. was saying. I've had to drop classes before because I couldn't understand that teacher (and therefore couldn't learn). (NSNT)

> I actually knew one of the international TAs that was being rated, so when I listened to his speech I knew for sure how he looked like. That influenced how I rated him as well. It's bad. Although I tried to rate him as honestly as I can. (NST)

Rater Native Speaker Status

Raters' own native language status seemed to influence their ratings of NNS's proficiency. The quotes below support the quantitative finding of this study that NNS raters were harsher than NS raters. NSs and NNSs showed different perceptions of the importance of accent in speech. That is, NNS raters' stringency in their ratings seemed to be determined by their judgments of accentedness (i.e., degree of accent) in other NNSs' speech.

> I tried to be very patient and very critical when I listened to the speeches. I thought that no matter how proficient the person is in English, or how comprehensible, or how effective he/she is in teaching, if he/she has a strong foreign accent, he/she would be difficult to understand from native American English speakers. So I couldn't give them high scores. (NNSNT)

> There are many native English-speaking people who do not have high language proficiency- this doesn't seem to affect most people's understand as much as our accentedness would. (NNSNT)

On the other hand, a NS expressed a greater degree of leniency.

> I find it easy to understand most Indian-native speaking individuals even if they have a thick accent. Also, I know a Cuban family that speaks with very strong accents, but I understand their English easily. It really depends on which origin/accent is strong as to whether or not I can comprehend. (NST)

The following NNS rater's comment illustrates not only the influence of native language status but also the impact of the rater's formal linguistic training in university courses.

> I just stopped thinking about what my preconceived notions of the English language is supposed to be. The class that I just took talked about standard English and its aspect. I kept thinking about my own standards about good English. (NNST)

No rater comments addressed their own teaching experience nor their general attitudes (stereotypes) toward NNSs.

Statements About Ratings Related to Speaker's Speech Characteristics

A majority of raters who offered comments reported that the most important component of rating ITAs' speech performances was related to accentedness. Almost all the raters who provided open-ended remarks commented on speakers' pronunciation. Particularly the data collected from the post-rating open-ended comments showed that raters were especially aware of speakers' slow rate of speech, hesitant manner, and monotonous intonation. Speech rate was very frequently commented on by both NS raters and NNS raters.

> Speaking at a normal rate is also important. Speaking slowly isn't necessarily a bad thing, but too slowly and the students minds start to wander. (NST)

> This teacher spoke extremely slow. As a result, it was a little difficult to follow him and it made it seem as if he were not familiar with the material. (NNSNT)

> I feel as though he talked way too slow. In a typical undergraduate classroom, most of the students would probably fall asleep. They might also feel like they would be better off just reading the textbook than listening to him speak. (NST)

> His rate of speech was very frustrating. Very hesitant. Pronunciation of words was okay but I do not think that he is qualified to teach a class in English yet. I definitely would be frustrated going to class every day if I had him as an instructor. (NSNT)

Raters also commonly pointed out ITAs' pausing patterns including length of silent pauses, frequency of filled pauses (e.g., 'ah' or 'eh'), and repetition of words. Irregular use of silent pauses and frequent uses of filled pauses seemed to be the negative factors with which raters found uncomfortable.

> He used "Uhhh.." too many times and he sounded very unsure of what he was saying. He stops as if he forgot. This teacher would frustrate alot of undergraduate students. He takes too long to get the message across (NSNT)

> The "ums" and long pauses are off putting. I might lose attention if were in the classroom. This instructor was very hesitant in his speaking and made it hard to follow what he wanted to say. (NST)

> This teacher has great English proficiency, however has many pauses in the sampling of speech. If the pause were for dramatic effect, it may help undergraduate students pay attention better to what is coming next in what he says. (NSNT)

> He speaks slowly and says "ah" alot. You can kind of tell when he's thinking about what he's going to say next and seems to repeat certain phrases alot. (NNSNT)

Finally, raters' comments—although framed in lay terminology--addressed intonation-related aspects, volume, segmental enunciation, and lexical stress.

This teacher has a really bad pronunciation and intonation. It's difficult to catch what he says, not attractive. It is common in undergraduate classroom I think. (NST)

Tone was upbeat, instructor sounded interested in material however it was difficult to follow the lecture. (NSNT)

His pronunciation of English is acceptable, but because of his sudden change of soft and loud volume in speaking, it makes it frustrating to follow what he is saying. Moreover, the monotone in the delivery of his sentences makes his ideas extremely boring. (NSNT)

Mispronunciation is confusing, and the rate of speech is pretty slow. i would not consider someone who mispronounces key golf words like, "wood" to be knowledgeable on the subject. (NSNT)

Only a few miscellaneous responses were related to grammatical errors, poor organization, or the ITA's lack of confidence.

Rater Impressions of the Socio-Psychological Intervention (Training)

The online interview and open-ended questionnaire responses provided insight into the impact of the one-hour intercultural intervention between the subsample of raters and ITAs. The following excerpts support the quantitative findings that the intervention did influence raters' perceptions of ITAs' oral performances.

There wasn't much of a difference between phase one and phase two except that there was an informal meeting with the TA's and when i was doing phase two I could put a face of the TA to the speech sample because I had met him during the meeting (NST)

I did notice a bit of difference between Phase I and Phase II much, but maybe because I met ITAs at the informal meeting or something, I kind of felt a bit more comfortable with the ITAs' accent. (NST)

I really enjoyed the informal meeting because I felt that it was good to intereact with TAs I had never met. However, I wish I had interacted more with the other groups and the other TAs. (NNST)

I enjoyed the informal meeting. The meeting allowed me to express my ideas and learn from other individuals who come from different backgrounds than I. I didn't have any problem understanding ITAs' English then. It made you think a little bit more about Phase II, and picturing faces with accents. (NNST)

The representative quotes above indicate that the intercultural meeting between raters and ITAs engendered a more mindful and less prejudiced mind set among at least some of the raters. Overall, these qualitative data supported conclusions drawn from the quantitative findings in this study. As self-reported by raters, rater background factors and experience did influence ratings of ITAs' speech. Those factors included amount of exposure to ITAs' accents, quality of experience in taking courses taught by ITAs, native/non-native English language speaker status,

and formal training in linguistics. Also, raters were mostly aware of speaker's speech characteristics related to speech rate, pause structures, and intonation. Finally, the one-hour intercultural meeting enabled raters to adopt a more mindful and personalized mind set in rating ITAs' speech.

A few raters used the open-ended comments as an opportunity to evaluate the research procedure itself. Some raters expressed a negative reaction to the online rating system. One said that he/she had to enter a "4" for most of the ratings because it was impossible to judge so many specific characteristics without knowing the speaker. Most of the raters, however, agreed that the online rating procedures were user-friendly, convenient, and interesting.

Summary of the Results

The study has presented a new approach to assessing rater characteristics in speech evaluations and comparing the impact of rater background (biases) with the impact of objective features of pronunciation. The findings of the study suggested that raters from different language backgrounds had different perceptions of ITA accented speech. The results of regression analysis for each rater background (5 variables) and attitudinal factors (2 variables) as well as speaker's acoustic factors (3 clusters) on assessment outcomes are summarized in Tables 4.29 and 4.30 (multiple linear regression analysis in 4.29 and Mixed Random Coefficent Modeling analysis in 4.30). Among the 7 rater background variables, rater's teaching experience and rater's past negative experience in ITA courses were the most active and potent predictors in ITAs' oral performance judgments as seen in Table 4.29. The former was positively related to the oral performance assessments whereas the latter was negatively associated with the outcome variables.

Table 4.29

Summary of the Findings from Linear Regressions on Seven Dependent Variables

Variables	NNS status	Linguistic sophistication	Teaching experience	Time spent with NNS	Negative past experience in ITA course	Superiority -RLS	Social attractiveness -RLS	Acoustic fluency	Irregular boundary	Hesitation marker
Oral proficiency ratings	−		+		−			+		
Instructional competence ratings			+		−			+		−
Comprehensi--bility ratings			+		−					
Accent standardness ratings			+		−	−		+	−	
Superiority ratings	−	−	+					+		
Social attractiveness ratings			+		−		−			
Rater leniency (speaker ability) ratings	−			−	−			+		

Note: (1) +/− = Directionality of regression coefficients (i.e., + = positive beta weights; − = negative *Beta* weights)

(2) Statistically significant predictor variables marked ($p < .05$)

Table 4.30
Summary of the Findings from Mixed Random Coefficient Modeling on Six Dependent Variables

Variables	NNS status	Linguistic sophistication	Teaching experience	Time spent with NNS	Negative past experience in ITA course	Superiority -RLS	Social attractiveness -RLS	Acoustic fluency	Irregular boundary	Hesitation marker
Oral proficiency ratings	−		+	−	−			+	−	−
Instructional competence ratings			+	−	−			+	−	−
Comprehensi--bility ratings			+	−	−			+	−	−
Accent standardness ratings			+		−	−		+	−	−
Superiority ratings	−	−	+	−	−	−		+	−	−
Social attractiveness ratings		−	+	−	−		−	+	−	

Note: (1) +/− = Directionality of regression coefficients (i.e., + = positive beta weights; − = negative *Beta* weights)
(2) Statistically significant predictor variables marked ($p < .05$)

Among the three acoustic clusters variables, the acoustic fluency factor showed the strongest effects in predicting oral performance assessment throughout 7 outcome ratings.

The results of the mixed random coefficient regression as shown in Table 4.30 are generally consistent with those of the linear regression for the prediction of rater background and acoustic factors on ITAs' oral performance ratings. All through the 7 dependent variables, the teaching experience variable showed a strong positive relationship while rater's past negative experience in ITA course exerted a strong negative impact. Besides, due to the efficient reduction of error variance and increase in degrees of freedom, this mixed model analysis made the F-tests more powerful; all the three prosody cluster variables turned out to be statistically significant predictors of most of the rating outcome variables: oral proficiency judgments, judgments of instructional competence, perceived comprehensibility, and perceptions of heavy accent, of social superiority, and of social attractiveness. The acoustic fluency factor exhibited a positive association, and irregular boundary and hesitation marker exhibited negative associations.

Finally, from the results of the 2 (time) x 2 (training) repeated ANOVA, it was found that there was interaction between time and training in the ratings of instructional competence ratings and comprehensibility ratings. That is, trained raters found the ITAs more comprehensible and more instructionally competent after spending an hour of intercultural contact with them.

CHAPTER FIVE: DISCUSSION and CONCLUSION

Introduction

Non-native speakers of English (NNSs) are often subjected to evaluations of their spoken English that have profound consequences for their education, employment, and even citizenship. International teaching assistants (ITAs) face high stakes oral assessments on a regular basis, as they are under constant scrutiny from their native English speaking (NS) students, who often voice dissatisfaction with ITAs' oral proficiency and thus judge their teaching competence likewise harshly. This pattern of oral assessment is especially problematic because NS judgments of NNS speech are notoriously biased. NS listeners often hear what they expect to hear, rather than accurately perceive NNS speech. And what they expect to hear is often quite unsatisfactory. To obtain a "true score" estimate of NNS oral proficiency, one would need to rely on objective acoustically measures of accentedness.

Accordingly, the primary purpose of this study was to ascertain the proportion of variance in NS ratings of NNS speech attributable to measurable parameters of accentedness and the proportion attributable to potentially biasing rater characteristics. In other words, the study examined whether factors extraneous to speaker proficiency like rater background and attitudes exerted as much impact on perceived oral proficiency as did trait-relevant factors like speech rate, pausing, and intonation patterns. It further sought to determine the degree to which training (a brief socio-psychological intervention) could mitigate bias in NS listeners' ratings of NNS speech. The rating scores were analyzed using descriptive and inferential statistics. An online interview and open-questionnaire responses were used to better understand the raters' judgment processes.

This study has been guided by the following research questions:

1. What is the relative impact of rater background characteristics on ratings of L2 oral performance?

2. What is the impact of objectively measured suprasegmental characteristics of accented English on ratings of L2 oral performance?

3. To what extent does a course of training (a brief socio-psychological intervention function) affect ratings of L2 oral performance?

Accordingly, the following hypotheses were developed and tested:

H1: Oral proficiency ratings are inversely proportional to measured propensity to linguistic stereotyping.

H2: Rater background characteristics account for significant variance in ratings of oral proficiency.

H2a: Oral proficiency ratings conducted by native speakers of English differ from ratings conducted by NNS raters (nondirectional hypothesis).

H2b: Oral proficiency ratings are directly proportional to the amount of self-reported contact by raters with NNS friends and acquaintances.

H2c: Oral proficiency ratings are directly proportional to the amount of rater formal training in language and linguistics.

H2d: Oral proficiency ratings are directly proportional to the amount of rater's experience in teaching/tutoring English as a second language or foreign language.

H2e: Oral proficiency ratings are inversely proportional to the amount of rater's negative experience in taking courses taught by ITAs.

H3: The following acoustical properties of speaker's vocal productions account for significant variance in ratings of oral proficiency

H3a: The fluency factor directly predicts rated oral proficiency.

H3b: The level of the irregular boundary markers inversely predicts rated oral proficiency.

H3c: The incidence of the hesitation marking inversely predicts rated oral proficiency.

H3d: The acoustic parameters of speech rate, pauses, stress, and intonation contribute individually unique and statistically significant variance in predicting oral proficiency ratings.

H4: In ratings of NNS oral proficiency, the cluster of rater background characteristics and rater linguistic stereotyping, and the cluster of measured speaker acoustical properties both contribute unique and statistically significant variance in predicting oral proficiency ratings.

H5. Raters who received a socio-cultural *sensitization* intervention (training) are more lenient in oral proficiency ratings as compared with (a) their ratings prior to the intervention and (b) raters who did not receive such intervention.

This chapter summarizes the results of this study and considers their implications by investigating the relations between rater background and speech characteristics in ratings of oral performances in World Englishes. The findings of the study were compared and discussed with the results of a companion ETS-funded study of TOEFL® oral proficiency scoring where helpful, in order to affirm the conclusions drawn from this study. First an overview of the findings is provided. Then, each of the 12 hypotheses made for this study is discussed with descriptions and examples. Finally, this chapter concludes with implications, a proposed model of speech rating, limitations of the study, and recommendations for further research.

Overview of the Findings

The findings of this study indicate that rater background and attitudinal factors as well as suprasegmental parameters of speech samples did contribute substantial variance to ratings of ITAs' oral performances. That is, about 20-30% of the variance (R^2) in all of the seven rating dimensions--oral proficiency ratings, instructional competence ratings, comprehensibility ratings, accent standardness ratings, superiority ratings, social attractiveness ratings, and rater leniency scores ─ was attributable collectively to rater background and attitudinal factors. Approximately 60-70 % of the variance (R^2) in ITAs' oral performance ratings ─ ratings of oral proficiency, instructional competence, accent standardness, and superiority ─ was attributable collectively to objectively measured prosodic pronunciation factors. Note, however, that the acoustically measured suprasegmental characteristics did not significantly predict comprehensibility and social attractiveness ratings in the linear regression analysis.

The goal in language assessment is to reduce sources of variability that are extraneous or irrelevant to the learner's language performance to the greatest possible degree in order to mirror the candidate's true ability (Wigglesworth, 2001). The sources of trait-irrelevant variability in the assessment of L2 oral performance are diverse. The finding that up to 30% of the variance in oral performance ratings is caused by rater background and attitudes suggests that the judgment of listeners is indeed suspect. Consequently, it can result in measurement error. On the other hand, it is reassuring to find that these speech ratings were by no means completely independent of factors that are logically and empirically linked to true oral proficiency. At least partial validity of the ratings was warranted by strong relations with objectively measurable linguistic features. Suprasegmental elements of pronunciation contributed significantly to oral proficiency and other speaker ability ratings, which implies that raters actually did evaluate NNSs' speech samples to a

large degree (60-70 %) in a manner corresponding to relevant features of the speech and speaker's ability.

A complementary study, an ETS-funded study of TOEFL® speaking scores (Rubin, Kang & Pickering, 2008), showed similar results. About 20% of the variance in holistic ratings, in rater severity scores, and in the deviation between naïve raters' and ETS-trained raters' holistic scores was attributable to naïve raters' background and attitudinal factors. Also, about 60-75% of the variance of iBT TOEFL ® speaking scores was attributable to suprasegmental pronunciation factors. What is interesting, however, is that despite differences between these two studies in terms of speech samples used, raters recruited, and rating tools implemented, the results of these two studies were not dissimilar (but not exactly identical either). Rater background and attitudinal factors were slightly more potent in this present study.

One possible reason for the potency of rater background and attitudes in the present study is that raters here were all undergraduate students, and they rated speech performances of ITAs who could very well have been their instructors. These speech samples (college lectures) and this rating task (responding to an instructor) were contextually quite relevant to the undergraduate raters. Undergraduate listener's mostly negative expectations in evaluating ITAs' comprehensibility and teaching competence have been well documented (e.g., Rubin, 2002; see review in Williams, 2006). In contrast, a majority of rater participants in the ETS-funded companion study were graduate students or non-students such as teachers or researchers. The speech performances may have been high stakes for the examinees, but were less contextualized with respect to most of the raters' lives. TOEFL® raters can maintain more of a professional distance from the objects of their judgment. Thus raters who scored high stakes English

proficiency tests might have engaged in different rating processes than those who graded ordinary classroom discourse of NNS teachers.

The findings of this study confirm previous research findings that ratings of speaking skills are susceptible to rater expectation and stereotype, because listeners can be so prone to render social inferences about speakers on the basis of speech (Bradac, Cargile, & Hallet, 2001; Piché, Michlin, Rubin & Sullivan, 1977). This study's findings are particularly compatible with the view that ratings of speaker accent are distorted by listeners' expectations (Nisbett & Wilson, 1977). That perceptual distortion is especially potent within the frequently troubled relations between undergraduate students and ITAs (Smith, Strom, & Muthuswamy, 2005). The two reverse lingustic streotyping (RLS) measures—superiority and social attractiveness--were found to be strongly related to ratings of ITAs' accent standardness and social attractivness in particular. Note that in the MRCM mixed model anaysis, the RLS measures were strongly and inversely related to superiority ratings as well.

On another note, the intercultural sensitization intervention employed as rater training in this study exerted impact on key ratings of ITA oral performances. The repeated measures ANOVA (time x training status) showed that the "trained" raters found the ITA speech more comprehensible and more instructionally competent at Phase II, after participating in a collaborative discussion activity, compared with their ratings prior to the sensitization activity. The mean scores assigned by other raters, who did not experience the sensitization activity, did not change from Phase I to Phase II. This impact on rating behaviors constitutes a surprising finding, because the intervention was so short in duration (i.e., only for an hour) and quite limited in intensity to bring about any profound change.

The psycho-social intervention in this study was designed to meet several of the criteria empirically verified as effective in prejudice reduction (e.g., equal status among participants, collaboration necessary for success, opportunity to get to know the "other" personally (Voci, 2003). Yet in cases of serious and intractable conflict, even the most elaborately designed intercultural contact has little effect on prejudice reduction (see review in Rubin & Lanutti, 2001). Nonetheless, a measurable impact was apparently achieved in the present instance. This result lends support to the contact hypothesis (Allport, 1954), which holds that interactional contact between two groups has positive effects on intergroup attitudes and can reduce predjudice under certain conditions. As seen in this study, informal and pleasant contact with interpersonal intimacy and equality can bring a positive change in undergraduate attitudes toward ITAs and consequently influence undergraduates' perceptions of ITA speech performances. Open ended comments suggested that as a result of the brief interaction, the U.S. undergraduates felt that they had achieved greater familiarity with the ITAs and their speech patterns. The success of this brief intervention lends credence to the view that comprehension of ITA speech is in part a function of the undergraduate's motivation to listen with an adaptive mind set (Rubin, 2002).

With respect to the psycho-social intervention, , it is important to note that training in this study did not employ typical instruction about the scoring rubric and then calibration with anchor-point speech samples (e.g., Weigle, 1994). For example, the companion ETS-funded iBT TOEFL® study informed a subset of raters about evaluative criteria and standards in that a subset of participants was exposed to similar online training as that available to iBT TOEFL® raters. After training, in the ETS study, the impact of the extraneous rater variables was reduced by about 75% for holistic ratings. Rating outcomes which prior to training had diverged

dramatically from ETS-endorsed scores converged significantly after training, and trained raters showed higher levels of reliability (Rubin, Kang, & Pickering, under review).

After training pertaining to rating rubrics and anchor points, raters in earlier studies tended to be more consistent in applying rating criteria to ESL composition (Weigle, 1984). However, trained raters (rather than naive ones) in an earlier study involving oral assessment in particular showed great variability in their severity of ratings even though they tended to increase their internal consistency in assigning ratings as they gained experience and repeated training (Bonk & Ockey 2003). Similarly, this study examined the extent to which the inter-cultural meeting could affect raters' judgments of ITAs' oral performance.

Even though the term, 'training', was interchangeably used with 'socio-cultural intervention' in this study, the primary interest of the training effect in the present study was to investigate whether raters who received the intervention perceived NNSs more positively than did raters who did not have the same structured interactional experience. Therefore, implications for the training effects extend beyond the field of language assessment *per se*, to the broader disciplines such as inter-cultural communication or cross-cultural psychology (e.g., Gudykunst, 1991).

Evaluation of Hypotheses

H1: Oral proficiency ratings are inversely proportional to measured propensity to linguistic stereotyping.

The findings of this study partially support H1 in that oral proficiency ratings on some of the dependent variables—but not on all—were inversely proportional to measured propensity to linguistic stereotyping. The degree of linguistic stereotyping manifested by listeners was measured using the Speech Evaluation Scale (Zahn & Hopper, 1985). The reverse linguistic

stereotyping (RLS) measures were the indices of attitudinal factors used in this study. Whereas the linguistic stereotype hypothesis posits that listeners ascribe to speakers stereotyped traits on the basis of their speech (Lambert, et al, 1972), RLS is based on the converse notion that listeners ascribe stereotyped characteristics to speech ("hear" nonstandard accent where none may be present, for example) based on social information about the speaker's identity.

The Speech Evaluation Scale (SEI) from which the RLS measures were derived exhibited internal consistency reliability within each of the three subscales, a finding which was parallel with the original three-factor-loaded structures reported by Zahn and Hopper (1985).

Among the two RLS measures further employed in this study, superiority-RLS was strongly and inversely related to perceived accent standardness. And social attractiveness-RLS was inversely related to ITAs' social attractiveness ratings. High RLS scores represent high stereotyping activity; i.e., higher RLS scores imply that raters more negatively stereotyped NNSs on these dimensions. Therefore, raters who showed a tendency of negative stereotyping of NNSs− for example, considering NNSs as lower class, uneducated, unintelligent, illiterate, and so forth─ rated ITAs speech more deviated from standard accents. In addition, with regard to social attractiveness-RLS, the more negatively raters stereotyped NNSs ─ considering them as unfriendly, cold, hostile, dishonest ─ the more harshly they rated ITAs in terms of superiority. In other words, raters with higher stereotyping proclivities tended to perceive ITAs as not likeable, not kind, and not appealing. However, RLS measures did *not* significantly predict oral proficiency ratings, instructional competence ratings, or comprehensibility ratings.

H2: Rater background characteristics account for significant variance in ratings of oral proficiency.

In addition to the two RLS measures described above, five rater background factors were selected to test a model of the effect of trait-irrelevant rater background factors on ratings outcomes. Those selected background factors were non-native speaker status, language sophistication, amount of contact with NNSs, amount of teaching or tutoring experience, and negative experience in ITAs' courses. Each of these rater background variables showed statistically significant relations with one or more of the seven outcomes variables.

H2a: Oral proficiency ratings conducted by native speakers of English differ from ratings conducted by NNS raters (nondirectional hypothesis).

NNS speaker status was a significant predictor of ITA speech ratings. Nonnative speakers were found to be significantly harsher in oral proficiency ratings and superiority ratings than were native speakers. These results are consistent with Brown's (1995) findings that Japanese raters (NNSs) were substantially harsher than NS raters of English in evaluating the pronunciation of Japanese English language learners.

It is not surprising that NNS assessors, who have gone through a complex and arduous learning procedure themselves, are less tolerant of other's mistakes. Santos (1988) reported that when NNSs rated other NNSs' writing skills, the raters' own efforts in attaining a high level of proficiency led them to attribute writing errors to a lack of commitment on the learner's part. An excerpt from a NNS raters' online interview in the present study is a case in point.

Excerpt 5.1

…if he/she has a strong foreign accent, he/she would be difficult to understand from native American English speakers. So I couldn't give them high scores. (NNS2)

Excerpt 1 comports with the view that ESL/EFL learners who have struggled to acquire an Inner Circle English pronunciation (Pickering, 2006) as a target may undervalue accented Englishes which are different from that norm. In contrast, NS raters may have a more casual view of what it means to attain proficiency in English, and not worry about nonnative features as long as they do not seriously impede communication.

NNSs' tendency to severe scoring standards is also consistent with psycholinguistic findings. Recent studies (e.g, Deterding & Kirkpatrick, 2006; Kirkpatrick, 2007) showed that NNSs' pronunciation features ─ if they were not shared by the listener's own native language ─ hindered comprehensibility for NNS listeners. Moreover, another psycholinguistic study measuring speech intelligibility, found that Dutch listeners did not even benefit from hearing their own non-native accent in a second language, and instead they found the native English speakers more intelligible (Van Wijngaarden, Steeneken, & Houtgast, 2002). Thus, because English language learners may experience greater difficulty than NSs in understanding English produce by NNSs, they might be expected to rate NNS speech performances more harshly, as was found in the present study,

H2b: Oral proficiency ratings are directly proportional to the amount of self-reported contact by raters with NNS friends and acquaintances

Based on the findings of this study, this hypothesis is not accepted, given that the more time raters spent with NNSs, the *harsher* they became in the oral proficiency ratings. For example, the more time raters spent with NNSs, the lower were their leniency scores (derived from FACET Rasch analysis on oral proficiency rating scores). Time spend with NNSs was also found to be inversely associated with other rating outcomes such as judged instructional competence, comprehensibility, superiority, and social attractiveness ratings in the alternative MRCM analysis results.

This finding about the impact of contact with NNSs is inconsistent with some earlier studies about the role of familiarity with NNS accents. Previous research findings concurred with the view that amount of interaction with speakers of specific languages or particular varieties of World Englishes facilitate listening comprehension of those English varieties (Clark & Garrett, 2004; Field, 2003; Gass and Varonis, 1984). Individuals unfamiliar with a particular World English variety generally perceive a higher degree of L2 foreign accent than those who are familiar with that particular variety (Thompson, 1991).

However, unlike this present study, listeners in most of those previous studies were native speakers who listened to NNSs' and then rated it on either comprehensibility or intelligibility. They did not, in those earlier studies, rate accented speech for proficiency, as was done in the present study. Also, contact with NNS accents in this study was indexed just as hours per week spent with NNSs. It did not specify the types of interaction or the particular language varieties to which raters were exposed. In the particular locale in which this study was

conducted, most NNS contact would be with native speakers of Mexican or Guatemalan Spanish, a World English variety not included in the speech samples presented for evaluation here.

Some previous studies (e.g., Mattran, 1977) concur with the present findings in that extensive experience in NSs' evaluation on NNSs' speech was not necessarily positively related to accurate judgments. In other words, Mattran's study showed there was no statistical difference found in ratings of NNSs' speech, between NSs who had no systematic experience in dealing with NNSs and NSs who were linguistically sophisticated and experienced. In fact, another study found that familiarity with accented English (or speakers even sharing listeners' native language) did not automatically facilitate listeners' listening comprehension, but resulted in rather complex outcomes of variance depending on the L1 of the listeners (Major et al., 2002). Overall, this hypothesis needs to be further tested in the future, for example, by investigating the impact of this NNS' contact variable for each specific contact language separately.

H2c: Oral proficiency ratings are directly proportional to the amount of rater formal training in language and linguistics

The results of the study failed to establish a significant relation between oral proficiency ratings and rater formal training in language and linguistics. However, the amount of rater language and linguistic training was inversely related to one social impression measure. In the linear regression model, the degree of language sophistication showed strong and negative impact on superiority ratings. This negative relationship was extended to social attractiveness ratings when all the predictor variables (i.e., the rater background and attitudinal factors and suprasegmental factors) were analyzed in the mixed random coefficient model.

Even linguistically naïve listeners are quite sensitive to cross-linguistic differences in prosodic patterns when judging foreign accents (Anderson-Hsieh & Koehler, 1988; Brennan & Brennan, 1981; Munro 1995). Yet, in this study, the higher language sophistication was indexed, the harsher were raters' social impressions of NNSs. Data from the present study do not offer any reasons why linguistically more trained raters tended to have more negative impressions than less trained raters in terms of stereotyping outcome ratings. To the contrary, one would think that ability to contrast structures across languages would inoculate listeners against making social judgments based on speech patterns. In this study, linguistic sophistication was derived by summing (a) the number of college classes in linguistics, applied linguistics, or TESL endorsement classes and (b) years of foreign language study. It can be speculated, however, that raters who themselves have put great effort in learning foreign languages may set up higher standards and expectations for ITAs' oral performances. Also, with linguistic sophistication may come a mind set that compels the listener to be especially discerning in evaluating speech. If a high degree of education in linguistics and language equates with harsh judgments of NNS speech, this finding suggests that greater emphasis needs to be placed in language courses on descriptive rather than prescriptive approaches to language variation.

The following excerpts are open-ended responses taken from raters who showed particularly high scores in the index of language sophistication. Raters' concerns included traits that could explain their rating process of superiority judgment such as 'uninteresting, cold, and unclear".

Excerpt 5. 2

Very harsh and unapproachable sounding. did not say "excuse me" after sneezing twice. very dry and **uninteresting** lecturer. I wouldn't want him as an instructor. (NS1: 15 linguistic and TESOL classes taken and 9 years of foreign language studied)

Excerpt 5.3

It would be nice for this teacher to express ideas more clear and be more attentive for the students needs. His speech contains no emotion; rather, its monotone and boring. His speech was too disjointed and sounded a bit cold.

(NS2: 4 linguistic and TESOL classes taken and 11 years of foreign language studied)

H2d: Oral proficiency ratings are directly proportional to the amount of rater's experience in teaching/tutoring English as a second language or foreign language

Amount of previous teaching experience was a potent predictor in this study. ITAs' oral proficiency ratings were directly proportional to the amount of rater's experience in teaching/tutoring English as a second language or foreign language. Therefore, H2d is affirmed. Indeed, amount of teaching experience significantly and positively predicted most of the rating outcomes in this study, viz., oral proficiency ratings, instructional competence ratings, comprehensibility ratings, accent standardness ratings, superiority ratings, and social attractiveness ratings.

Undergraduates who have taught languages in the past seemed to be more lenient raters of ITAs' oral performances. This result is substantiated by Barnwell's (1989) study, which reported the non-teaching raters were relatively harsher than the teaching rater group. In contrast,

this teaching experience variable did not show any effect in iBT TOEFL ratings in the ETS-funded study.

H2e: Oral proficiency ratings are inversely proportional to the amount of rater's negative experience in taking courses taught by ITAs

H2e was confirmed; oral proficiency ratings were inversely proportional to the amount of rater's negative experience in taking courses taught by ITAs. In fact, raters' negative experience in ITAs' courses was very strongly and negatively predictive of most of the oral performance assessment measures: low proficiency judgments, judgments of instructional incompetence, perceived incomprehensibility, perceptions of heavy accent, of social inferiority, and of social unattractiveness. The result is not unexpected since biases constructed through previous experience with ITAs can affect raters' general expectations of ITAs' speech performances. This expectation effect holds true, even when the actual speech stimulus is thoroughly Standard American English (Rubin, 1992). Previous research conducted (Rubin, 2002; Lindemann, 2002) documents that these student complaints are frequently more a function of students' stereotyped expectations than of instructors' objective language performance. Therefore, raters with the negative attitudes and expectations established through previous experience with ITAs could negatively stereotype ITAs in general. Indeed, it is highly likely that those prior negative experiences with ITAs were themselves the products of negative self-fulfilling prophesies.

It has been known that undergraduate students quite often object to being taught and graded by ITAs (Fitch & Morgan, 2003). In addition, 40% of undergraduates at some point in their educations dropped or switched classes because the instructor was a NNS (Rubin & Smith, 1990). Excerpt 5.5 taken from the open-ended questionnaire responses also illustrates rater's

critical attitude toward ITAs. This person's rating scores for all of the 6 outcome variables were extremely low, compared to the average scores of other raters.

> Excerpt 5.4
>
> The most important thing to me (in this rating) was to be able to understand what the professor/T.A. was saying. I've had to drop classes before because I couldn't understand that teacher (and therefore couldn't learn). (NS3)

H3: The acoustical properties of speaker's vocal productions will account for significant variance in ratings of oral proficiency

Nonnative temporal (e.g., pause structures) and tonal patterns (e.g., tone choices), account, at least in part, for native listeners' perceptions of L2 English learners' speech as "accented" (Shah, 2002). The present study included acoustic measures of 12 different pronunciation parameters. They were clustered though hierarchical cluster analysis into three prosodic factors: acoustic fluency, irregular boundary, and hesitation markers. Collectively they accounted for a very substantial amount of variance in ratings: up to 70% on some measures. This variance can be considered true score variance with respect to speech assessment. The factor which accounted for the greatest amount of variance in the system of acoustically measured prosodic variables was acoustic fluency. Hesitation markers were strongly and inversely related to instructional competence ratings. Qualitative findings revealed that many raters were quite attuned to prosodic factors when making their evaluations of proficiency and comprehensibility of NNS speech.

H3a: The fluency factor directly predicts rated oral proficiency

The acoustic fluency cluster was comprised of typical measures of speech rate; namely, articulation rate, syllable per second, mean length of run, and phonation time ratio. It also included number of prominent syllables per run – pace. A prominent syllable is a stressed syllable which determines listeners' perceptions of NNSs' speech (Field, 2005). In addition, the fluency cluster captured an element of intonation, the overall pitch range. Finally, (in fact, rather unexpectedly), it encompassed the number of silent pauses, which would be expected to weigh negatively in calculating the composite variable. Note, however, from Table 4.9 that the frequency of silent pauses was very weakly correlated with most of the other suprasegmental variables. In terms of silent pasue production, research has shown that NNSs' pause patterns appeared to be more frequent, longer, and irregular than those in NSs (Anderson-Hsieh & Venkatagiri, 1995; Pickering, 1999; Riggenbach, 1991; Rounds, 1987). However, some recent research findings showed that there were no statitistical differences to be found in the amount of silent pause production between advanced and the low-intermediate learners of English (Kormos & Denes, 2004) or between NS TAs and NNS TAs (Kang, 2008). Even though the number of silent pauses was grouped into the acoustic fluency in this study based on the hierarchical cluster analysis, the effect of this variable on rating outcomes was small. Overall, the relationships with oral perforamnce ratings and pause productions may require futher resrach.

The acoustic fluency factor directly and positively predicted ITAs' oral proficeincy ratings. It was the most potent variable in predicting all other 6 remaining rating scores. Results concerning rate measures concur with previous studies investigating fluency (Anderson-Hsieh & Kohler, 1992; Ejzenberg, 2000; Freed, 2000; Riggenbach 1991, 2002). Similarly, the stress measure, pace, was previously found to be a reliable predictor of fluency judgments (Kormos &

Denes, 2004). The pitch range is one of the major intonation features that affect NSs' comprehension on NNSs' speech. NNSs tend to show a compressed pitch range and a lack of variety in pitch levels which can lead to a succession of mostly high or mostly low strings of syllables (Mennen, 1998; Pickering, 1999; Wennerstrom, 2000). For example, Finnish speakers of English use significantly narrower pitch ranges than native speakers of English (Hirvonen 1967; Toivanen 2001).

The results are pareralled with those of the ETS-funded project, the intra-run fluency whose components were simlar to acoustic fluency in this present study, i..e, most of the speech rate meaures, pace, and pitch range measaures. The most potent clustered variable was this intra-run fluency which was positively related to holistic ratings of iBT TOEFL speech. In addition, in reviewing the online intereview and open-ended questionnaire responses, the most frequent phrases that raters used to describe speech characteristics of ITAs samples were either 'slow/fast speech rate' or 'mono-tone'.

H3b: The level of the irregular boundary makers inversely predicts rated oral proficiency.

The irregular boundary cluster, which consisted of the proportion of prominent words-*space*, mean length of silent pauses, and the proportion of irregular topic boundary, were inversely related to the ratings of ITAs' oral proficiency. Thus hypothesis H3b is rightly affirmed. In the linear regression model, the irregular boundary cluster was most strongly related to ratings of accent standardness. In the random mixed model analysis, because of its greater power, irregular boundary marking was found to be inversely related to nearly all of the rating outcome measures. The regularity of pauses was certainly a matter of conscious concern for raters. The following open-ended comment was provided by a NS rater.

Excerpt 5.5

"This teacher … has many long pauses in the sampling of speech. If the pause were for dramatic effect, it may help undergraduate students pay attention better to what is coming next in what he says." (NS4)

Often in ITAs' speech performances, pauses (particularly long pauses of 0.8 or above) occurred in the middle of a sentence in a structurally and rhetorically haphazard way. In contrast, more effective instructors employ strategic silence to create a certain intentional rhetorical or dramatic effect, such as the pregnant pause that comes just before a punch line is delivered or a major point is made in a lecture (Rounds, 1987). Examples from both ITAs' speech and US TAs' speech are as follow: (Note: '//' stands for a run which is determined by pauses of 0.1 and above; numbers in brackets represent the length of pauses in seconds).

Expert 5.6 (ITA7)

So // **(.86)** // this is visual represation, // (. 27) // representation of mathematical proof. // (**.86**) // And uh (. 42) // (**. 83**) // this uh (. 27) thing // (. 29) // connects // (. 17) // some of ideas // (. 33) // to each of uh (. 12) like // (. 13) // that alge, algebra and the geometry are connected to // (1.03) // uh (.27), to each other // (. 58) // that means of // (. 16) // this concept. // (**. 92**) // And this is help students // (. 13) // to understand // (**. 91**) // the, // (.13) // thuh (.43) essence of the proof, and // (**1.18**) // helps students to understand another way, // (**.88**) // many different ways of // (. 13) // proving // (**2.22**).//

Excerpt 5.7 (US TA2)

…for example, you could have something depending on whether it's raining or not, // (.51) and depending on where you are one day it could rain, // (.37) // you know, and in some places it might rain ten percent of the time, // (.40) // in other places it might rain a hundred, you know, that kind of thing. (.19) So this says, this says nothing about // (.12) // how frequently, // (.27) // just because there's three cases doesn't mean it's gonna // (.34) // execute like that. // (**1.28**) // But the default case often is uh (.47) // (.60) ….

As seen above, the ITAs' long pauses are very frequently and irregularly produced as compared to the US TA's pause pattern. The speech flow is not smooth with many long silences at locations other than clause boundaries. The topic pause boundaries are defined as pauses of 0.8 seconds or longer which clearly coincide with major semantic breaks (Brown, 1977; Brown & Yule, 1983). The ITA topic pauses in Excerpt 5.6 were produced without any patterns and sometimes extremely long (i.e., 2 seconds or longer).

H3c: The incidence of the hesitation marking inversely predicts rated oral proficiency

In the linear regression model, the hesitation markers did not show any significant effects on oral proficiency ratings, but exerted an inversely proportional effect on instructional competence ratings. Hesitation markers were weakly related to other outcome variables such as comprehensibility ratings, accent standardness ratings, and linguistic stereotyping ratings. The effect of this hesitation marker seems to be less potent than the other two acoustic clustered factors in ratings of NNSs' oral proficiency ratings. Perhaps, hesitation features are somewhat based on individual speaking style. Flucher (1996) argues that low and high- proficiecny learners

create different impressions on listeners, not because their use of hesitations markers is different, but because they hesitate for different reasons. However, please note that in the mixed model analysis, this variable illustrated significant relationships with most of the outcome ratings. In other words, this hesitation maker factor does affect oral performance ratings, but the results of inferential statistics may vary depending on the statistical power achieved.

A rater's following comment on their rating procedures helps interpreting the strong relationships between the hesitation marker and the instructional competence ratings.

Excerpt 5.8

"He used "Uhhh.." too many times and he sounded very unsure of what he was saying. He stops as if he forgot .This teacher would frustrate alot of undergraduate students. He takes too long to get the message across." (NSNT)

Listening to frequent and long hesitation markers produced by ITAs, raters tended to question the level of ITAs' content knowledge rather than accusing them of low language proficiency ability.

H3d: The acoustic parameters of speech rate, pauses, stress, and intonation contribute individually unique and statistically significant variance in predicting oral proficiency ratings.

As discussed above, the strongest predictor of ITAs' oral proficiency ratings was the acoustic fluency factor, which included acoustical measures representing all the four categories: speech rate measures, stress measures (pace), silent pause measures, and the overall pitch range. Seven outcome rating scores were highly correlated with the acoustic fluency factor, as indicated

in Table 4.12 in Chapter 4. However, the effects of irregular boundaries and hesitation marking were distinct, and varied depending on what rating outcomes were regressed against these acoustic factors. Therefore Hypothesis H3d is not affirmed, in that the four types of suprasegmental features did not remain distinct, as predicted.

One of the most interesting findings of the Hierarchical Cluster Analysis is that ITAs' discourse-contingent stress—*space* (proportion of prominent stress words) was clustered into the irregular boundary factor, and was negatively associated with rated oral proficiency. Wennerstrom's (2000) study reported that a typical pattern of low fluency speakers was to associate high pitch with all words, regardless of their function. To be precise, low-fluency speakers tend to give relatively equal pitch to each word regardless of its role in the discourse structure, which leads to many sequential high-pitch words (i.e., stressed words in this study) or a flat monotonous string of words. One of the distinctive characteristics in ITA speech performances was that low proficiency ITAs tended to place stress on many function words or articles such as "be" or "the" in addition to the content words, whereas native speakers tend to distress those function words. The ITA stress pattern makes their speech more accented and less comprehensible.

Hierarchical Cluster Analysis had been conducted to reduce the data by aggregating the 12 acoustical measures of prosody into a smaller number of clustered variables (three). But to tease out the distinct impact of each of those twelve components of the clustered variables, a post hoc analysis was conducted in which correlations were run among the seven rating outcome variables and the twelve disaggregated individual suprasegmental measures. The correlations[11] revealed that most of the dependent variables were particularly highly associated with the

[11] The 6 (dependent variables) x 12 (suprasegmental measures) correlation matrix appears in Appendix K

following disaggregated suprasegmental features: speech rate, articulation rate, mean length of run, space, ratio of irregular topic boundary, and overall pitch range.

The contribution of each individual acoustic measure to oral language assessment can be confirmed by oral proficiency testing that the ITAs underwent independent of this research project. Table 3.4 in Chapter 3 showed PRAAT analysis values of ITAs' oral performance and their English proficiency test scores assigned by TOEFL® raters. ITAs (e.g., ITA1, ITA5, ITA7, and ITA8)--the four who received the top scores in their TAST (a free standing speaking section of the iBT TOEFL® that was available at the time) speaking proficiency tests--manifested the highest values of all the ITAs in speech rate (syllable per second), articulation rate, mean length of run, phonation time ratio, and overall pitch range. ITA7, who scored the highest in the TAST test, was the most fluent speaker of all in terms of high speech rate and articulation rate; and his pitch range was the widest of all. Similarly, acoustic features reported for native TAs in Table 3.4 support this pattern.

H4: In ratings of NNS oral proficiency, the cluster of rater background characteristics, rater linguistic stereotyping, and the cluster of measured speaker acoustical properties all contribute unique and statistically significant variance in predicting oral proficiency ratings

In predicting oral proficiency ratings, rater's native language status, teaching experience, and previous experience in ITAs' courses in particular contributed statistically significant variance. The rater's native language status was the most powerful predictor of iBT TOEFL rating scores (Rubin, Kang, & Pickering, under review). The amount of NNSs' contact and linguistic sophistication were weakly related to oral proficiency ratings. But variance in rater severity-leniency ability scores was significantly contributed by the former and variance in

superiority ratings by the latter. The rater attitudinal factors showed no effects on ratings of oral proficiency ratings per se, but strongly affected accent standardness ratings or the linguistic stereotyping ratings. In terms of acoustical properties to predict oral proficiency ratings, the acoustic fluency was the most potent variable which was significantly related to all of the 6 judgment tasks. The irregular boundary showed moderate effects but the hesitation markers exerted no effects in ratings of ITAs' oral proficiency.

H5: Raters who received a socio-cultural sensitization intervention (training) are more lenient in oral proficiency ratings as compared with (a) prior to the intervention and (b) raters who did not receive such intervention

The intercultural sensitization intervention, functioning as a kind of rater training, was designed with the hope that the trained raters could return to the rating task differently attentive to ITA speech characteristics rather than with an overwhelming negative disposition toward ITA speech. The training was a social-psychological inoculation against linguistic stereotyping, consisting of a one hour informal meeting between a random sample of ITAs and a random sample of raters. Each of the trained group members participated in a collaborative problem-solving task and exchanged their cultural backgrounds and academic commitments with each other. As seen in excerpts in Chapter 4 (e.g., "*I did notice a bit of difference between Phase I and Phase II much … because I met ITAs at the informal meeting …I kind of felt a bit more comfortable with the ITAs' accent*"), undergraduate raters who participated in the intervention did seem to attend to ITAs' speech differently at Phase II.

The repeated measures ANOVA results revealed significant time by group interaction effects for instructional competence ratings and comprehensibility ratings. Members of the

"trained" group found ITA speech more comprehensible and more instructional competent after they received an hour of a socio-physiological intervention, as compared with prior to the intervention. On the other hand, ratings of instructional competence and comprehensibility in non-training group remained relatively steady between Phase I and II. Other outcome rating scores; namely, oral proficiency rating, accent standardness, superiority rating, and social attractiveness rating of the trained group did not reveal similar interaction effects.

The post hoc comparison of cell means for the significant interaction showed that the untrained group raters at time 1 were significantly more lenient (gave higher scores) than trained-group raters at time1. However, at time 2 — after the intervention was administered — the two groups no longer differed. The trained group raters became considerably more lenient in their ratings at time 2, after the intervention.

The findings of this study comport with the view that undergraduates must learn more about intercultural communication (Inglis, 1993) and improve their listening skills (Rubin, 1992; Rubin & Smith, 1990). In addition, the results of this study suggest that the factors influencing undergraduates' perceptions of ITAs such as anxiety, prejudice, and social stereotyping are malleable. Consequently, those negative stereotyping factors can be adjusted through effective prejudice-reducing contact. Earlier studies (Yook, 1999; Yook & Albert, 1999) have already shown that structured inter-group contact between ITAs and undergraduate students can improve undergraduates' ability to empathize and to take the perspective of their ITAs. As undergraduates showed greater empathy for ITAs through the inter-cultural contact in those studies, their comprehension of course materials seemed to have increased (Yook & Albert, 1999). In this respect, the findings of this study concurs that effective intercultural contact can mitigate negative attitudes toward NNSs and it can affect NNSs' performance ability

A Model of Speech Rating

Models in language testing are often presented to better understand the assessment process and the roles of variables that affect rating scores assigned. In Chapter 2, models of McNamara (1996) and Skehan (1998) were criticized due to the absence of accounts related to trait-irrelevant rater background variables in the assessment process. Even though those models clearly recognized the rater as an important factor that influence test scores, they failed to detail in which way and to what extent rater characteristics affect the rating of L2 oral performances. Skehan's (1998) model has been often used in language testing research to investigate variables' impact on test scores in L2 oral tests (Fulcher, 2003). Therefore, a new model can be proposed in order to fill the gap in influential models of oral performance assessment by acknowledging the impact of rater training and rater characteristics. The proposed model is primarily an emendation of Skehan's (1998) model structure introduced in Chapter 2. In addition, it is supported by Kim's (2005) proposed model in which the importance of testing purpose is taken into account in ratings of oral performances.

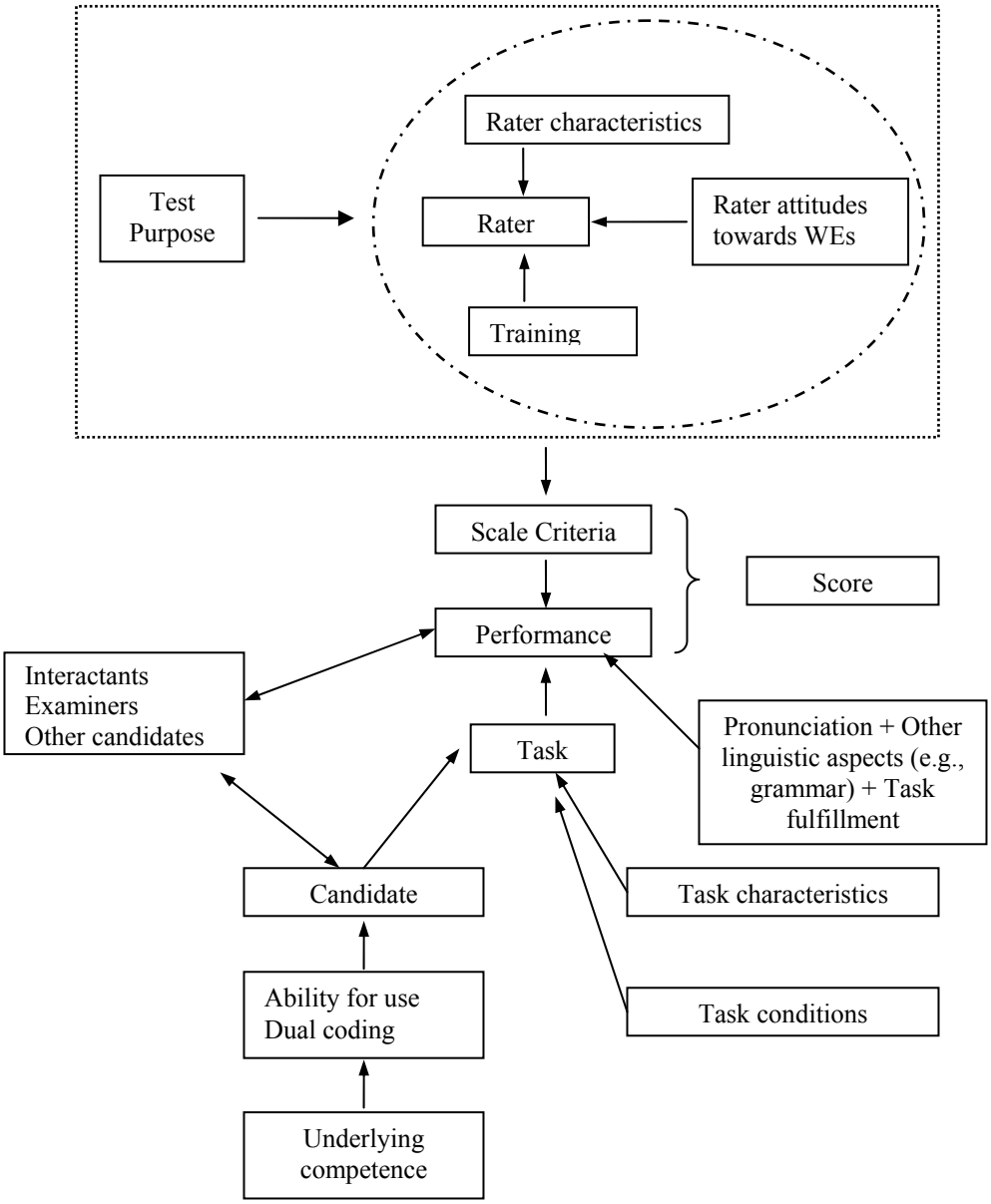


Figure 5.1
A Model of Oral Performance Assessment

The model in Figure 5.1, as currently proposed, does elaborate on the rater background and attitudinal variables that play important roles in the rating process. Furthermore, the proposed model points out that a rater training process can be incorporated into the overall assessment operation so that subsequent methods of screening and selecting raters can be

developed, if necessary, according to test purpose. Test scores are closely related to raters as well as tasks and rating scales.

In addition, the proposed model details the assessment elements of speaker's performances in the process of ratings. According to the findings of this study, over 60% of variance in oral performance assessment was attributable to speaker's pronunciation factors. Consequently, ratings of oral performance should consider various linguistic features (e.g., pronunciation), along with more rhetorical factors like task fulfillment, that determine speaker's oral proficiency.

Implication of the Study

Rater effects in the scoring of oral proficiency examinations constitute measurement error because speech assessment is consummately sensitive to listener expectations and social stereotypes. Rating discrepancy caused by rater characteristics constitutes a serious impediment to assuring test validity, thereby incurring the mistrust of the language assessment process itself. Previous research has well documented rater differences in severity (e.g., McNamara, 1996) or salience of rating criteria (e.g., McNamara 1990, 1996). Statistical methods (based on Rasch models and G-theory; e.g., Lynch & McNamara, 1998) have been developed to statistically control for such rater deviations. Yet were it possible to ascertain individual attitudinal and experiential characteristics that predisposed raters toward greater or lesser accuracy in rating speech samples, corresponding methods of screening, selecting, and training raters could be devised.

The present study represents an innovative approach to assessing rater characteristics that are likely biasing factors in speech evaluations and comparing the impact of those "nuisance"

rater effects with the impact of features of pronunciation which are legitimately components of "true score" variance. Consequently, the findings of this study challenge the validity of oral proficiency rating. Similarly, this study has implications for the interpretations of the assessment of non-native speakers' English language oral performances, for rater training, and for the evaluation of NNS' English language oral proficiency. In fact, it has been recently argued that international tests of English language proficiency are unfair to speakers of non-standards forms of English (Hamp-Lyons & Alan, 2008). Therefore, collaborative projects among researchers in language assessment, World Englishes, and linguistic analysis are needed to better develop assessment criteria, and to implement assessment training.

As to oral proficiency assessment practice, implications drawn from the findings of this study are as follow. First, recruitment of oral assessors should take rater's NS/NNS status into account, because if raters are non-native speakers, higher stringency is to be expected. In any event, test administrators should recognize that pools of raters comprised of both NSs and NNSs will not be homogenous with respect to rating standards.

Second, raters with professional experience as teachers have a high potential to be relatively lenient raters. The present study offers no benchmark against which to evaluate the "accuracy" of raters, so it is quite possible that the trend toward relative leniency in this context may (or may not) represent rater "accuracy."

Third, raters' formal preparation in languages and linguistics may not be directly relevant to NNSs' oral proficiency rating scores, and it potentially affects oral performance ratings of other types such as judgment of stereotyped expectations (i.e., superiority) about speakers. The more academically prepared the rater is, the more analytical and stringent he/she can become, especially in social perceptions of NNSs' superiority.

Fourth, raters of NNS oral proficiency may or may not vary dramatically among themselves in the amount of contact they have with NNSs. This factor of familiarity with NNSs' varieties needs further investigation for the effects on oral rating outcomes.

Fifth, student raters' previous experience in classes taught by NNSs can substantially affect their speech rating procedures. Negative experience can make raters much severe and critical in ratings of NNSs' oral proficiency. Test administrators may wish to screen candidates for rating jobs to eliminate those who report that grades in a class have been hurt by an ITA.

Sixth, practices for effective intercultural contact (e.g., collaborative problem solving) as a training can mitigate negative attitudes toward NNSs. No delayed posttest was run on study outcome, however, so it is unknown whether that prejudice reduction was just short term and just limited to this particular oral rating task.

Seventh, raters who hold negatively stereotyped expectations about NNSs' superiority tend to find NNSs' speech more heavily accented. That is, once raters construct negative stereotyped expectations about NNSs' social attractiveness, they tend to extend their biases to judgments of other NNSs and find specific NNSs less attractive socially.

Eighth, elements of acoustic fluency such as speech rate, syllable stress, and pitch range should be emphasized in the development of speech assessment criteria. Raters should be trained to heighten their sensitivity to these aspects of NNSs' speech.

Finally, even though frequency, location, and length of pauses are trait-relevant properties in NNSs' speech evaluation, raters should consider that NNSs speech involves individual characteristics and contextual information rather than just the level of English proficiency.

Moreover, the present research has important implications for applied linguists such as English-as-Second-Language (ESL) teachers. Acoustical differences (pauses, speech rates, stress, and intonation) between NNSs and NSs, analyzing the ITAs speech samples, can be found to be directly connected to accentedness; hence, this information can be used in ESL and accent-modification programs. That is, ESL/pronunciation programs can profitably focus on teaching and practicing prosodic aspects of speech (intonation, stress, rhythm, rate, and volume). Programs can teach NNSs to reduce pause duration, regulate the speech rate, make durational distinctions between stressed and unstressed words, and vary intonation of their speech, so that NNSs' production can be perceived as more native-like in their speech. Attending to these prosodic features will make a great improvement in perceived oral proficiency, and even outcomes such as perceived teaching competence.

Finally, because ITAs' in-class presentations were used for speech samples and because an inter-cultural intervention did positively affect ratings of oral performances, the findings of this study can be of interest to both ITAs and U.S. undergraduates. There is an urgent need to assess ITAs oral English proficiency and instructional ability due to the increasing number of international and non-native speakers of English in the teaching force (Noor, 1995). Effective assessment of ITAs' linguistic and instructional competence should occur with a variety of raters (Williams, 2006). In this study, undergraduates at universities served as raters because they were the intended audience and natural evaluators for ITA English discourse. Once the ITA learns the signficance of student ratings, undergraduates should be incorporated into the training process so that the ITA becomes comfortable interacting with undergraduates (Williams, 2006). Thus, the involvement of undergraduates and ITAs in this study mutually benefits both the ITA and the students as the two seek to gain a better understanding of the other's frame of reference.

Ultimately, this study can be a model of research that provides opportunities to improve undergraduates' comprehension of World Englishes. The responsibility for effective communication between native speakers and nonnative speakers lies not only with the latter as speakers, but also with the former as active, responsive, and empathic listeners.

Limitations and Recommendations for Further Research

The primary goal of this study was to find out the impact of rater background and attitudinal factors as well as acoustic properties of NNSs' speech on ratings of L2 oral performances. Even though numerous implications can be drawn from the findings of this study listed earlier, the present study has the following limitations to be considered.

One of the major limitations of this study is that it used only 11 ITAs with approximately 4-5 minutes of in-class presentations as speech samples. The present study can be certainly expanded by analyzing speech samples from a larger number of ITAs using a longer series of presentations or actual in-class lectures. A similar limitation regarding the sample size involves raters. If sample sizes were adequate both for raters and speakers, the study could be designed to explore the impact of rater background and attitudinal factors separately each for NS raters and NNSs. As in the results of the ETS study as well as those in this study, variation in rater native language status has been established as a major factor in oral proficiency ratings. Then, the variable, the amount of contact with NNSs showed a negative relationship with the oral proficiency ratings wherein the effect might be affected by the involvement of NNSs. In fact, one third of raters were composed of NNSs who tended to be generally more stringent than NSs in ratings of NNSs' oral performances. Therefore, it would be interesting to find out the difference

of the impact of rater background characteristics, separated by native language status, on oral assessment.

The speech sample selected for this study creates more restrictions on this study. Due to the nature of the speech sample involving various contents of international teaching assistants' (ITAs) teaching presentations, nonsystematic variance could be introduced. Unlike iBT TOEFL speech samples which consisted of examinees' responses to officially calibrated iBT TOEFL® speaking tasks, the present study used course-required presentations with a high potential of variability in terms of the nature of the content. Even though there was a control over the segment selection, i.e, only narrative description included, eliminating any interactive parts of the presentation, the contextual components of each speech performance could vary.

These speech sample limitations can be further related to the interpretation of the findings of this study. There may be restrictions on generalizing the results of the study due to the speech samples selected for acoustic analysis. The acoustic analyses of speech performances were only completed with male ITAs. Male speakers were selected only in order to avoid compounding effects generated by gender such as vocal pitch differences (Wennerstrom, 2001). Therefore, if female speakers (or both male and female speakers) were used, the results of the cluster analysis of suprasegmental variables reported here would be different. Next, the speech samples to be rated belong to ITAs who had relatively high levels of English proficiency. Results might not generalize to ratings of lower level English language learners (ELLs). Therefore, research is recommended where larger number of speech performance samples at different level of language proficiency included.

Another main limitation is associated with the acoustic speech analysis. The present study used only suprasegmental features as indices of linguistic characteristics. However, further

research can be developed to investigate the contribution of other linguistic elements of oral proficiency to rating scores such as grammatical accuracy and lexical uses in NNSs' oral performances. Very recently, for example, Iwashit et al., (2008) investigated the relationship between the features of the spoken language (grammatical accuracy and complexity, vocabulary, pronunciation, and fluency) produced by test-takers and holistic scores awarded by raters to these performances. They found that features from each category helped distinguish overall levels of performance, with particular features of vocabulary and fluency having the strongest impact. Therefore, to better predict oral proficiency ratings, an extensive comparison study between the factors of rater background and attitudes and the factors of other linguistic components of oral proficiency can be conducted. Also, non-suprasegmental parts of pronunciation can be included for further analysis to see relative contributions of prosodic and non-prosodic features of speech to ratings of oral proficiency.

The acoustic analysis limitation can be expanded to the issues of measurement reliability. The suprasegmental measures were completed by one single analysist, the researcher, as an experienced phonetician and applied linguist. This single-person measurement can be related to the individual subjectivity, even with the objectively measured instrumentation analysis. In the future study, reliability of the acoustic measures itself can be improved, by computing inter-rater reliability with two or three analysists involved in the acoustic measures.

Moreover, it is important to note a limitation related to the inter-cultural training intervention. The training effects (i.e., the socio-psychological intervention) require a careful interpretation. In this study, the training group selected showed relatively low rating scores at time 1 than the non-training group at time 1. In other words, members of the trained group were harsher than those of the untrained group in the two outcome ratings (instructional competence

and comprehensibility ratings) at time 1. Therefore, it can be questionable if the increase in the rating group from time 1 to time 2 was due to the true rating effect or due to the regression to the mean. Because the training group consisted of relatively harsh raters, changes in rating scores from time 1 to time 2 might have appeared drastic. As a result, the training method needs to be replicated with a complete random selection to confirm the effects of intercultural intervention in ITAs speech ratings.

In addition, as for the accurate interpretation of the training effects, experimental mortality should be taken into consideration by looking at 7 raters who had dropped at time 2 from time 1. The results of t-test, comparing mean differences between trained group and untrained group at time 1 separately for 70 entire raters and 63 raters (excluding 7 dropouts who did not complete the time 2 ratings) each, showed that experimental mortality apparently did not affect the interaction that indicated a training effect on instructional competence rating. But mortality from the non-trained group following pretest might be one explanation for the interaction on comprehensibility rating. Therefore, it does require that the conclusion with respect to comprehensibility be considered tentative at this time.

Further, limitations of this study can arise from the overall research resign itself. Because the task was artificial and low stakes, raters might have moderately little motivation to rate the speakers. Besides, the researcher of the present study utilized a rating instrument called 'Composite Speech Evaluation Instrument', which included analytic linguistic 7-point scales, to rate the samples of L2 oral performances. The findings might be different if other scales and rating instruments were selected.

Finally, the statistical procedures applied in this study require further research. The study attempted to conduct an integrative statistical analysis, as an alternative method, including

both the trait-relevant pronunciation variables and the trait-irrelevant rater background and attitudinal variables in a single regression. A random coefficient mixed modeling was computed by using the SAS PROC MIXED model analysis. This model was believed to be appropriate for this study because there were two crossed factors (i.e., raters and speakers), rather than one nested in the other, the data structure of this study had "covariates" on both the speakers and raters (the attributes). Although the analysis was completed after extensive consultation with expert statisticians in the Mixed Modeling analysis, the interpretation of the results of this analysis may still remain in questions due to the unprecedented nature of data which were widely crossed.

In fact, the random mixed modeling usually requires a larger sample size than the one in the current study. Multi-level logistic regression models use a big sample size because when the sample size is small there may not be sufficient variation to estimate a random effect, thus leading to non-convergence (Moineddin, Matheson, & Glazier, 2007). Moineddin et al. suggest sample sizes with at least a minimum of 100 groups and 50 individuals per group. Therefore, further research is invited for the appropriateness of the Mixed Modeling analysis for this type of study by using a bigger sample size.

REFERENCE

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

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*, 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.

Amichai-Hamburger, Y., & McKenna, K. Y. A. (2006). The contact hypothesis reconsidered: Interacting via the Internet. *Journal of Computer-Mediated Communication, 11*(3), article 7. http://jcmc.indiana.edu/vol11/issue3/amichai-hamburger.html

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

Bachman, L. F. (1997). Generalizability theory. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7; pp. 255-262). Dordrecht: Kluwer Academic Publishers.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing practice*. Cambridge: Cambridge University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-247.

Bailey, K. M. (1985). If I had known what I know now: Performance testing of foreign teaching

assistants. In P. Hauptman,, R. LeBlanc, & M.B. Wesche (Eds.), *Second language performance testing* (pp. 153-172). Ottawa: University of Ottawa Press.

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7), pp. 275-287. Dordrecht: Kluwer Academic Publishers.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267-285.

Barnwell, D. (1989). Naïve native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*, 152-163.

Bauer, G. (1996). Addressing special considerations when working with international teaching assistants. In J. D. Nyquist & D. H. Wulff (Eds.), *Working effectively with graduate assistants* (pp. 84-103). Thousand Oaks, CA: Sage.

Bejar, I. I. (1985). *A preliminary study of raters for the test of spoken English* (TOEFL Research Report No. 18). Princeton, JN: Educational Testing Service.

Bent, T., & Bradlow, A. (2003). The interlanguage speech intelligibility benefit. *Journal of The Acoustical Society of America, 114*, 1600-1610.

Bond, Z. (1999). *Slips of the ear: Errors in the perception of casual conversation*. San Diego, CA: Academic Press.

Boersma, P. & Weenink, D. (2007), PRAAT for Windows, 2007. http://www.fon.hum.uva.nl/praat/download_win.html

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral dicussion task. *Language Testing, 20*(1), 89-110.

Boulet, J. R., van Zanten, M., McKinley, D. W., & Gary, N. E. (2001). Evaluating the spoken

English proficiency of graduates of foreign medical schools. *Medical Education, 35*, 767-773.

Boztli, İ.(2006*). Academic oral presentation skills of instructors: Perceptions of the final project presentation rating scale used in the DML at METU*. (Publication no. http://www.metu.edu.tr/home/wwwmld/newsletter4.0/index.htm). Retrieved October 2, 2006.

Bradac, J. J., Cargile, A. C., & Hallett, J. S. (2001). Language attitudes: Retrospect, conspect, and prospect. In W. P. Robinson and H. Giles, (Eds.), *the new handbook of language and social psychology* (pp. 137-158). Chichester, England: John Wiley.

Brazil, D. (1997). *The communicative value of intonation in English*: Cambridge University Press.

Brennan, E., & Brennan, J. (1981). Accent scaling and language attitudes: Reactions to Mexican-American English speech. *Language and Speech, 24*, 207-221.

Bresnahan, M. I., & Kim, M. S. (1993). Factors of receptivity and resistance toward international teaching assistants, *Journal of Asian Pacific Communication, 4*, 1-12.

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

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.) *Current developments in language testing* (pp. 139-164). Singapore: Regional Language Centre.

Briggs, S. (1994). Using performance assessment methods to screen ITAs. In C. Madden & C. L. Myers (Eds.), *Discourse and performance of international teaching assistants*. (pp. 63-80). Alexandria, VA: TESOL Publications.

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

Brown, A., & Hill, K. (2007). Interviewer style and candidate performance in the IELTS oral interview. In L. Taylor and P. Falvery. (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 37-57). Cambridge: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English for Academic Purposes speaking tasks* (ETS Research Reports RR-05-05). Princeton, NJ: Educational Testing Service.

Brown, G. (1977). *Listening to spoken English*. London: Longman.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.

Cargile, A. C. (2002). Speaker evaluation measures of language attitudes: Evidence of information-processing effects. *Language Awareness*, *11*, 178-191.

Canagarajah, A. S. (2006). TESOL at Forty: What Are the Issues? *TESOL Quarterly, 40*(1), 9-34.

Canale, M. (1988). The measurement of communicative competence. *Annual Review of Applied Linguistics, 8,* 67-84.

Chafe, W. (1980). The deployment of consciousness in the production of a narrative. In W.L. Chafe (Ed.) *The pear stories* (pp. 9-50), Norwood N.J.: Ablex.

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

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*, 389-391.

Chapelle, C. (1988). Field independence: a source of language test variation? *Language Testing,*

5, 62-82.

Chaudron, C. (1988). *Second language classrooms*. Cambridge: Cambridge University Press.

Civikly, J. M., & Muchisky, D. M. (1991). A collaborative approach to ITA training: The ITAs, faculty, TAs, undergraduate interns, and undergraduate students. In J. Nyquist, R. D. Abbott, D. H. Wulff, & J. Sprague (Eds.), *Preparing the professorate of tomorrow to teach: Selected readings in TA training* (pp. 356-360). Dubuque, IA: Kendall/Hung.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign accented English. *Journal of the Acoustical Society of America, 116*, 3647-3658

Cluver, A. D. (2000). Changing language attitudes: The stigmatizatoin of Khoekoegowap in Namibia. *Language Problems & Language Planning 24*(1), 77-100.

Cohen, A. D. (1987). On taking tests: What the students report. *Language Testing, 1*, 70-81.

Cohen, A. D. (1994). *Assessing language ability in the classroom*. Boston: Heinle & Heinle.

Cohen, A. D., & Reed, D. J. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, A. K. Hill, N. Iwashita, T. McNamara, & K. O'Loughlin. (Eds.), *Experimenting with uncertainty: Essays in honor of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cole, R. A., & Jakimik, J. (1978). Perceptibility of phonetic features in fluent speech. *Journal of The Acoustical Society of America, 64*(1), 44-56.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed method research.* Thousand Oaks, CA: Sage Publications.

Croker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL:

Holt, Rinehart and Winston, Inc.

Crystal, D. (2003). *English as a global language*. Cambridge: Cambridge University Press.

Crystal, D. (2003). *A dictionary of linguistics and phonetics*. Malden, MA: Blackwell Publishing.

Cutler, A., Dahan, D., & Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40,* 141–201.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Dailey-O'Cain, J. (2000) The sociolinguistic distribution of and attitudes toward focuser like and quotative like. *Journal of Sociolinguistic*s, 4, 60–80.

Damron, J. (2003). What's the problem? A new perspective on ITA communication. *Journal of Graduate Teaching Assistant Development, 9*, 81-88.

Davies, C. E. (2003). How English-learners joke with native speakers: An interactional sociolinguistic perspective on humor as collaborative discourse across cultures. *Journal of Pragmatics, 35*, 1361-1385.

Derwing, T. M. (1990). Speech rate is no simple matter. *Studies in Second Language Acquisition, 12*(3), 303-313.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1-16.

Derwing, T., & Munro, M. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics 22*: 324-227.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*, 379-397

Derwing, T. M., & Rossiter, M. J. (2003). The Effects of Pronunciation Instruction on the

Accuracy,Fluency, and Complexity of L2 Accented Speech. *Applied Language Learning, 13*(1), 1-17.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Testing, 54*(4), 655-679.

Deterding, D. & Kirkpatrick, A. (2006). Intelligibility and an emerging ASEAN English lingua franca , *World Englishes 25*, 391-410.

Douglas, D. (2001). Language for Specific Purposes assessment criteria: Where do they come from? *Language Testing, 18*(1), 171-185.

Dunkel, P. (1991). Listening in the native and second/foreign language: Toward an integration of research and practice. *TESOL Quarterly, 25*, 431-457.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197-221.

Edwards, J. R. (1982). Language attitudes and their implications among English speakers. In E. B. Ryan and H. Giles (eds.), *Attitudes towards language variation* (pp.20-33). London: Edward Arnold.

Ejzenberg, R., 2000. The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–314). Ann Arbor: The University of Michigan Press,.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing, 18*(1), 149-170.

Engber, C. (1987). Summary of the discussion session. In A. Valdman (Ed), *Proceedings of the symposium on the evaluation of foreign language proficiency* (pp. 254-268). Bloomington, IN: Committee for Research and Development in Language Instruction.

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis* (pp. 261-287). Mahwah, NJ: Erlbaum.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL composition* (ETS Research Reports RR-03-17). Princeton, JN: Educational Testing Service.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.

Field, J. (2003). The fuzzy notion of 'intelligibility': A headache for pronunciation teachers and oral testers. *IATEFL Special Interest Groups Newsletter,* 34-38.

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*, 399-423.

Fitch, F., & Morgan, S. E. (2003). "Not a lick of English": Constructing the ITA identity through student narratives. *Communication Education, 52*, 297-310.

Flege, J. E. (1991). Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. *Journal of Acoustical Society of America, 89*(1), 395-411.

Flege, J. E., Munro, M. J., & Fox, R. A. (1994). Auditory and categorical effects on cross-language vowel perception. *Journal of the Acoustical Society of America, 95(6)*: 3623-3641

Flege, J. E., & Port, R. (1981). Cross-language phonetic inferences: Arabic to English. *Language and Speech, 24*, 125-146.

Fulcher, G., 1996. Does thick description lead to smart tests? A data-based approach to rating

scale construction. *Language Testing 13*, 208–238.

Forbes, H. D. (2004). Ethnic conflict and the contact hypothesis. In Y-Tg Lee, C. McCauley, and F. Moghaddam (Eds), T*he psychology of ethnic and cultural conflict* (*pp.* 390-411). Portsmouth, NH: Greenwood Publisching Group.

Fox, W. S., & Gay, G. (1994). Functions and effects of international teaching assistants. *The Review of Higher Education, 18*(1), 1-24.

Freed, B.F., 2000. Is fluency, like beauty, the eyes, of the beholder? In: Riggenbach, H. (Ed.), *Perspectives on fluency* (pp. 243–265). Michigan: the University of Michigan Press.

Fulcher, G. (1999) Ethics in language testing, *TAESIG Newsletter* 1 (1), 1-4. Retrieved March 10, 2007 from http://taesig.8m.com/news1.html

Fulcher, G. (2003). *Testing second language speaking.* Harlow: Person.

Gallego, J.C. (1990). The intelligibility of three non-native English speaking teaching assistants: An analysis of student reported communication. *Issues in Applied Linguistics*, *1*, (2), 219-237.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, *64*, 428-433.

Gardner, R.C., & Lambert, W.E. 1972. *Attitudes and motivation in second language learning*. Rowley, MA: Newbury House.

Gass, S. M., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34*, 65-89.

Gayle, G. (1980). *A descriptive analysis of second language teaching styles in the oral approach.* Ottawa: University of Ottawa Press.

Grant, L. (2001). *Well said: Pronunciation for clear communication*. Boston: Heinle & Heinle.

Griffiths, R. (1992). Speech rate and listening comprehension: Further evidence of the relationship. *TESOL Quarterly, 26*, 385-391.

Grover, C., Jamieson, D., & Dobrovolsky, M. (1987). Intonation in English, French, and German: Perception and production. *Language and Speech*, *30,* 277-295.

Gudykunst, W. B. (1991). *Bridging differences: Effective intergroup communication*. Newbury Park, CA: Sage.

Hadden, B. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, *41*, 1-24.

Hale, G. A. (1988). Student major field and text context: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5*(1), 49-61.

Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *The Modern Language Journal*, *79*, 223-234.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*, 579-591.

Hamp-Lyons, L., & Davies, A. (2008). The English of English tests: Bias revisited. *World Englishes, 27* (1), 27-39.

Hirvonen, P. (1967) *On the problems met by Finnish students in learning the rising interrogative intonation in English*. Department of Phonetics: University of Turku.

Hogg, R., & McGully, C. B. (1987). *Metrical phonology: A coursebook*. Cambridge, England: Cambridge University Press.

Hocking, J. E., Stacks, D. W., & McDermott, S. T. (2003). *Communication research.* Boston, M.A.: Pearson Education, Inc.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Hamp-Lyons, L. (2002). 'An Interview with Liz Hamp-Lyons', by T. Newfields, *JALT Testing & Evaluation SIG Newsletter*, Vol. 6 (2), 3-4. Retrieved March 10, 2002 from http://www.jalt.org/test/ham_new.htm

Inglis, M. (1993). The communicator style measure applied to nonnative speaking teaching assistants. *International Journal of Intercultural Relations*, *17*, 89-105.

Ingram, J.C.L., & Park, S-G. (1997) Cross-language vowel perception and production by Japanese and Korean speakers of English. *Journal of Phonetics, 25*, 343-370.

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

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach. *Language Learning, 51*(3), 401-436.

Jenkins, J. (2006). The spread of EIL: A testing time for testers. *English Language Teaching Journal*, *60* (1), 42-50.

Jenkins, J. J., Strange, W., & Polka, L. (1995). Not everyone can tell a "rock" from a "lock": Assessing individual differences in speech perception. In D. Lubinski and R. V. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods, and findings* (pp. 297-325), Palo Alto, Davies-Black Publishing.

Jenkins, S. (2000). Cultural and linguistic miscues: A case study of international teaching assistant and academic faculty miscommunication. *International Journal of Intercultural Relations, 24*(4), 477.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of non-verbal, paralinguistic and verbal behaviors in assessment decisions. *The Modern Language Journal, 67*, 90-107.

Juffs, A. (1990). Tone, syllable structure and interlanguage phonology: Chinese learners' stress errors. International Review of Applied Linguistics, 28, 99-117.

Kachru, B. B. (1997). Past imperfect: The other side of English in Asia. In L. E. Smith, and M. L. Forman (Eds.), *World Englishes 2000: Resources for research and teaching* (pp. 209-251). Hawaii: University of Hawaii.

Kang, O. (2008). The effect of rater background characteristics on the rating of International Teaching Assistants Speaking Proficiency. *Spaan Fellow Working Paper*, 6.

Kang, O., Rubin, D., & Pickering, L. (under review). Judgments of ELL proficiency in oral English and acoustic measures of accentedness. *Modern Language Journal.*

Kelch, K. (1985). Modified input as an aid to comprehension. . *Studies in Second Language Acquisition, 7*, 81-89.

Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart and Winston, Inc.

Kim, H.-J. (2005). *World Englishes and language testing: The influence of rater variability in the assessment process of English language oral proficiency.* Doctoral Dissertation, University of Iowa, 2005.

Kim, H.-J. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers College Working Papers in TESOL and Applied Linguistics, 6*(1), 1-37.

Kirkpatrick, A. (2007). *World Englishes*. Cambridge: Cambridge University Press.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of

second language learners. *System, 32*, 145-164.

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

Kreft, I., & deLeeuw, J. (2006). *Introduction to multilevel modeling*. London: Sage Publications.

Kunnan, A. J. (Ed.)(1998). Special issue: structural equation modeling. *Language Testing, 15*(3).

Lantolf, J. P., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis *Modern Language Journal*, 69 (4), 337-345.

Lambert, W. E., Hodgson, R.C., Gardner, R.C., & Fillenbaum, S. (1960). Evaluational reactions to spoken language. *Journal of Abnormal and Social Psychology,* 60, 44-51.

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32, 505-524.

Linacre, J. M. (1989/1993*). Many-facet Rasch measurement and the challenges to measurement*. Unpublished manuscript. Department of Education, University of Chicago.

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

Lindemann, S. (2000). *Non-native speaker "incompetence" as a construction of the native listener: Attitudes and their relationship to perception and comprehension of Korean-accented English*. Unpublished Doctoral Dissertation, the University of Michigan, 2000.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States. *Language in Society, 31*, 419-441.

Lindemann, S. (2003). Koreans, Chinese, or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistic*s, 7, 3, 348-364.

Lindemann, S. (2005). Who speaks "broken English"? US undergraduates' perceptions of non-native English, *International Journal of Applied Linguistics, 15* (2). 187-213.

Lippi-Green, R. (1997). *English with an Accent: Language, ideology, and discrimination in the United States*. New York, Routledge.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., & Schabenberger, O. (2007). *SAS for mixed models*. Cary, NC: SAS Institute, inc.

Llurda, E. (2000). Effects of intelligibility and speaking rate on judgments of non-native speakers' personalities. *International Review of Applied Linguistics*, *38*, 288-299.

Luhman, R. (1990) Appalachian English stereotypes: Language attitudes in Kentucky. *Language in Society* 19, 331–348.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testin*g, *12*, 54–71.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18* (4), 351-372.

Lynch, B. K., & Davidson, F. (1997). Crierion-referenced language test development: Linking curricula, teachers and tests. *TESOLl Quarterly, 28*, 727-743.

Lynch, B. K., & McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics, 26*(4), 381-400.

Major, R., Fitzmaurice, S., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly 36,* 173-190.

Mattran, K. J. (1997). Native speaker reactions to speakers of ESL: Implications for adult basic education oral English proficiency testing. *TESOL Quarterly*, 11, 407-414.

McKay, S. (2002). *Teaching English as an international language: rethinking goals and approaches*. Oxford: Oxford University Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264: 746-748.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52-75.

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

McNamara, T. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18,* 446-466.

McNamara, T. (2002). Language testing and social policy. The message of the Shibboleth', Retrieved March 11, 2004 from http://bear.soe.berkeley.edu/measurement/docs/CommentaryPorterMcNamara.pdf

McNamara, T. F., & Adams, R. J. (1991). Exploring rater characteristics with Rasch techniques *Paper presented at the 13th annual Language Testing Research Colloquium*, Princeton, NJ: Educational Testing Service.

McNamara, T. F., & Adams, R. J. (1994). Exploring rater characteristics with Rasch techniques. [ERIC Document Reproduction Service ED 345 498]. In *Selected papers of the 13th Language Testing Research Colloquium (LTRC)*. Princeton, New Jersey: Educational Testing Service.

Mennen, I. (1998). Second language acquisition of intonation: the case of peak alignment, *Chicago Linguistic Society, 34,* 327- 341.

Miller, N. (2002). Personalization and the promise of contact theory. *Journal*

of Social Issues, 58(2), 387–410.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol*. 16, 7-34.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition, 17*, 17-34.

Munro, M. J. (1998). The effects of noise on the intelligibility of forieng-accented speech. *Studies in Second Language Acquisition, 20*, 198-154.

Munro, M. T. D., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, v. 45*, 73-97.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*(no. 2), 159-182.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system.* (ETS Research Reports RR-00-06). Princeton, NJ: Educational Testing Service.

Nakajima, S., & Allen, J. (1993). A study on prosody and discourse structure in cooperative dialogues. *Phonetica*, *50,* 197-210.

Nakamura, Y. (1995). Making speaking tests valid: Practical consideration in a classroom setting. In J.D. Brown & S. O. Yamashita (Eds.). *Language Testing in Japan* (pp. 136-148). Tokyo, Japan: The Japan Association for Language Teaching.

Naslund, D. T. (1993). The /s/ phoneme: A gender issue. Unpublished manuscript, University of Minnesota, Duluth.

Nisbett, R.E., & Wilson, T.D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality & Social Psychology, 35*, 250-256.

Nguyen, B. B.-D. (1993). Accent discrimination and the test of spoken English: a call for an objective assessment of the comprehensibility of nonnative speaker. *Asian Law Journal, 81*, 1324-1361.

Noor, Z. M. (1995). *Measures of spoken assessment in ESL: A comparison of three measures of assessing the oral proficiency of international teaching assistants.* Doctoral Dissertation, the State University of New York at Buffalo, 1995.

Norton, B., & Toohey, K. (2004). *Critical pedagogies and language learning.* Cambridge: Cambridge University Press

Oller, J. W. (1976). Evidence of a general language proficiency factor: an expectancy grammar. *Die neuren Sprachen, 76*, 165-174.

Pae, T.-i. (2001). International teaching assistant programs and world Englishes perspectives. *Journal of Graduate Teaching Assistant Development, 8*(2), 71-77.

Pawley, A., & Syder, F. (2000). *The one-clause-at-a-time hypothesis*. In: Heidi Riggenbach (ed.). Perspectives on fluency. Ann Arbor: University of Michigan Press, 163-199.

Pica, T., Barnes, G. A. & Finger, A. G. (1990*). Teaching matters: Skills and strategies for international teaching assistants*. New York: Newbury House.

Piché, G.L., Michlin, M.L., Rubin, D.L., & Sullivan, A. (1977). Effects of dialect-ethnicity, social class, and quality of written compositions on teachers' subjective evaluations of children. *Communication Monographs*, *44*, 60-72.

Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants.* Unpublished Doctoral Dissertation.

University of Florida, 1999.

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

Pickering, L. (2002). Patterns of Intonation in cross-cultural communication: Exchanging structure in NS TA and ITA classroom interaction. *Crossroads of Language, Interaction, and Cult*ure, 4, 1-17.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19.

Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistic*s, *25*, 219-233.

Pickering, L., & Wiltshire, C. (2000). Pitch accent in Indian-English teaching discourse. *World Englishes, 19* (2), 173-184.

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

Polio, C. & Gass, S. (1998). The role of interaction in native speaker comprehension of nonnative speech. *The Modern Language Journal, 82*,(3), 308-319**.**

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, 4, 72-92.

Pollitt, A., & Murray, N. L. (1995). What raters really pay attention to? In M. Milanovic & Saville (Eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 74-91). Cambridge, England: Cambridge University Press

Powers, D. E., Schedl, M. A., Wilson Leung, S. W., & Butler, F. A. (1999). Validating the

revised Test of Spoken English against a criterion of communicative success. *Language Testing, 16*(4), 339-425.

Purnell, T., Idsardi W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology* 18(1), 10-30.

Rao, V., & Woolcock, M. (2003). Integrating qualitative and quantitative approaches in program evaluation. In F. J. Bourguignon & L. P. da Silva (Eds.), *The impact of economic policies on poverty and income distribution: Evaluation techniques and tools* (pp. 165-190). New York: The World Bank and Oxford University Press.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition, 23*, 497-526.

Riggenbach, H., 1991. Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes, 14*, 423–441.

Riniolo, T.C., Johnson, K.C., Sherman, T.R. & Misso, J.A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology*, 133 (1), 19-35.

Ross, S. (2001). Program-defining evaluation in a decade of eclecticism. In J. C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 167-196). Cambridge: Cambridge University Press.

Rounds, P. (1987). Characterizing successful classroom discourse for NNS teaching assistant

training. *TESOL Quarterly, 21*, 643-672.

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduate's judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Rubin, D. L. (1993). The other half of international teaching assistant training: Classroom communication workshops for international students. *Innovative Higher Education, 17*, 183-193.

Rubin, D.L. (2002). Help! My professor (or doctor or boss) doesn't talk English! In J. Martin, T. Nakayama, and L. Flores (Eds.), *Readings in intercultural communication: Experiences and contexts* (pp. 127-137). Boston: McGraw Hill.

Rubin, D.L., & Lannutti, P. (2001). Frameworks for assessing contact as a tool for reducing prejudice. In V.H. Milhouse, M.K. Asante , and P.O. Nwosu (Eds.), *Transcultural realities: Interdisciplinary perspectives on cross-cultural relations* (pp. 313-326). Beverly Hills : Sage.

Rubin, D., Kang, O., & Pickering, L. (under review). *Relative impact of rater intercultural and language background, rater attitudes, rater training, and measurable elements of pronunciation on TOEFL iBT speaking proficiency scoring*. (ETS Research Reports*)* Princeton, NJ: Educational Testing Service.

Rubin, D.L., & Lannutti, P. (2001). Frameworks for assessing contact as a tool for reducing prejudice. In V.H. Milhouse, M.K. Asante, and P.O. Nwosu (Eds.), *Transcultural realities: Interdisciplinary perspectives on cross-cultural relations* (pp. 313-326). Beverly Hills: Sage.

Rubin, D.L., & Smith, K. (l990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of non-native English speaking teaching assistants. *International Journal of Intercultural Relations, 14*, 337-353.

Ryan, E. B., Carranza, M. A., & Moffie, R. W. (1977). Reactions toward varying degrees of accentedness in speech of Spanish-English. *Language and Speech, 20*(3), 267-273.

Ryan, E. B., & Sebastian, R. J. (1980). The effects of speech style and social class background on social judgments of speakers. *British Journal of Social and Clinical Psychology*, *19,* 229-233.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110(4)*, 474-494.

Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly*, 22, 69-90.

Sarkisian, E., & Maurer, W. (1998). International TA and beyond: Out of the program and into the classroom. In M. Marincovich, J. Prostko, and F. Stout, (Eds.), *the professional development of graduate teaching assistants* (pp. 163-180). Boston: Anker Press.

Sarwark, S., Smith, J., MacCallum, R., & Cascallar, E.C. (1995). *A study of characteristics of the SPEAK Test. (*Research Report RR 94-47*.)* Princeton, N.J., Educational Testing Service.

Schmid, P. M., & Yeni-Komshian, G. H. (1999). The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language and Hearing Research, 42*, 56-64.

Schmidt, A. M., & Flege, J. E. (1996). Speaking rate effects on stops produced by Spanish and English monoliguals and Spanish/Enlgish bilinguials. *Phonetica, 53*(162-179).

Schuetze-Coburn, S., Shapley, M., & Weber, E. G. (1991). Units of intonation in discourse: a comparison of acoustic and auditory analyses. *Language And Speech, 34 (Pt 3)*, 207-234.

Seligman, C. R., G. R. Tucker, and W. E. Lambert (1972). The effects of speech style and other attributes on teachers' attitudes toward pupils. *Language in Society* 1: 131-142.

Shah, A. P. (2002). *Temporal characteristics of Spanish-accented English: Acoustic measures and their correlation with accentedness ratings.* Doctoral Dissertation, the City University of New York, 2002.

Shah, A. P. (2004). Production and perception correlates of Spanish-accented English. *From Sound to Sens*e, June 11-June 13, MIT. Retrieved October 02, 2007, from http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Friday%20Posters/FA-Shah-STS.pdf.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Newbury Park, CA: Sage.

Sherif, M. (1966). *Group conflict and co-operation: Their social psychology.* London: Routledge and Kegan Paul.

Shohamy, E. (1983). The stability of oral proficiency assessments on the oral interview testing procedure. *Language Learning*, *33*, 527-39.

Shohamy, E. (1993). The power of tests: the impact of language tests on teaching and learning, *NFLC Occasional Papers*. Washington, DC: National Foreign Language Center.

Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Longman.

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

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323-355.

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task based learning. *Language Teaching Research*, 1(3), 185-211.

Skehan, P. (1998). A cognitive approach to language learning. Oxford: Oxford University Press.

Smith, L., & Nelson, C. (1985). International intelligibility of English: Directions and resources. *World Englishes, 4,* 333-342.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (pp. 75–90). Urbana: University of Illinois Press.

Smith, R.A., Strom, R.E. & Muthuswamy, N. (2005). Undergraduates' rating of domestic and international teaching assistants: Timing of data collection and communication intervention. *Journal of Intercultural Communication and Research, 34* (1), 3-21.

Spolsky, B. (1978). *Approaches to language testing. (*Advances in Language Testing Series 2) Arlington, VA: Center for Applied Linguistics.

Spolsky, B., 1995. *Measured words: The development of objective language testing*, Oxford: Oxford University Press.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*, 347-364.

Stockburger, D. W. (1998). Multivariate statistics: concepts, models, and applications. WWW version 1.0, Missouri State University. Retrieved April 07, 2008, from http://www.psychstat.missouristate.edu/multibook/mlt00.htm,.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology 18*(1), 86-100.

Strand, E. A., & Johnson, K.. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference* (pp. 14-26). Berlin: Mouton.

Stuart, A., Ord, K., & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics* **2A**. London: Arnold.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech, 37*(21-43).

Taylor, L. (2006). The changing landscape of English: Implications for language assessment. *English Language Teaching Journa*l, 60 (1), 51-60.

Templer, B. (2004). High-stakes testing at high fees: Notes and queries on the international English proficiency assessment market. *Journal for Critical Education Policy Studies,* 2 (1). Retrieved August 22, 2005, from www.jceps.com/index.php?pageID=article&articleID=21.

Tench. P. (1996). *The intonation systems of English*. London: Cassell.

Thompson, W.F. (2003). Is music unique among human activities? *Bulletin of Psychology and the Arts*, *4*(1), 38-39.

Toivanen, J. (2001). *Perspectives on Intonation: English, Finnish and English Spoken by Finnish.* Frankfurt Main: Peter Lang.

Towell, R., Hawkins, R., Bazergui, N., 1996. The development of fluency in advanced learners of French.applied linguistics 17, 84–119.

Thompson, I. (1991). Foreign accents revisited: the English pronunciation of Russian immigrants, *Language Learning*, 41, 177-204.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing*, 16(1), 82-111.

van Wijngaarden, S.; Steeneken, H.; & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for nonnative listeners. *Journal of the Acoustical Society of America, 111*, 1906-1916.

Vaissiere, J. (1995). Phonetic explanation for cross-linguisitc prosodic similarities. *Phonetica, 52*, 123-130.

Vanderplank, R., 1993. Pacing and spacing as predictors of difficulty in speaking and understanding English. *English Language Teaching Journal, 47,* 117–125.

Voci, A. (2003). Intergroup contact and prejudice toward immigrants in Italy: The mediational role of anxiety and the moderational role of group salience. *Group Processes and Intergroup Relations, 6* (1), 37-54.

von Hippel, W., Sekaquaptewa, D., & Vargas, P. (1995). On the role of encoding processes in stereotype maintenance. In M. P. Sanna (Ed.), *Advances in Experimental Social Psychology* (pp. 177-254*),* San Diego, Academic Press.

Voss, B. (1984). *Slips of the ear: Investigation into the speech perception behavior of German speakers of English.* Tubingen: Narr.

Watt, D. (1997). *The phonology and semiology of intonation in English: An instrumental and systemic perspective.* Bloomington: Indiana University Linguistics Club Publications.

Weigle, S. (1994). Using FACETS to model rater training effects. Paper presented at the 16th Language Testing Research Colloquium, Washington, DC.

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

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

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks second language learning, teaching and testing* (pp. 186-209). Essex, UK: Longman.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of nonnative speakers. *Applied Linguistics* 15: 399-421.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in Second Language Acquisition,* 42, 1-13.

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

Wennerstrom, A. (2001). *The music of everyday speech*. New York: Oxford University Press.

Williams, G. (2006). *Cultural, professional and personal influences on the teaching identity development of international teaching assistants.* Unpublished Doctoral Dissertation, University of Georgia, 2006.

Williams, R. M. (1964). *Strangers next door: Ethnic relations in American communities.* Englewood Cliffs, NJ: Prentice-Hall.

Yano, Y. (2001). World Englishes in 2000 and beyond. *World English, 20*, 119-131.

Yook, E. (1999). An investigation of audience receptiveness to non-native teaching assistants.

*Journal of the Association for Communication Administratio*n, 28, 71-77.

Yook, E., & Albert, R. (1999). Perceptions of international teaching assistants: The interrelatedness of intercultural training, cognition, and emotion, *Communication Education*, *48,* 1-17.

Yule, G. (1980). Speakers' topics and major paratones. *Lingua, 52*, 33-47.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review, 57* (4), 573-590.

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The Speech Evaluation Instrument. *Journal of Language and Social Psychology*, *4*, 113-124.

Appendix A

Language Background Questionnaire

1. Your name (pseudonym):
2. Date of birth:
3. Gender: ○ Male ○Female
4. What is your native language (mother tongue)?
5. How many languages can you speak fluently?
6. Where were you born? (city, province, country)
7. How would you describe your ethnic or racial background?
8. What is the highest level of education achieved by your father?
   - ○ less than high school diploma
   - ○ high school graduation
   - ○ post high school technical certificate
   - ○ some college
   - ○ bachelors degree
   - ○ professional and graduate degree
   - ○ don't know
9. What is your major, or proposed major, field of study?
10. What is your present or intended minor, if any?
11. What is the year or the class of your program?
12. Your estimated overall GPA:
13. Your intended career:
14. Have you studied a foreign language?

If YES:

10a. Language(s) studied in high school

	Language(s) Taken	Highest level Course taken	Number of year studied	Final Grade (If applicable)
i				
ii				

10b. Language(s) studied in college

	Language(s) Taken	Highest level Course taken	Number of year studied	Final Grade (If applicable)
i				
ii				
iii				

10c. Language(s) studied at another place

	Language(s) Taken	Highest level Course taken	Number of year studied	Final Grade (If applicable)
i				
ii				

15. How would you evaluate your current foreign language skills, if any?

	Language(s)	Proficiency		
		Speaking	Listening	Writing
i				
ii				
iii				

16. Have you taken any formal college courses in Linguistics/Applied Linguistics/Second Language (L2)/ Grammar Study/ ESL (English as a Second Language)?

    If YES, please specify all the classes that you have taken.

17. Have you taught any courses such as Linguistics/Applied Linguistics/Second Language (L2)/ Grammar Study/ ESL classes?

    If YES, please specify the months/years, the level, and the site of your teaching experience.

18. Have you taught any ESL students either formally or informally?

    If YES, please specify the months/years, the level, the site, and the student number of your teaching experience.

19. Do you interact with non-native speakers of English (speakers whose first language is not English)?

If YES:

19a. Please indicate the degree of contact you have with nonnative speakers of English. (Please circle one number for each on the following rating scale)

	Very frequent/ Daily or almost Daily				Very infrequent Several times a year or less
A. Friends or social acquaintances ……	5	4	3	2	1
B. Colleagues or business acquaintances..	5	4	3	2	1
C. Teachers/Teaching assistants ……..	5	4	3	2	1
D. Other……..……..	5	4	3	2	1

Please specify-_____

19b. During a "typical" week, approximately how many nonnative speakers of English do you come in contact with?

19c. In a typical encounter, how long do you have contact with a nonnative speaker of English?

19c. Approximately how many hours per week do you spend communicating with non-native speakers of English?

19d. What languages do they speak?

20. Approximately what percentage of your current good friends are nonnative speaker of English? How many are they?

_____% _____

21. Approximately what percentage of your current daily acquaintances were nonnative speaker of English? How many are they?

_____% _____

22. In how many courses (including current courses, courses dropped, course completed) have you been instructed by a teacher who was not a native speaker of English?

23. On how many occasions—if any—have you decided not to enroll in a particular section of a course (or to drop or withdraw) because the instructor was not a native speaker of English?

24. On how many occasions – if any – do you feel that your final grade is a course was hurt because the instructor was not a native speaker of English?

25. Have you traveled or lived outside of the US?

If YES:

25a. In what country?

25b. How long (weeks/months/years)?

25c. How old were you when you returned to the US?

25d. Please narrate if you have any significant life-time exposures to non-native speakers of English.

26. Do you have normal hearing?

27. How good are you at understanding foreign accents? (Please circle one number for each on the following rating scale)

very poor				excellent
1	2	3	4	5

28. How would you describe your geographic location?

29. How would you describe your personality? (e.g., introvert, extrovert, etc)

30. Other comments ?

Appendix B

Measure of Speaker Comprehensibility

The speaker to whom I just listened ....

was easy to understand___/___/___/___/___/___/___ was hard to understand

was incomprehensible ___/___/___/___/___/___/___was highly comprehensible

was unclear___/___/___/___/___/___/___ was clear

required little effort to understand___/___/___/___/___/___/___required lots of effort to understand

made it simple to grasp the meaning___/___/___/___/___/___/___made it difficult to grasp the meaning

Appendix C-1

Linguistic Stereotyping Measures

C.1 Speech Evaluation Instrument (SEI)

Speech Evaluation Instrument

1. How would you rate the instructor you just heard?

Advantaged	___/___/___/___/___/___/___	Disadvantaged
Kind	___/___/___/___/___/___/___	Unkind
Lazy	___/___/___/___/___/___/___	Energetic
Poor (Assertive)	___/___/___/___/___/___/___	Rich (Unassertive)
Unclear	___/___/___/___/___/___/___	Clear
Complete	___/___/___/___/___/___/___	incomplete
White collar	___/___/___/___/___/___/___	Blue collar
Unsure	___/___/___/___/___/___/___	Confident
Intelligent	___/___/___/___/___/___/___	Unintelligent
Fluent	___/___/___/___/___/___/___	Influent
Attractive	___/___/___/___/___/___/___	Unattractive
Qualified	___/___/___/___/___/___/___	Unqualified
Unfriendly	___/___/___/___/___/___/___	Friendly
Disorganized	___/___/___/___/___/___/___	Organized
Cold	___/___/___/___/___/___/___	Warm
Uneducated	___/___/___/___/___/___/___	Educated
Unappealing	___/___/___/___/___/___/___	Appealing
Illiterate	___/___/___/___/___/___/___	Literate
Likeable	___/___/___/___/___/___/___	Likable

Lower class	___/___/___/___/___/___/___	Upper class
Passive	___/___/___/___/___/___/___	attractive
Bad	___/___/___/___/___/___/___	good
Person of color	___/___/___/___/___/___/___	Caucasian
Sour	___/___/___/___/___/___/___	Sweet
Effective teacher	___/___/___/___/___/___/___	Poor teacher
Experienced	___/___/___/___/___/___/___	Inexperienced
Speaks with foreign accent	___/___/___/___/___/___/___	Speaks with American accent
Talkative	___/___/___/___/___/___/___	Shy
Hostile	___/___/___/___/___/___/___	Good natured
Aggressive	___/___/___/___/___/___/___	Unaggressive
Nice	___/___/___/___/___/___/___	Awful
Enthusiastic	___/___/___/___/___/___/___	Hesitant
Strong	___/___/___/___/___/___/___	Weak
Considerate	___/___/___/___/___/___/___	Inconsiderate
Honest	___/___/___/___/___/___/___	Dishonest

How would you rate the lecture you just heard?

Formal	___/___/___/___/___/___/___	Informal
Boring	___/___/___/___/___/___/___	Interesting
Confusing	___/___/___/___/___/___/___	Clear
Easy to understand	___/___/___/___/___/___/___	Difficult to understand

Appendix C-2

Linguistic Stereotyping Measures

Sample Cloze Test

Galaxies and Clusters (1748 words—TEXT1)
Hydron Spinrad University of California, Berkley

Galaxies—vast collections of billions of stars—are the basic building blocks of the universe. These grand objects are so _____in size that they simply dwarf all _____ experience. The amount of light and other energies they give off defy any _____ at everyday comparison. Yet since we _____ that the existence of other galaxies in the 1920s, telescopes of increasing _____ and sophistication have shown us not just a ___, not hundreds or _____, but hundreds of billions of these grand star systems in _____ direction we look.

Galaxies also ____ out to be "gregarious"—they generally __ not appear to live alone. We ____ them gathered together in smaller groups or _____ clusters, and those groups and clusters _____ tend to join together in immense _____ astronomers call superclusters.

Furthermore, the _____ of galaxies is rapidly expanding on the _____ scale. All the galaxies are _____ form one another, so that in the _____ it has taken you to _____ this page, the distance to a _____ of galaxies has opened up by another full _____ miles! This expansion is a fundamental _____ of the universe which illuminates and illuminated by our study of galaxies.

Appendix C-3

Linguistic Stereotyping Measures

Sample Lecture script

Galaxies and Clusters (1748 words—TEXT1)
Hydron Spinrad University of California, Berkley

Galaxies—vast collections of billions of stars—are the basic building blocks of the universe. These grand objects are so enormous in size that they simply dwarf all human experience. The amount of light and other energies they give off defy any attempt at everyday comparison. Yet since we confirmed that the existence of other galaxies in the 1920s, telescopes of increasing power and sophistication have shown us not just a few, not hundreds or thousands, but hundreds of billions of these grand star systems in every direction we look.

Galaxies also turn out to be "gregarious"—they generally do not appear to live alone. We find them gathered together in smaller groups or larger clusters, and those groups and clusters themselves tend to join together in immense structures astronomers call superclusters. Figuer [6.3] is a photograph that I took at the Cerro Tololo Interamerican Observatory in Chile where a 4-meter (158—inch) diameter telescope site atop a peak in the Andes and affords astronomers beautifully clear views of the night sky. The photograph shows a rich cluster of galaxies unromantically called Str 0431-616 (a name derived from its position in the sky). The impressive collection of galaxies is visible only from the Earth's Southern Hemisphere by the way. As you look at the

Appendix D

Composite Speech Evaluation Instrument

I. How would you rate the instructor you just heard?

Advantaged	___/___/___/___/___/___/___	Disadvantaged
Kind	___/___/___/___/___/___/___	Unkind
Lazy	___/___/___/___/___/___/___	Energetic
Poor	___/___/___/___/___/___/___	Rich
Unclear	___/___/___/___/___/___/___	Clear
Assertive	___/___/___/___/___/___/___	Unassertive
Complete	___/___/___/___/___/___/___	incomplete
White collar	___/___/___/___/___/___/___	Blue collar
Unsure	___/___/___/___/___/___/___	Confident
Has native-like intonation	___/___/___/___/___/___/___	Has unfamiliar intonation
Fluent	___/___/___/___/___/___/___	Influent
Approachable	___/___/___/___/___/___/___	Unapproachable
Enthusiastic	___/___/___/___/___/___/___	Hesitant
Attractive	___/___/___/___/___/___/___	Unattractive
Easy to interact with students	___/___/___/___/___/___/___	Difficult to interact with students
Qualified	___/___/___/___/___/___/___	Unqualified
Unfriendly	___/___/___/___/___/___/___	Friendly
Disorganized	___/___/___/___/___/___/___	Organized
Cold	___/___/___/___/___/___/___	Warm
Self-confident	___/___/___/___/___/___/___	insecure

Uneducated	___/___/___/___/___/___/___	Educated
Unappealing	___/___/___/___/___/___/___	Appealing
Illiterate	___/___/___/___/___/___/___	Literate
Likeable	___/___/___/___/___/___/___	Likable
Lower class	___/___/___/___/___/___/___	Upper class

Has a slow speaking rate	___/___/___/___/___/___/___	Has a fast speaking rate
Intelligent	___/___/___/___/___/___/___	Unintelligent
Motivate students to learn	___/___/___/___/___/___/___	Does not motivate students to learn
Passive	___/___/___/___/___/___/___	active
Bad	___/___/___/___/___/___/___	good

Person of color	___/___/___/___/___/___/___	Caucasian
Sour	___/___/___/___/___/___/___	Sweet
Effective teacher	___/___/___/___/___/___/___	Poor teacher
Experienced	___/___/___/___/___/___/___	Inexperienced
Speaks with foreign accent	___/___/___/___/___/___/___	Speaks with American accent

Talkative	___/___/___/___/___/___/___	Shy
Hostile	___/___/___/___/___/___/___	Good natured
Aggressive	___/___/___/___/___/___/___	Unaggressive
Help students learn	___/___/___/___/___/___/___	Does not help students learn
Nice	___/___/___/___/___/___/___	Awful

Strong	___/___/___/___/___/___/___	Weak
Considerate	___/___/___/___/___/___/___	Inconsiderate
Honest	___/___/___/___/___/___/___	Dishonest
Speak with acceptable pronunciation	___/___/___/___/___/___/___	Speak with unacceptable pronunciation
Facilitate the learning process	___/___/___/___/___/___/___	Does not facilitate the learning process

II. How would you rate the lecture you just heard?

Formal	___/___/___/___/___/___/___	Informal
Boring	___/___/___/___/___/___/___	Interesting
Confusing	___/___/___/___/___/___/___	Clear
Easy to understand	___/___/___/___/___/___/___	Difficult to understand

III. How would you rate the instructor's English you just heard?

*Please answer the following items in terms of the instructor's English-language proficiency ONLY.

1	2	3	4	5	6	7
Low Proficiency			Moderate proficiency			High proficiency

1. The instructor's pronunciation/accent of English….

Did not interfere with understanding	___/___/___/___/___/___/___	Interfere completely with understanding
	1 2 3 4 5 6 7	

2. The instructor's English grammar….

Did not interfere — Interfere completely

with understanding ___/___/___/___/___/___/___ with understanding

3. The instructor's English vocabulary….

Did not interfere — Interfere completely

with understanding ___/___/___/___/___/___/___ with understanding

4. The instructor's Rate of speech

Did not interfere — Interfere completely

with understanding ___/___/___/___/___/___/___ with understanding

5. The instructor's overall ability to communicate in English….

Did not interfere — Interfere completely

with understanding ___/___/___/___/___/___/___ with understanding.

6. The instructor's English contained.…

efficient words and expressions — no efficient words

of ideas ___/___/___/___/___/___/___ and expressions of ideas

7. The instructor's English contained….

clear organization — poor organization

and cohesive device ___/___/___/___/___/___/___ no use of cohesive devices

8. What would be the most important factor to determine the instructor's English proficiency? (* Circle more than one, if necessary)

- *Grammatical accuracy*
- *The use of sophisticated vocabulary*
- *Native-like pronunciation*
- *The successful use of communicative skills*
- *Fluency in speaking*
- *Functional ability as instructor*

*** Please write down any comments, if any, on the instructor's English proficiency.

IV. What country do you think the instructor comes from?

Africa, Argentina, Australia, Brazil, Cambodia, Canada, China, Columbia, Costa Rica England, Egypt, France, Germany, Greece, India, Iran, Iraq, Italy, Japan, Mexico, Nepal, Peru, Russia, Singapore, Saudi Arabia, South Korea, Spain, Sudan, Taiwan, Thailand, Turkey, the US, Vietnam, etc.

APPENDIX E

INTERVIEW QUESTIONS

- Could you tell us your overall impression about the online rating? Any difference between Phase I and Phase II?
- How was your impression about the informal meeting, one hour intercultural intervention with international teaching assistants?
- Please tell us some interesting cases of your ratings, if any. What and why you provided that rating?
- Please explain the way you approached and performed the assessment of the speech samples online. How did you feel about this online speech rating?
- What components (English language proficiency, instructional completence, instruction quality, and accent standardness) do you believe were the most important in your rating of the speech samples? Why?

    - English language proficiency (e.g., 1=low proficiency, 7=high proficiency)
    - Instruction competence (e.g.,1= effective teacher, 7=poor teacher)
    - Comprehensibility (e.g., 1= easy to understand, 7= difficult to understand)
    - Accent standardness (e.g., 1= speak with foreign accent, 7= speak with American accent)

- Could you tell us your overall impression about the online rating? Any difference between Phase I and Phase II?
- Please tell us some interesting cases of your ratings, if any. What and why you provided that rating?
- Please explain the way you approached and performed the assessment of the speech samples online. How did you feel about this online speech rating?
- What components (English language proficiency, instructional completence, instruction quality, and accent standardness) do you believe were the most important in your rating of the speech samples? Why?

    - English language proficiency (e.g., 1=low proficiency, 7=high proficiency)
    - Instruction competence (e.g.,1= effective teacher, 7=poor teacher)
    - Comprehensibility (e.g., 1= easy to understand, 7= difficult to understand)
    - Accent standardness (e.g., 1= speak with foreign accent, 7= speak with American accent)

Appendix F

INTERCULTURAL SENSITIZER SURVEY

First, please introduce yourself briefly to your partners (your name, country or hometown, program of study). Your group can choose one of the following three discussion topics to discuss.

I. Discussion Topic

Please ask your partner to remember a time where a miscommunication occurred because of different cultures: either he/she was misinterpreted or he/she did not know how to interpret the other person's behavior. Please share your answers to the following questions:

1. What happened? (What occurred? Where? When?)
2. How did I interpret the message or behavior then? Would I interpret it the same now?

II. Discussion Topic

You and your partners should guess, individually, what message each person in these pictures below is trying to convey. Share your *guesses* (you may not recognize these). Then, learn at least 2 new nonverbal messages that you could produce if you visited your partners' countries or areas.

III. Discussion Topic

Please share with your partners about some proverbs in your culture that you have often heard from your parents. How do those proverbs influence your life? Would you want to teach your children those proverbs also? What are the most important values that you want to teach your children? (As you share with your partners, please try to find out how your partners think).

Appendix G-1

Mystery Puzzles

Robbery

The robbery was discovered at 8:00 a.m. on Friday, November 12. The bank had closed at 5 p.m. the previous day.	Dirsey Flowers was carrying $500 when police apprehended him and had thrown a package into the river as the police approached.
Miss Ellington stated that her brother Howard, when strolling to Taylor's Diner for coffee about 11:00 p.m. on Thursday, November 11, had see Mr. Smith running from the bank.	The president of the bank, Mr. Albert Greenbags left before the robbery was discovered. He was arrested by authorities at the Mexico City airport at noon on Friday, November 12.
Mr. Greenbags was the only person who had a key to the vault.	The front door of the bank had been open with a key.
Mr. Greenbag's half-brother, Arthur Nodough, had always been jealous of his brother.	A strange, hippie-type person had been hanging around the bank on Thursday, November 11, watching employees and customers.
Anastasia Wallflower of East Birdwatch, Wisconsin, said that she had brought $500 worth of genuine Indian love beads from Dirsey Flowers for resale in her boutique in downtown East Birdwatch.	An acme employee, Howard Ellington, said that a hippie had been hanging around the construction company on Wednesday afternoon.
A substantial amount of dynamite had been stolen from the Acme Construction Company on Wednesday, November 10.	Miss Margaret Ellington, a teller of the bank, discovered the robbery.
The airline clerk confirmed the time of Smith's arrival.	The vault of the bank had been blasted open by dynamite.
Arthur Nodough appeared in Chicago on Monday, November 8, waving a lot of	When police tried to locate the janitor of the bank, Elwood Smith, he had apparently

money.	disappeared.
Mr. Smith was found by the F.B.I. in Dogwalk, Georgia, on November 12. He had arrived there via Southern Airlines Flight 414 at 5:00 p.m. on the 11th.	Anastasia said that Dirsey had spent the night of November 11th at the home of her parents and left after a pleasant breakfast on the morning of the 12th.
In addition to keeping payroll records, Mr. Ellington was in charge of the dynamite supplies of the Acme Construction Company.	The president of the bank had been having trouble with his wife, who spent all of his money. He had frequently talked of leaving her.
There were on planes out of Dogwalk between 4:00 p.m. and 7:00 a.m.	The only keys to the bank were held by the janitor and the president of the bank.
Miss Ellington said that Smith had often flirted with her.	Mr. Smith's father, a gold prospector in Alaska, had died in September.
Miss Ellington often borrowed the president's key to open the bank early when she had an extra amount of work to do.	Mr. Greenbags waited in the terminal at O'Hare Field in Chicago for 16 hours because of engine trouble on the plane he was to take to Mexico City.
The hippie type-type character, whose name was Dirsey Flowers and who had recently dropped out of Southwest Arkansas State Teachers College, was found by police in East Birdwatch, about ten miles from Minnetonka.	

APPENDIX G-2

MYSTERY PUZZLES

Murderer

Miss Smith saw Mr. Kelley go to Mr. Jones' apartment building at 11:55 p.m.	The elevator operator reported to police that he saw Mr. Kelley at 12:15 a.m.
When the elevator man saw Mr. Kelley, Mr. Kelley was bleeding slightly, but he did not seem too badly hurt.	Mr. Kelley had been dead for one hour when his body was found, according to a medical expert working with police.
Mr. Kelley's blood stains were found in Mr. Scott's car.	The elevator man went off duty at 12:30 a.m.
Mr. Jones shot at one intruder in his apartment building at 12:00 midnight.	The bullet taken from Mr. Kelley's thigh matched the gun owned by Mr. Jones.
Mr. Kelley's body was found at 1:30 a.m.	Mr. Kelley's body was found in the park.
Mr. Kelley had destroyed Mr. Jones' business by stealing all his customers.	Mr. Kelley's blood stains were found on the carpet in the hall outside Mr. Jones' apartment.
The elevator man saw Mr. Kelley go to Mr. Scott's room at 12:25 a.m.	Only one bullet had been fired from Mr. Jones' gun.
A knife with Mr. Kelley's blood on it was found in Miss Smith's yard.	The knife found in Miss Smith's yard had Mr. Scott's fingerprints on it.
When he was discovered dead, Mr. Kelley had a bullet hole in thigh and a knife	When police tried to locate Mr. Jones after the murder, they discovered that he had

wound in his back.	disappeared.
The elevator man said that Miss Smith was in the lobby of the apartment building when he went off duty.	The elevator man saw Mr. Kelley's wife go to Mr. Scott's apartment at 11:30 p.m.
Mr. Kelley's wife disappeared after the murder.	Police were unable to locate Mr. Scott after the murder.
It was obvious from the condition of Mr. Kelley's body that it has been dragged a long distance.	Mr. Jones had told Mr. Kelly that he was going to kill him.
The elevator operator said that Mr. Kelley's wife frequently left the building with Mr. Scott.	Miss Smith often followed Mr. Kelley.

APPENDIX H

PARTICIPANTS NEEDED FOR RESEARCH IN SPOKEN ENGLISH ASSESSMENT

You can earn from $24-$56 for 3-7 hours of your time listening to speech samples. No special background is necessary.

Any Undergraduate Student may be eligible

As a participant in this study, you would be asked to anonymous *questionnaires; speech samples evaluation; a meeting with Teaching Assistants (if required)*

For more information about this study, please contact:

Okim Kang
Professor Don Rubin
Dept of Language & Literacy Education
at
Call 706 542-7174 or (Email: *speech@uga.edu*)

This study has been reviewed by, and received ethics clearance through, Institutional Review Board, University of Georgia

University of Georgia

Appendix I

Debriefing Statement—Rating of Teaching Assistants' Speaking Proficiency

Thank you very much for participating in the research study entitled "Rating of Teaching Assistants' Speaking Proficiency" As you may know, courses on campus are often taught by teaching assistants from different linguistic and cultural backgrounds. Their English proficiency as well as their teaching performance can influence the volume and composition of the US undergraduate. The speech samples that you heard and rated in this project are a part of the oral presentations made by these International Teaching Assistants (ITAs).

While everyone can agree that assessing spoken English is important, objective speech evaluation turns out to be a difficult feat to accomplish for several reasons. Listeners often carry unconscious expectations and biases—positive as well as negative--into speech rating situations. In addition, some people's backgrounds might make them more qualified to accurately hear differences in pronunciation.

The study in which you have just participated considered a number of rater variables such as rater foreign language proficiency, contact with non-native speakers of English, and formal study in linguistics. It also included a measure of stereotyped response to non-native speakers. In addition to considering such rater variables, the study also examined the impact of speaker accent as measured by acoustical or sound patterns.

Furthermore, some of the participants in this study received training (an informal meeting with ITAs) in speech rating, but others did not. We will therefore be able to ascertain whether training (the informal interaction with ITAs) can minimize the effects of rater background on the scores they assign.

After we have finished analyzing these data, we hope that we will have a better picture of what kinds of people make the most accurate judges of spoken English and what kind of training can improve judges' accuracy in rating speaking proficiency.

Your participation to this research was essential, and so we are truly obliged to you. Should you wish to receive a copy of the final results of this study, please send an email to speech@uga.edu to request it.

Okim Kang, Doctoral Candidate
Donald Rubin, Academic Advisor

Appendix J

Goodness of Fit Indices for the Random Coefficient Regression Models

Conjoint effects of rater and speaker predictors on oral proficiency ratings

Fit Statistics

-2 Log Likelihood	4591.3
AIC (smaller is better)	4617.3
AICC (smaller is better)	4617.9
BIC (smaller is better)	4646.5

Conjoint effects of rater and speaker predictors on instructional competence ratings

Fit Statistics

-2 Log Likelihood	4707.2
AIC (smaller is better)	4735.2
AICC (smaller is better)	4735.9
BIC (smaller is better)	4766.7

Conjoint effects of rater and speaker predictors on comprehensibility ratings

Fit Statistics

-2 Log Likelihood	4353.3
AIC (smaller is better)	4379.3
AICC (smaller is better)	4379.9
BIC (smaller is better)	4408.6

Conjoint effects of rater and speaker predictors on accent standardness ratings

Fit Statistics

-2 Log Likelihood	3399.1
AIC (smaller is better)	3425.1
AICC (smaller is better)	3425.7
BIC (smaller is better)	3454.3

Conjoint effects of rater and speaker predictors on superiority Ratings

Fit Statistics

-2 Log Likelihood	4488.6
AIC (smaller is better)	4514.6
AICC (smaller is better)	4515.2
BIC (smaller is better)	4543.9

Conjoint effects of rater and speaker predictors on social attractiveness

Fit Statistics

-2 Log Likelihood	4628.2
AIC (smaller is better)	4654.2
AICC (smaller is better)	4654.8
BIC (smaller is better)	4683.4

Appendix K

The 6 X 12 Correlation Matrix

Correlations Among six Dependent Variables and Twelve Acoustic Suprasegmental Measures

Predictors	Oral Proficiency	Instructional competence	Compre-hensibility	Accented-ness	Superiority	Social attractive-ness
Speech Rate	.27**	.28**	.24**	.29**	.19**	.14**
Articulation Rate	.32**	.35**	.31**	.36**	.25**	.20**
Mean Length of Rruns	.34**	.29**	.33**	.28**	.27**	.15**
Phonation Time Ratio	.16**	.17**	.13**	.15**	.10	.06
# of Silent Pauses per Minute	.03	.09*	-.00	.09*	.01	.05
Mean Length of Pauses	-.09*	-.13**	-.06	-.11**	-.05	-.05
#r of Filled Pauses per Minute	-.26**	-.28**	-.22**	-.28**	-.23**	-.09*
Mean Length of Filled Pauses	-.27**	-.27**	-.27**	-.24**	-.23**	-.12**
Pace	.15**	.18**	.15**	.17**	.10*	.08*
Space Ratio	-.35**	-.38**	-.34**	-.37**	-.29**	-.23*8
irregular topic boundary	-.29**	-.32**	-.23**	-.29**	-.28**	-.18**
Overall pitch range	.31**	.34**	.24**	.34**	.30**	.21**

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