RATER EFFECTS IN ITA TESTING: ESL TEACHERS' VERSUS AMERICAN UNDERGRADUATES' JUDGMENTS OF ACCENTEDNESS, COMPREHENSIBILITY, AND ORAL PROFICIENCY

By

Ching-Ni Hsieh

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Second Language Studies

2011

ABSTRACT

RATER EFFECTS IN ITA TESTING: ESL TEACHERS' VERSUS AMERICAN UNDERGRADUATES' JUDGMENTS OF ACCENTEDNESS, COMPREHENSIBILITY, AND ORAL PROFICIENCY

By

Ching-Ni Hsieh

Second language (L2) oral performance assessment always involves raters' subjective judgments and is thus subject to rater variability. The variability due to rater characteristics has important consequential impacts on decision-making processes, particularly in high-stakes testing situations (Bachman, Lynch, & Mason, 1995; A. Brown, 1995; Engelhard & Myford, 2003; Lumley & McNamara, 1995; McNamara, 1996).

The purposes of this dissertation study were twofold. First, I wanted to examine rater severity effects across two groups of raters, English-as-a-Second-Language (ESL) teachers and American undergraduate students, when raters evaluated international teaching assistants' (ITAs) oral proficiency, accentedness, and comprehensibility. Second, I wanted to identify and compare rater orientations, that is, factors that drew raters' attention when judging the examinees' oral performances. I employed both quantitative and qualitative methodologies to address these issues concerning rater effects and rater orientations in the performance testing of ITAs at a large Midwestern university.

Thirteen ESL teachers and 32 American undergraduate students participated in this study. They evaluated 28 potential ITAs' oral responses to the Speaking Proficiency English Assessment Kit (SPEAK). Raters evaluated the examinees' oral proficiency, accentedness, and comprehensibility, using three separate holistic rating scales. Raters also provided concurrent written comments regarding their rating criteria and participated in one-on-one interviews that explored raters' rating orientations. I employed a many-facet Rasch measurement analysis to examine and compare rater severity across rater groups using the computer program FACETS. I compared the written comments across groups to identify major rating criteria employed by the ESL teachers and the undergraduates. I analyzed the interview data to explore the reasons for rating discrepancies across groups.

Results of the study suggested that the ESL teachers and the undergraduate raters did not differ in severity with respect to their ratings of oral proficiency. However, the comparisons of ratings in accentedness and comprehensibility were both statistically significant. The undergraduate raters were harsher than the teacher raters in their evaluations of examinees' accentedness and comprehensibility. Additionally, the analysis of the written comments identified six major rating criteria: linguistic resources, phonology, fluency, content, global assessment, and nonlinguistic factors. Cross-group comparisons of the rating criteria indicated that the undergraduate raters tended to evaluate the examinees' oral performances more globally than the ESL teachers did. In contrast, the ESL teachers tended to use a wider variety of rating criteria and commented more frequently on specific linguistic features. The interview protocols revealed that raters' experience with accented speech, perceptions of accent as an important rating criterion, and approaches to rating (i.e. analytical or global), had important bearings on raters' judgments of ITA speech.

Acknowledgements

I am heartily thankful to my supervisor, Dr. Paula Winke. I have been amazingly fortunate to have a supervisor who granted me the freedom to explore on my own, and at the same time provided me with the guidance to recover when my steps faltered. Her constant encouragement and support helped me tremendously throughout my research.

I would like to thank all my committee members. Dr. Susan Gass has been the most admirable program director and professor. I am genuinely thankful to her for granting me the funding support throughout my study. Dr. Debra Hardison's insightful comments at different stages of my research were stimulating and helped me stay focused. Dr. Debra Friedman has always been there to listen and give advice. I am grateful to her for all the inspirational discussions that helped me generate ideas. I am indebted to Dr. Dan Reed for his continuous guidance on my work and research in language testing and assessment.

Many friends and colleagues have shared with me the ups and downs through my graduate study. I warmly appreciate their friendship and belief in me. My family back home in Taiwan has been a constant source of love, concern, and strength in my life. Most importantly, none of this would have been possible without the love, care, and help of my boyfriend, Vineet, to whom this dissertation is dedicated. I would like to express my heart-felt gratitude to him.

Finally, I would like to thank all the teachers and students who participated in this study and to acknowledge the financial support from a Spaan Fellowship for Studies in Second or Foreign Language Assessment from the English Language Institute at the University of

Michigan, a Russell N. Campbell Doctoral Dissertation Grant from The International Research Foundation, and a Dissertation Completion Fellowship from Michigan State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Second language (L2) oral performance assessment always involves raters' subjective ratings and is thus subject to rater variability. The term *rater variability* refers to variations in scores that raters give that are associated with rater characteristics, but not with examinees' actual performance or ability (Engelhard & Myford, 2003; McNamara, 1996; Myford & Wolfe, 2000). The variability due to rater characteristics has been identified as *rater effects* and has important consequential impacts on decision-making processes, particularly in high-stakes testing situations (Bachman et al., 1995; Barrett, 2001; Engelhard & Myford, 2003; Myford & Wolfe, 2000; Schaefer, 2008; Weigle, 1998). These rater effects are part of what is considered as construct-irrelevant variance and may obscure the construct being measured (Congdon & McQueen, 2000). These effects, therefore, call into question the validity and fairness of performance assessments (Kunnan, 2000, 2005; McNamara, 2000; Meiron & Schick, 2000; Messick, 1989; Weir, 2005).

Research on rater effects in L2 performance assessment has shown that rater effects can be manifested in different forms, such as central tendency, halo effect, and differential rater severity (Eckes, 2005). A central tendency effect is exhibited when raters avoid using the extreme categories of a rating scale and instead prefer using the categories near the midpoint of the scale. A halo effect manifests itself when raters transfer their perceptions of particular features of examinee performance to another, possibly unrelated, features and provide highly similar ratings across those features on the basis of an overall impression or evaluation (Bond & Fox, 2007; Eckes, in press; Linacre, 1989). However, the most prevalent rater effect in performance assessment is *rater severity*. This effect occurs when raters provide ratings that are consistently either too harsh or too lenient as compared to other raters (Bachman et al., 1995; Eckes, in press; Lumley & McNamara, 1995; McNamara, 1996; Weigle, 1998).

Several other factors also influence the ratings of performance assessments. For example, raters may differ in the way they apply the rating criteria or vary in the degree to which they weigh specific linguistic or non-linguistic features of the performance and thus derive different ratings for the same performance or derive the same ratings for different reasons (A. Brown, Iwashita, & McNamara, 2005; Eckes, in press; Papajohn, 2002; Weigle, 1999). Rater background variables, such as raters' occupations (A. Brown, 1995; Chalhoub-Deville, 1995; Hadden, 1991; Meiron & Schick, 2000), gender (O'Loughlin, 2007), first languages (Chalhoub-Deville, 1995; Johnson & Lim, 2009; Kim, 2009; Xi & Mollaun, 2009), second languages (Winke, Gass, & Myford, in press) and rating experience (Cumming, 1990; Weigle, 1994, 1999) may also influence how raters determine their ratings.

Since many important decisions are made based on raters' judgments of test takers' test performances in high-stakes test settings, research studies concerning how to minimize measurement errors resulting from rater effects are crucial. To this end, studies that examine the sources of rater effects and explore rater orientations (i.e. factors that draw raters' attention while rating) are most relevant. Results of such research can inform our understanding of the exact nature of rater variability and help us tackle practical problems regarding rater training.

An examination of the relevant literature indicates that research on rater effects and rater orientations has predominantly focused on L2 writing assessment (e.g., Cumming, 1990; Cumming, Kantor, & Powers, 2002; Eckes, 2005; Lumley, 2002, 2005; Milanovic, Saville, & Shuhong, 1996; Schaefer, 2008; Weigle, 1994, 1998, 1999). There are as of yet relatively few studies that have investigated the effects of rater variability in ratings and how raters make their rating decisions in L2 speaking assessment, despite a growing interest in general in how different rater effects influence the quality of ratings and what raters actually do (A. Brown, 2007; A. Brown & Hill, 2007; Chalhoub-Deville, 1995; McNamara & Lumley, 1997).

While high-stakes language tests, such as the TOEFL® (Test of English as a Foreign Language^{TM}) iBT (internet-based test) (Xi & Mollaun, 2009) or the International English Language Testing System (IELTS) (A. Brown, 2007; A. Brown & Hill, 2007; Merrylees & McDowell, 2007), have been the focus of many investigations on rater effects, published research on rater effects and rater orientations in L2 oral performance assessment has paid little attention to one particular high-stakes testing situation, the English language tests for those who hope to become international teaching assistants (ITAs) at higher educational institutions in North America. This testing situation, henceforth referred to as *ITA testing*, should also be considered when examining rater effects involved in the rating process because the screening of qualified ITAs whose English proficiency is sufficient for instruction and pronunciation is comprehensible to linguistically naïve undergraduates is important for the undergraduates' learning.

Part and parcel to ITA testing is the assumption that the official raters are acting as de facto representatives of the undergraduate student population at their institution, the population from which any class of students an ITA would teach would be drawn. Underlying this assumption is that if an official rater deems the speech of a potential ITA as insufficient (too low in terms of overall proficiency or comprehensibility), then undergraduates would not be able to learn from this person very well due to explicit speech issues, regardless of the person's affability, subject-area content knowledge, or teaching style. In turn, it is assumed that any international student a typical undergraduate cannot understand cannot pass the ITA exam. Any university with an ITA testing program should periodically check that their official raters judge potential ITAs' speech on par with how undergraduates would. Discrepancies in how these two groups rate potential ITAs should be investigated, and how such discrepancies may impact the reliability and validity of ITA testing programs needs theoretical discussion.

The ITA Problem

The growing percentage of ITAs at U.S. universities has raised serious nationwide concerns with the English proficiency of the ITAs and how their proficiency impacts the quality of undergraduate education (Bailey, 1983, 1984b; Chiang, 2009; Muthuswamy, Smith, & Strom, 2004, May; Plakans, 1997; Tyler, 1992). According to a 2006 survey that examined the international graduate applications, admissions, and enrollment trends conducted by the Council of Graduate Schools, 53 percent of graduate students in engineering, biology, and physics in U.S. universities were foreign-born (Council of Graduate Schools, 2007). The high percentage of international students in these technical fields brings a diverse instructional team in many research-oriented universities. It is now a common practice for research-oriented universities to rely on foreign-born graduate students to serve as ITAs and teach basic undergraduate courses, such as mathematics, chemistry, and statistics. ITAs often take roles as lecturers, provide help in the labs, and give out grades. They are sometimes the primary source of input for the incomprehensible course materials. When undergraduates encounter ITAs whose English or pronunciation one can not understand, the communication gap may frustrate the undergraduates or even push students away from a potential major or taking more advanced courses in a given department. Undergraduates' complaints about ITAs' insufficient language ability and heavy foreign accents are often heard among college campus and reported in the news (e.g., Bannon, 2005, June 24; Finder, 2005, June 25; Ruderman, 2000).

In response to American undergraduates' complaints about the difficulty in understanding their foreign TAs, ITAs' speech comprehensibility and English oral proficiency, along with their teaching competence, have come under regulation by more than 20 states since the1980s (see Appendix A). These state statutes and regulations concerning ITA English proficiency aim to ensure that ITAs have sufficient communicative competence to perform their instructional duties (K. Brown, Fishman, & Jones, 1990; Dick & Robinson, 1994; Hoekje & Linnell, 1994; Monoson & Thomas, 1993). The state mandates vary in several ways. Many are in the form of legislative statutes while others are directed by state governing boards or governors. In states both with mandated and nonmandated regulations concerning ITAs, many research universities have used structured or standardized language tests, such as the Test of Spoken English (TSE)[1] or the Speaking Proficiency English Assessment Kit (SPEAK), to asses ITAs' language competence as recommended by state statutes (Monoson & Thomas, 1993; Oppenheim, 1997, March). (A list of 50 universities that used the SPEAK test as an ITA screening tool at the time of this dissertation research is shown in Appendix B.) Instead of using commercially available speaking tests, many universities that employed ITAs developed their own in-house ITA assessment instruments. These test vary in their forms, such as the in the forms of oral interviews (e.g., University of Michigan), oral presentations (e.g., Northeastern University), and teaching simulations (e.g., Brown University).

While the state mandates have seen substantial improvements in the selection of and preparation for qualified ITAs, problems concerning the communication between ITAs and their

[1] The TSE and its institutional version, the SPEAK, are both English oral proficiency test developed by the Educational Testing Service (ETS). They measure the ability of nonnative speakers of English to communicate orally in a North American English context and are widely used as an ITA screening tool.

students remain. In 2005, the North Dakota State Representative Bette Grande proposed a new bill, providing that if a student complained in writing that his or her foreign-born instructor did not speak English clearly and with good pronunciation, that student would be entitled to withdraw from the class with no academic or financial penalty and would even get a refund (Gravois, 2005, April 8). This new law was passed in March 2005 and required universities in North Dakota State not only to test ITAs' English speaking skills, but also to notify students how to file and resolve complaints about their ITAs' speech. The new proposal would remove ITAs from teaching roles if 10 percent of their class complained that they did not speak clearly. However, subsequent controversies and debates regarding the new policy arose because many people argued that student evaluations of ITAs were not necessarily reliable measures of ITAs' speech and may not be fair to the ITAs being tested. Others suggested that communication was a two-way street and the undergraduates should take responsibilities in the communication process and make effort to understand their ITAs (Teicher, 2005).

ITA Research

Given the background regarding American undergraduates' complaints about ITA speech and the establishment of ITA testing state regulations, it is important to understand the linguistic features of ITA speech that cause comprehension difficulty and to examine how ITA speaking proficiency is evaluated because an understanding of these issues can inform ITA programs with regard to their process and ensure that ITA testing is properly executed. Over the past three decades, a number of applied linguists have devoted substantial amount of research effort to identify the problematic features of language use by ITAs in instructional contexts (Bailey, 1984b; Bryd & Constantinides, 1992; Hoekje & Linnell, 1994; Hoekje & Williams, 1992; Rounds, 1987; Tyler, 1992; Williams, 1992). Following Bailey's seminal research on the communicative problems of foreign TAs in the early 1980s (Bailey, 1983, 1984a, 1984b), a wide range of L2 speech features, in particular foreign accent, and their impact on native listeners' comprehensibility have been explored and identified (e.g., Anderson-Hsieh & Koehler, 1988; Hinofotis & Bailey, 1981; Pickering, 2004; Williams, 1992). Researchers have also examined different assessment tools used to screen ITAs. For example, Douglas and Smith (1997) investigated the theoretical underpinnings of the TSE/SPEAK. Hoekje and Linnell (1994) examined the authenticity of language tasks in the SPEAK test. Plough, Briggs, and van Bonn (2010) examined the rating criteria employed by raters who assessed foreign TAs' speaking proficiency. Xi (2008) explored the feasibility of using TOEFL iBT Speaking test as an alternative ITA screening tool. Several of these studies and others will be reviewed more thoroughly in Chapter 2 of this dissertation.

Although research on issues related to ITAs has established that English language proficiency of ITAs and ITAs' accented speech may affect American undergraduates' comprehension, very few studies thus far have examined the issues of rater variability and rater orientations within an ITA testing situation. For example, Myford and Wolfe (2000) examined four sources of rater variability within the TSE. The study showed that TSE raters differed somewhat in the levels of severity they exercised when they rated the examinee performances. More specifically, they found that if the examinees' scores were adjusted for the differences they found in rater severity, then the scores of two-thirds of the examinees would have differed from their raw score averages by 0.5 to 3.6 raw score points. These differences could have substantial consequences for examines whose scores lie within the cut score range.

The paucity of research on rater effects and rater orientations in ITA testing may be partially due to the fact that most researchers in the ITA field are ITA trainers or applied linguists who specialize in speech production or cross-cultural communication. The majority of the ITA research is oriented toward the English language and teacher training of ITAs (Bauer, 1996; Bryd & Constantinides, 1992; Jia & Bergerson, 2008; Landa, 1988; Okoth & Mupinga, 2007, February; Pica, Barnes, & Finger, 1990a, 1990b) and addresses issues such as ITAs' communication problems (Bailey, 1984a; Dick & Robinson, 1994), ITAs' cross-cultural awareness (Gorsuch, 2003; Yook & Albert, 1999), ITAs' identities (LoCastro & Tapper, 2008), and undergraduates' perceptions and attitudes toward ITA speech (K. Brown, 1992; Dalle & Inglis, 1989, March; Muthuswamy et al., 2004, May; Oppenheim, 1996; Plakans, 1997; Rao, 1995, May; Rubin, 1992). Other ITA research focuses on the examination of specific aspects of speech production, such as intelligibility (Isaacs, 2008), primary stress (Hahn, 2004), and discourse structuring cues (Tyler, 1992).

The scarcity of research in the area of rater effects in ITA testing signals a gap in the ITA and language testing literature. To bridge this gap, research into rater effects in ITA testing is critical and needed because decisions made by raters who evaluate ITAs' oral performance has important bearings on the quality of the undergraduate courses taught by ITAs. Results of empirical studies that examine rater effects in specific ITA testing contexts could provide insights into how raters who screen ITAs behave and could identify specific rating concerns associated with ITA language tests.

Context of the Study

Given the increasing research interest in rater effects in L2 speaking assessment (A. Brown, 1995; Chalhoub-Deville, 1995; Eckes, 2005; Lumley & McNamara, 1995; Winke et al., in press) and the paucity of research on rater effects in ITA testing, I carried out this dissertation study to examine rater effects and rater orientations in the ITA testing context at Michigan State

University (MSU). Echoing Issacs' (2008) and Morley's (1994) call for the inclusion of undergraduate students in the ITA screening and test validation process, I included two groups of raters, (a) experienced English-as-a-Second-Language (ESL) teachers who were trained raters that assessed ITAs' oral proficiency professionally and (b) native-English-speaking American undergraduate students at MSU. These raters evaluated potential ITAs' oral responses to the SPEAK test made available by the Testing Office at MSU. Since ITAs' communication problems, as indicated in the bulk of the ITA literature (e.g., Bailey, 1984; Hoekje & Linnell, 1994; Hoekje & Williams, 1992; Rubin, 1992), are often associated with not only ITAs' oral proficiency, but also their pronunciation problems, particularly their foreign accent, I examined rater variability between the two groups of raters in terms of examinees' oral proficiency, degree of foreign accent (accentedness), and perceived comprehensibility.

In this study, *oral proficiency* is operationalized as an examinee's global communicative competence to function at an instructional setting in U.S. higher educational institutions (Douglas & Smith, 1997). The construct definition of accentedness follows Munro and Derwing's work on L2 speech perception and production (Derwing & Munro, 2009; Munro & Derwing, 1995a). According to Derwing and Munro (2009), *accentedness* is defined as "how different a pattern of speech sounds compared to the local variety" (p. 478). Lastly, the construct of comprehensibility is defined as the listeners' estimation of how easy or difficult it is to understand a given speaker (Derwing & Munro, 1997; Munro & Derwing, 1995a).

The inclusion of the constructs of *accentedness* and *comprehensibility* in this study deserves further justification. The conceptions of accentedness and comprehensibility are important sub-dimensions of L2 speech (Munro & Derwing, 1995a, 1999, 2001; Munro, Derwing, & Morton, 2006). A foreign accent is a normal feature of L2 leaning among those who acquire the L2 after puberty (Flege, Munro, & MacKay, 1995; Munro & Derwing, 1995a; Piske, MacKay, & Flege, 2001). Although foreign accent is a salient indicator of a speaker's linguistic background, the impact of foreign accent on listeners' comprehension is complex and compounded by an array of factors. Research has shown that a strong foreign accent does not necessarily entail decreased comprehensibility or intelligibility (Munro & Derwing, 1995a) and that even if an L2 speaker has a strong accent, his or her speech could still be highly comprehensible and intelligible (Munro, Derwing, & Morton, 2006). This complicated interrelationships among accentedness, comprehensibility, and intelligibility have important theoretical implications for construct definition in oral performance assessment and are relevant to the ITA testing context because complaints about ITAs' speech are often discussed in terms of ITAs' pronunciation problems (e.g., Bailey, 1984; Rubin, 1992), in particular, foreign accent and its impact on comprehensibility. Heavy foreign accent, as many studies have shown (Derwing & Munro, 2009; Kang, 2010; Munro, Derwing, & Sato, 2006), has often been deemed as the cause of ITAs' poor communication skills and the main source of undergraduates' comprehension difficulty.

Significance of the Study

This dissertation project closely relates to contemporary scholarship in the fields of ITA research and language testing. First of all, I examined the quality of human ratings in oral performance assessment within a high-stakes testing context. Such investigation is a research priority in performance-based language assessment since high quality ratings are essential for drawing reliable and valid inferences about a test taker's performance (Lumley & McNamara, 1995). Secondly, I addressed issues regarding the comparability and characteristics of oral performance ratings awarded by two rater groups, ESL teachers and American undergraduates, in ITA testing. This research agenda echoes the call for the inclusion of linguistically naïve undergraduate raters in the screening of ITAs in local test validation studies (Isaacs, 2008; Morley, 1994). Thirdly, I employed a mixed-method design in data collection and analysis, a current trend in research methodology within the fields of language testing and second language acquisition. In particular, this design offers a comprehensive and diverse illustration of raters' rating behaviors and allows both the examination of the *quantitative data* (the scores) raters awarded to the speech samples and the exploration of the *thought processes* raters underwent while rating. This methodology provides a deeper understanding of rater effects in performance assessment as different aspects of rater behaviors can be elicited by different methods. Most importantly, this study is distinguished from previous research by its breadth of focus not only on rater variability in oral proficiency ratings, but also on the ratings of two aspects of L2 speech: accentedness and comprehensibility, both of which have important bearings on our understanding of U.S. undergraduates' complaints about ITA speech (Bailey, 1984b; Rounds, 1987; Yule & Hofffman, 1990). This investigation is further supported by data drawn from evaluation criteria commented on by the ESL teachers and the American undergraduates. Results of this dissertation study will provide validity evidence to the rating processes in a specific ITA testing situation as well as shed light on the research of rater effects in the field of language testing and assessment.

Organization of Chapters

The goal of this study was to examine rater effects, particularly differences, or lack thereof, in rater severity between ESL teachers and undergraduate students in terms of their ratings of ITAs' speech samples on oral proficiency, accentedness, and comprehensibility. The secondary goal of this study was to determine factors in the test takers' speech that drew the raters' attention while evaluating the test takers' performance. Chapter 1 presents an overview of the dissertation study. Chapter 2 presents a review of relevant literature and provides a synthesis of research on rater effects, rater orientations, and L2 speech perception and production. Chapter 3 describes the methodology of the study. Chapter 4 reports the results of the quantitative data analysis. Chapter 5 reports the findings and discussions of the qualitative data analysis. Chapter 6 discusses the implications of the research results, addresses the limitations of the study, and makes recommendations for future research. Chapter 7 draws conclusions of the dissertation study.

Chapter Two: Literature Review

The main purpose of this study was to examine rater variability associated with the characteristics of two groups of raters, ESL teachers and American undergraduates, on their ratings of potential ITAs' oral proficiency, accentedness, and comprehensibility. The second purpose of the study was to explore rater orientations, that is, what raters attended to while judging ITAs' performances on speaking tasks. This chapter reviews research in (a) rater effects, (b) rater orientations, and (c) accentedness and comprehensibility in associations with ITAs' communication problems. The literature reviewed below serves as the theoretical underpinnings guiding the research presented in this dissertation.

Rater Effects

Rater effects, such as rater severity or leniency, are often viewed as sources of systematic variance in ratings that are associated with raters and not with the examinees (Eckes, 2005, 2008; Hoyt, 2000; Myford & Wolfe, 2003). Researchers investigating language performance assessments have observed considerable differences in rater severity or leniency (Bachman et al., 1995; A. Brown, 1995; Eckes, 2005, 2008; Engelhard & Myford, 2003; Kim, 2009; Lumley & McNamara, 1995; Lynch & McNamara, 1998; Winke et al., in press). For example, Eckes (2005) employed a many-facet Rasch analysis to examine rater effects on the scoring of the writing and speaking sections of a German test and found that raters differed strongly in how severely they rated the examinees. Also applying a Rasch analysis to model rater bias patterns, Lumley and McNamara (1995; McNamara, 1990, 1996), in their validation research of two sub-tests of an occupational English test for health professionals, found that rater severity was not static; rather it changed over time.

Research on rater variability in L2 speaking tests suggests that different rater backgrounds may impact the way raters assess examinees' language ability and how raters weigh rating criteria (A. Brown, 1995; Elder, 1993; McNamara, 1996; Reed & Cohen, 2001). These studies have addressed issues such as (a) the aspects of the test performances on which raters focus (A. Brown et al., 2005; Orr, 2002) and (b) the impact of rater characteristics, such as teacher versus non-teacher raters (Chalhoub-Deville, 1995), raters' first language (L1) backgrounds (Kim, 2009; Xi & Mollaun, 2009), raters' L2 backgrounds (Winke et al., in press), and raters' countries of origin (Chalhoub-Deville & Wigglesworth, 2005). These idiosyncratic rater backgrounds have been shown to affect ratings to a large extent and should be further investigated.

Barnwell (1989) compared the rating behaviors between untrained, native speakers of Spanish and trained raters when they evaluated American students' performances on an oral interview in Spanish. The researcher reported that the untrained raters were more severe in their evaluations than the trained raters. Chalhoub-Deville (1995) compared the rating behaviors among native speakers of Arabic teaching Arabic in the U.S., non-teaching native speakers of Arabic living in the U.S., and non-teaching native speakers of Arabic living in Lebanon. She found that the non-teacher-rater group in Lebanon emphasized the grammar-pronunciation aspect more while the teacher-rater group in the U.S. was more diversified in terms of the features they employed to evaluate the subjects' L2 oral performance. Hadden (1991) compared ESL teachers and nonteacher raters' perceptions of Chinese speakers' spoken English. Teacher raters were found to be more severe than nonteacher raters in terms of students' linguistic ability, but no difference was found in the areas of comprehensibility, social acceptability, personality, and body language. Citing multiple studies on the effect of rater training (Lunz & Stahl, 1990;

Weigle, 1994), McNamara (1996) discussed cases of surprising differences between raw scores and scores adjusted for rater characteristics. His critical review on the impact of rater backgrounds in scoring suggested that rater training can make raters internally consistent, but cannot eliminate the variability in severity.

To summarize, what all of these studies imply is that raters of different backgrounds may exercise different degree of severity in their judgments of examinee performances and these differences in severity may in turn affect the scores they assign.

Rater Orientations

An import aspect of research on the validity of expert judgments concerns rater orientations, that is, factors that draw raters' attention. Douglas (1994) argues that raters may arrive at similar ratings for very different reasons and test takers may have qualitatively different performances and yet receive similar test scores. Thus far, research on rater orientation has mainly focused on writing assessment (Barkaoui, 2010a, 2010b, 2010c; J. D. Brown, 1991; Connor-Linton, 1995; Cumming et al., 2002; Delaruelle, 1997; Lumley, 2002; Mendelsohn & Cumming, 1987; Milanovic et al., 1996; O'Loughlin, 1994; Sakyi, 2000; Santos, 1988; Shi, 2001; Weigle, 1999). However, there is relatively little research on rater orientations in speaking assessment (A. Brown, 2007; A. Brown et al., 2005; Iwashita, Brown, McNamara, & O'Hagan, 2008; Meiron, 1998, April; Meiron & Schick, 2000; Nakatsuhara, 2008; Orr, 2002).

Research on rater orientations in speaking assessment has found mixed results in terms of how consistent or similar raters evaluated oral performances. Orr (2002) used verbal reports (Ericsson & Simon, 1993), a technique used to elicit individual's spoken thoughts, to investigate raters' decision-making processes. He had 32 raters watch two video recordings of the Cambridge First Certificate in English (FCE) Speaking test. While rating, raters provided verbal feedback (speak aloud) of what they were thinking—these reports were recorded and transcribed. Orr found that trained raters applied different standards in scoring and did not focus on the rating criteria in the same way. Raters varied in the amount of attention they paid to non-criterion aspects of the candidates' performances. The varied nature of the perceptions observed among the FCE raters has led to the author's concern about the possibility of deriving valid test scores in oral performance assessment.

Brown et al. (2005) conducted a comprehensive study to examine rater orientations on two types of TOEFL speaking tasks, the independent and integrated tasks,[2] at different levels of oral proficiency. Ten experienced ESL teachers listened to selected speech samples and provided verbal reports regarding the aspects of the examinees' oral performances to which they attended. Analysis of the verbal reports yielded five major categories: linguistic resources, phonology, fluency, content, and global assessment. Unlike Orr's (2002) findings that showed the varieties and contradictory nature of rater perceptions of examinee performances, Brown et al. concluded that expert raters tended to agree generally as to what aspects of performances were valued.

Two studies (Meiron, 1998, April; Papajohn, 2002) investigated rater behavior and rating orientations in ITA testing. Both used the SPEAK test. Meiron (1998, April) used verbal protocols, written retrospectives, and questionnaires with novice and experienced SPEAK raters to explore rater behaviors on a single SPEAK task, the picture narrative, in which test takers retold a story using a series of picture prompts. She found that raters may take different

[2] Integrated tasks require the test-takers to "process and transform a cognitively complex stimulus (e.g., a written text or a lecture) and integrate information from this source into the speaking performance" while the independent tasks require test-takers to "draw on their own knowledge or ideas to respond to a question or prompt … these tasks are often restricted to fairly bland topics that draw on test-takers' general knowledge" (Brown et al., 2005, p.1).

approaches to rating, such as focusing on certain self-generated features not specified in the rating rubric or weighing differentially on discrete features in the speech samples. Papajohn (2002) explored trained SPEAK raters' concepts of rating using concept mapping, a graphical method for showing meaningful relationships among concepts. He found that the nine trained SPEAK raters, even though they had gone through a standardized training process, still developed individualized concepts of the process of rating and emphasized the rating criteria to varying degrees. He suggested that the key concepts to emphasize in rater training should include holistic rating, the effort required from examiners and listeners, salient features of a response, reference to rating criteria, sustainability, and internalization of the rating criteria. His findings also implied that teaching experience and the way raters internalized the scoring criteria all came into play in the rating processes.

While Meiron (1998, April) and Papajohn (2002) both found various rating features among SPEAK raters, Meiron (1998, April) was mainly concerned with the validation of an existing scale and Papajohn emphasized how the rater trainers could use the information gained from the concept mapping to identify key rating criteria. The researchers, however, did not compare and determine specific performance features on which expert raters and untrained undergraduates focus when raters attempt to reach their judgments of the examinees' speaking ability. These features, nevertheless, are important for ITA testing and for understanding factors that hinder undergraduates' comprehension of ITA speech.

Aspects of L2 Speech with Relevance to ITA Testing

Oral proficiency. Examining rater severity and rater orientation in ITA testing across expert judges and novice raters requires an understanding of previous work that compares ESL teachers and U.S. undergraduates' judgments. Several researchers have compared ESL professionals' and U.S. undergraduates' evaluations of examinee's oral proficiency using standardized tests such as the TSE/SPEAK (Bejar, 1985; Clarke & Swinton, 1980; Powers, Shedl, Wilson-Leung, & Butler, 1999). Clarke and Swinton (1980) administered the TSE to ITAs at eight higher educational institutions in the U.S. The TSE scores, evaluated by ESL professionals, were found to be strong predictors of undergraduates' ratings of ITAs' English communicative competence in classroom lectures as well as in conversational situations. Results of the study revealed that examinees' comprehensibility of ITA speech was more closely related to the pronunciation and fluency aspects of the examinee's speech than to grammar. Powers et al. (1999) investigated the degree to which official TSE scores were predictive of U.S. undergraduates' ability to understand the messages conveyed by potential ITAs in different, U.S.-higher-educational institutions. The researchers found a strong relationship between the examinees' communicative competence and the undergraduates' ability to respond correctly or appropriately to the examinee's message.

Bejar's (1985) study of the TSE raters found that raters who were graduate students pursuing a masters or doctorate in teaching ESL exhibited a smaller degree of rating discrepancies compared to that of ESL professionals. Bejar suggested that the observed differences in the margins of rating discrepancies might be due to the fact that the student raters had daily contact with ITAs in their respective programs or institutions and thus may have became more familiar with the particular ITA's accented speech. Bejar also found that raters differed in severity, especially in their ratings of fluency, and suggested that raters who were too severe or too lenient should be excluded from the rater pool. Alternatively, he recommended that raters should be equated using rigorous psychometrical procedures to control for rating differences.

Other researchers had undergraduates evaluate ITAs' classroom performances and compare undergraduates' evaluations against expert raters' judgments of ITA speech. Orth (1983) compared American undergraduate students and experienced ESL teachers' evaluations of 10 ITAs' oral English proficiency. He had both groups of raters evaluate ITAs' tape-recorded lectures and found drastic difference in the ratings. The correlation in the ratings awarded by the undergraduates and the ESL teachers was very weak (*r* = .12). The undergraduates appeared to rate the ITAs less on linguistic features of their speech than on features of delivery and nonverbal aspects of communication. Orth also indicated that the undergraduates' ratings were biased by the grades they anticipated to receive from these ITAs.

Dalle and Inglis (1989, March) had undergraduates evaluate the intelligibility and clarity of 18 ITAs' classroom performances using the Speech Evaluation scale developed by Orth (1983). The mean ratings of the speech evaluation was moderately correlated (*r* = 0.6) with the scores the ITAs received on the SPEAK test. Similarly, Oppenheim (1998, March) compared ESL raters' ratings of ITAs' English oral proficiency against undergraduate raters' ratings of ITAs' linguistic skills. She found that the rater groups' assessments varied significantly. On the contrary, Saif (2002) recruited a panel of two ESL instructors and three undergraduate students to rate 26 ITAs' oral proficiency during or shortly after the ITAs took an oral proficiency test. The study revealed a high level of reliability among the five raters, suggesting that the linguistically naïve, untrained undergraduate raters were able to rate the ITAs' oral performance consistently as the ESL experts did. Different from Orth's (1983) study, the undergraduate raters at Saif's study did not know the ITAs they evaluated in person. Thus, it appears that they were able to judge the ITAs' performance more objectively without personal bias involved. However, it remains inconclusive whether novice undergraduate raters who do not have a direct relationship with the examinees may rate as objectively as observed in Saif's study and more research is needed.

Accentedness. Foreign accent has been one of the major topics for research in the field of L2 speech perception and production (Derwing & Munro, 1997, 2005, 2009; Derwing, Munro, & Wiebe, 1998; Flege, 1988a; Flege & Fletcher, 1992; Flege et al., 1995; Isaacs, 2008; Kang, 2010; Munro & Derwing, 1995a, 1995b, 1999, 2001). The degree to which native listeners perceive L2 learners as having accented speech varies widely by listener, although no accent is normally perceived as intrinsically best (Derwing & Munro, 2009).

Munro and Derwing (1995a) examined the interrelationships among accentedness, comprehensibility, and intelligibility in the speech of L2 learners. They had 18 English-speaking, undergraduate students listen to two, English-native speakers and 10 proficient, Chinese ESL speakers' narrations based on a one-page cartoon. Listeners transcribed three short excerpts of each speaker's narrative and evaluated the speaker's accentedness and comprehensibility on a 9-point, holistic scale. The researchers found that the speakers' utterances in English were highly intelligible and comprehensible although the listeners' ratings on accentedness were fairly widely dispersed along the scale, with a major proportion in the heavily accented range. The researchers suggested that the presence of a strong foreign accent may not necessarily result in reduced intelligibility or comprehensibility.

Researchers have identified a range of segmental and suprasegmental features of L2 speech that affect listeners' judgments of accentedness. Some of the most researched aspects include speaking rate (Munro & Derwing, 1998, 2001; Trofimovich & Baker, 2006), pausing (Kang, 2010; Trofimovich & Baker, 2006), stress (Juffs, 1990; Kang, 2008, 2010; Trofimovich & Baker, 2006; Zielinski, 2006, 2008), and intonation (Kang, 2008, 2010; Pickering, 2001, 2004). Each of these features plays a role in the perception of accented speech, although none of its own is solely responsible for a listener's judgements of a particular L2 speaker's speech.

In ITA-related research, several speech features have important effects on the kinds of conclusions that can be drawn about native listeners' judgements of ITA accented speech. For example, Kang (2010) analyzed 11 ITAs' in-class lectures acoustically and found that ITAs' accent ratings were best predicted by pitch range and word stress measures. Hahn (2004) found that native listeners tended to process ITA accented speech more easily when primary stress was correct although, contrastively, Trofimovich and Baker (2006) did not find stress timing to be a significant predictor of accent ratings in their Korean ITAs' sentence reading. They found instead that the speakers' pause duration and speech rate contributed more to foreign accent.

Foreign accent has been the main target of blame for American undergraduate students' difficulties in understanding ITA accented speech (Bailey, 1984a; Bauer, 1996; Bryd & Constantinides, 1992; Dalle & Inglis, 1989, March; Derwing & Munro, 2009; Kang, 2010; Landa, 1988; Munro, Derwing, & Sato, 2006; Oppenheim, 1996, April; Pica et al., 1990a; Rao, 1995, May; Rubin & Smith, 1990; Sebastian & Ryan, 1985). Poor pronunciation, in most cases referring to heavy foreign accent, has often been seen as the cause of ITAs' poor communication skills or deemed as a cover term that signals ITAs' ineffective classroom instructions (Rubin, 1992). In an early study, Hinofotis and Bailey (1981) compared American undergraduates and ESL teachers' evaluations of ITAs' oral performances using videotaped, mock lectures produced by ITAs at the University of California, Los Angeles. They found that both the undergraduates and the ESL teachers ranked *pronunciation* as the single most prominent failure in ITAs' overall communicative competence. Plakans (1997) conducted a large-scale survey to examine undergraduates' experiences with and attitudes toward ITAs at Iowa State University. She found that the two most common complaints about ITAs' language use were ITAs' poor pronunciation and their inability to understand and answer students' questions satisfactorily.

Research in language attitudes has indicated that native listeners' perceptions of a foreign accent may influence their attitudes toward and affective reactions to the speakers (Cargile & Giles, 1997; Lindemann, 2002; Lippi-Green, 1997). Cargile and Giles (1997) found that native listeners felt less pleasure after hearing a speaker with a Japanese accent than after hearing a speaker with a standard American accent. Lindemann (2002) showed that native listeners reacted negatively to speakers with a Korean accent while the speakers were conducting communication tasks.

Within the field of ITA research, foreign accent has long been shown to have a noteworthy impact on American undergraduates' attitudes toward ITAs (Bailey, 1984a; Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002; Fox & Gay, 1994; Rubin, 1992; Rubin & Smith, 1990). Rubin and Smith (1990) had two native speakers of Cantonese record a highly accented and a moderately accented version of simulated classroom lectures on two different topics. Undergraduate raters were recruited to evaluate one or the other of the recordings, which were accompanied by a photograph of either a European or an Asian instructor. Results of their study showed that when students were faced with the Asian instructor, they perceived higher levels of foreign accentedness and judged the speakers to be poor teachers. Rao (1995, May) explored American undergraduates' affective, cognitive, and behavioral responses when the students interacted with a foreign TA on the first day of class. Results of the study showed that when the undergraduates expected that their foreign TA's accent would be difficult to follow, they exhibited higher levels of anger and anxiety, evaluated the foreign TA less favorably on communication competence, and were more likely to drop the TA's class. Bresnahan, Ohashi, Nebashi, Liu, and Shearman (2002) examined undergraduates' attitudinal and affective responses toward American English and two conditions of foreign accent (intelligible versus unintelligible). Results revealed that the undergraduates exhibited more positive attitudinal and affective responses to American English as opposed to intelligible, foreign-accented speech.

To address native listeners' negative attitudes towards accented L2 speech, Derwing, Rosssiter, and Munro (2002) carried out an experiment that attempted to teach native speakers to listen to foreign-accented speech. Results of their experiment suggested that the undergraduates, through very limited instruction of a particular accent (Vietnamese), became more confident to understand foreign accents and felt more willing to communicate with individuals who spoke English as an L2. This study is particularly important for ITA testing and instruction programs because it suggests that initial complaints by undergraduates may be unduly harsh or judgmental due to personal bias or prejudice. However, effective training in how to listen to L2 speakers may contribute to positive changes in attitudes and willingness to interact with and listen to accented speech.

Comprehensibility. Research in L2 speech perception and production has shown that different aspects of phonemic segmentals and prosody features have different impacts on comprehensibility (Munro & Derwing, 1995a). In ITA-related research, a number of studies have sought to identify factors that affect American undergraduates' comprehension of ITA speech. These studies suggested factors such as speech rate (Munro & Derwing, 1998), discourse-level language use (Davies, Tyler, & Koran, 1989; Pica et al., 1990a; Tyler, 1992), intonation and tone (Kang, 2008; Pickering, 2004), accent familiarity (Rubin, 1992; Rubin & Smith, 1990), and personal emotions (Yook & Albert, 1999) are all attributable to comprehension difficulties in different ways. These factors will be reviewed below.

Speech rate. Speech rate has been shown to impact listeners' comprehensiblity of L2 speech. Anderson-Hsieh and Koehler (1988) investigated the effect of foreign accent and speech rate on native-speakers' comprehension of ITAs' class presentation. They found that ITAs' speech rate had an important effect on the undergraduates' comprehension of course materials. The increase in the ITAs' speaking rate from a regular to a fast rate was found to result in a decrease in the listeners' comprehension for the most heavily accented speakers. Munro and Derwing (1998) conducted two experiments to test the hypothesis that accented speech heard at a slower rate would sound less accented and more comprehensible than speech produced at a normal rate. Interestingly, they found that, in the first experiment, listeners preferred to listen to accented speech at slower rates, while during the second experiment listeners preferred some speeded passages, but none of the slowed ones. Taking these mixed findings into consideration, the researchers concluded that it was unclear whether faster speech or slower speech was more preferable for native listeners when they listened to accented speech (Derwing & Munro, 2001; Munro & Derwing, 2001).

Discourse structuring cues. Another aspect of L2 speech critical for comprehension relates to the use of discourse structuring devices. Tyler, Jefferies, and Davies (1988) examined ITAs' videotaped teaching demonstrations. They found that discourse structuring cues native speakers used to construct coherence or to orient their listeners to the relative importance among ideas in the discourse were absent in Chinese and Korean ITAs' speech. The ITAs instead constructed an undifferentiated, flat discourse structure. The undergraduates judged these ITAs' lecturers to be disorganized and unfocused. Tyler (1992) used a qualitative, discourse-analytic framework to analyze the discourse structure of two spoken texts produced by a native English speaker and a Chinese TA. The analysis revealed a variety of differences in the two speakers' use of discourse structuring devices, particularly in the areas of lexical discourse markers, lexical specificity, and syntactic incorporation. Tyler argued that the differences in the discourse-level patterns exhibited in the two speakers' spoken discourses were one of the factors that interfered with undergraduates' abilities to understand foreign TAs' speech. Williams (1992) examined 24 ITAs' planned and unplanned discourses at several U.S. research universities. She found that ITAs' planned discourses contained more markers that contributed significantly to undergraduates' comprehension, and stated that explicit use of discourse markers was critical for the comprehensibility of ITA speech.

Intonation and tone. The impact of intonation and tone choices on comprehensibility has also been researched in the ITA literature. Tyler et al. (1988) found that there were many inappropriate falling intonation in their Chinese and Korean ITAs' speech. Pickering (2001) examined how tone choices, that is, the choice of a sustained rising, falling, or level pitch movement, contributed to ITAs' communication failure in the classroom. The researcher compared the tone choice features of six Chinese ITAs and six American TAs by analyzing these TAs' classroom presentations during their regular course of teaching. Results of the study showed substantial differences in the number of specific tone choices and the way these tones were used between the two groups of TAs.

Accent familiarity. Familiarity with foreign accent is also an important factor that influences native listeners' comprehension of ITA speech. Gass and Varonis (1984) asked undergraduate students to transcribe sentences and to summarize a short story at a variety of familiarity conditions. The results demonstrated that familiarity with a topic, with accented speech in general, with a particular accent, and with a particular speaker all had an impact on listening comprehension. Rubin and Smith (1990) carried out an experiment to identify factors that predicted undergraduates' comprehension of accented speech. They used highly and moderately accented speech as well as native speakers' productions. The undergraduate participants who had taken courses from ITAs appeared to score higher on the comprehension tests.

Personal emotions. Undergraduates' comprehension of ITA speech is also associated with their emotional reactions toward the speaker. Yook and Albert (1999) examined the relationships among undergraduates' comprehension of ITA speech, evaluations of ITAs' language competence, and emotions. They had 422 American undergraduates view a 5-minute videotaped presentation by a male, Asian TA who majored in engineering at a large Midwestern university. After viewing the presentation, the students answered four open-ended questions tapping the content of the presented material to assess their comprehension. The students also listened to the speaker's responses to the SPEAK test, and rated the intensity of eight emotions while they viewed the video. The researchers found that positive emotions led to higher evaluations of the ITA's language competence, whereas negative emotions led to lower evaluations, suggesting that nonlanguage factors, such as personal emotions or reactions to accented speech, can impact native listeners' judgements of L2 speech.

To summarize, the studies reviewed in this chapter suggest that raters may vary substantially in the severity they exercise when evaluating oral performances, and that raters can derive similar (or different) ratings for different reasons. The reviewed studies have also provided evidence showing that American undergraduates and ESL teachers might agree more than they disagree with respect to their evaluations of ITAs' oral proficiency and yet may perceive accented speech in distinctive ways. A variety of factors that contribute to native listeners' judgements of L2 speech have also been established. However, these studies focused on oral proficiency, accentedness or comprehensibility, or a combination of accentedness and comprehensibility. A critical issue yet to be resolved concerns the relative effect of rater backgrounds on raters' judgements of these three measures.

Much more work is needed for two main reasons. First of all, none of the studies directly investigated the precise nature regarding the degree to which ESL teachers and American undergraduates differ in severity, or do not differ, with respect to their ratings of ITAs' oral performance. No known study has yielded systematic, detailed comparisons of judgements made by ESL teachers and U.S. undergraduates in order to determine the role of rater background in measures of oral proficiency, accentedness, and comprehensibility in ITA testing. If the speech features of ITAs are the predominate factors influencing raters' judgements of L2 speech, then rater backgrounds would only be a minor contributor to score discrepancies. And we would expect a high degree of agreement or little difference in rater severity among ESL teachers and undergraduates regarding scores on oral proficiency, accentedness, and comprehensibility associated with a particular ITA. In contrast, as much research has suggested, human judgements are highly subjective and affected by different rater characteristics. In this case, the ESL teachers and the undergraduates may exhibit a varying degree of rater severity and respond to the same oral response in inconsistent ways. Since the relative contributions of rater background effects remain unclear, in this study, I chose to compare ESL teachers and undergraduates' evaluations of ITA speech to determine whether raters from different backgrounds share similar judgements with regard to oral proficiency, accentedness, and comprehensibility.

Secondly, previous studies have not systematically examined ESL teachers and undergraduate raters' rating orientations while assessing ITA speech. As reviewed in the previous section, raters' rating orientations may differ depending on the raters' backgrounds and/or teaching and training experiences. Interpreting two groups of raters' evaluations of ITAs' oral proficiency requires an understanding of the factors on which they based their judgements. On the one hand, factors to which raters paid attention might reveal important features of the speech itself that were salient to the listeners' judgments regardless of backgrounds. On the other hand, factors to which raters pay attention might deviate from individual to individual because of familiarity with accents or teaching experience.

In this study, I compared ratings of oral proficiency, accentedness, and comprehensibility awarded by two groups of raters to address the issues listed above. I also explored the factors to which raters attended while they evaluated ITAs' oral proficiency.

The research questions guiding this study included:

1. Do ESL teachers and American undergraduate students differ in the severity with which they evaluate potential ITAs' oral proficiency, accentedness, and comprehensibility, respectively, and if so, to what extent?

2. What factors draw raters' attention while the raters evaluate potential ITAs' oral proficiency? Are different factors more or less salient to different rater groups?

To examine rater variability associated with the characteristics of these two groups of raters, I employed a many-facet Rasch measurement (MFRM) that can provide a fine-grained analysis of multiple variables potentially having an impact on the ratings (Bond & Fox, 2007; Eckes, in press; Linacre, 1989, 1998). To discern what raters focused on when evaluating the speech samples, and to investigate differences in rating orientations between the rater groups, raters provided concurrent written reports and participated in follow-up interviews that helped explain the potential causes for rater variability in scoring. The next chapter details the research design of the study.

CHAPTER 3: METHODOLOGY

Test Examinees

The data for this study consisted of 28 examinees' oral responses to the SPEAK test during operational SPEAK test administrations at MSU. The requested SPEAK test data included the recordings of the examinees' complete responses to the test. The SPEAK test data were pre-rated by official SPEAK raters. The examinees' actual SPEAK scores ranged from 40 to 55, with five-point increments. The cut score for a qualified ITA was 50 for MSU examinees.

The examinees were international graduate students who were seeking an ITA opportunity at MSU. They were 10 Chinese, 10 Korean, and 8 Arabic native speakers; 19 were males and 9 were females (see Table 1).

Table 1. Demographics of Examinees

Examinee ID	Gender	First Language	Actual SPEAK Score
1	M	Chinese	40
2	M	Chinese	40
3	F	Chinese	40
4	M	Chinese	45
5	M	Chinese	45
6	F	Chinese	45
7	M	Chinese	45
8	M	Chinese	50
9	M	Chinese	50
10	M	Chinese	50
11	M	Korean	40
12	M	Korean	40
13	M	Korean	40
14	F	Korean	45
15	F	Korean	45
16	M	Korean	45
17	F	Korean	45
18	F	Korean	50
19	M	Korean	50
20	F	Korean	50

Table 1 (Cont'd)

Examinee ID	Gender	First Language	Actual SPEAK Score
21	M	Arabic	40
22	F	Arabic	40
23	M	Arabic	40
24	M	Arabic	45
25	M	Arabic	45
26	M	Arabic	50
27	M	Arabic	50
28	F	Arabic	50

Rating Materials

The SPEAK test comprises twelve tasks, each of which is designed to elicit a particular speech act from the examinees. These speech acts include different language functions such as narrating, apologizing, persuading, recommending, and giving and supporting opinions. The test is administered aurally using prerecorded prompts and printed test booklets. The task types include descriptions of maps, story-telling based on a sequence of pictures, discussions of topics of general interests, descriptions of information presented in a simple graph, and presentations of information from a revised schedule. The time allotted to each response ranges from 30 to 90 seconds and the entire test takes around 20 minutes.

Three tasks from each examinee's response to the SPEAK test were chosen for ratings. These tasks included a picture description, a topic discussion, and a presentation on a revised schedule. The entire response time of these three tasks was approximately four minutes. There were a total of 84 speech samples (28 samples on each of the three tasks) for evaluation. These speech samples were saved on an online rating system for rater evaluations. All raters rated all 84 speech samples.

Rating Scales

Raters judged examinee performances using three sets of rating scales. The first one was the 5-point holistic SPEAK rating scale (see Appendix C) and was used to assess examinees' overall oral proficiency. Raters utilized this scale, ranging from 20 to 60 (20= no effective communication; no evidence of ability to perform task; 60= communication almost always effective; task performed very competently), with a 10-point increment. The ratings indicated raters' evaluations of an examinee's overall task performance with respect to each task. The second and third rating scales were both a 9-point holistic scale, following Munro and Derwing (1995a). I chose these two scales for the ratings of accentedness (1= no accent; 9= heavily accented) and comprehensibility (1= very easy to understand; 9 extremely difficult or impossible to understand), respectively.

Raters

Two rater groups participated in this study. The first rater group included 13 ESL teachers who were trained SPEAK raters at MSU. All the ESL teachers had some experience in rating speaking tests (e.g., SPEAK, placement tests, classroom-based achievement tests). There were 5 males and 8 females. Their ages ranged from 29 to 56 years (*M* = 39.9, *SD* = 9.1). The teacher raters all had academic backgrounds in language education or linguistics and experience teaching ESL at a level similar to the test examinees in the present study. Their years of ESL/EFL teaching experience ranged from 6 to 22 years (*M* = 12.5, *SD* = 6.1). The mean length of SPEAK rating experience was 4.5 years (*SD* = 5.5). All raters reported the nonnative accents they were familiar with. Raters were most familiar with the accents of Arabic, Chinese, Korean, Japanese, and Spanish speakers. Table 2 details the background information of the teacher raters.

Table 2. Demographics of ESL Teacher Rater Group

Rater ID	Gender	Age	L 1	Teaching experience (yrs)	Level of students[a]	SPEAK rating experience (yrs)	Accent familiarity[b]
T1	F	38	Japanese	11	2	0.5	1, 2, 3, 5
T2	F	39	English	18	1, 2, 3, 4	13	1, 2, 3, 4, 5, 7, 8
T3	M	56	English	26	1, 2, 3	16	1, 2, 4, 5, 7, 8
T4	F	30	English	8	1, 2, 3, 4	0.5	1, 2, 3, 4, 5, 6, 7, 8, 9
T5	F	29	English	11	1, 2, 3	0.5	1, 2, 3, 4
T6	M	35	English	10	1, 2, 3, 4	0.5	1, 2, 3, 4, 5, 6, 7, 11
T7	M	40	English	8	2, 3	0.5	1, 2, 3, 4, 5
T8	F	35	English	12	1, 2, 3	7	1, 2, 3, 4, 5, 6, 7, 12
T9	M	36	English	9	1, 2, 3	11	1, 2, 4, 5
T10	F	56	English	6	1, 2	4	1, 2, 4
T11	F	34	English	7	1, 2, 3	0.5	1, 2, 3, 4, 6, 12
T12	M	38	English	14	1, 2, 3	0.6	3, 4, 5, 13
T13	F	53	Turkish	22	1, 2, 3, 4	3	1, 2, 3, 5, 6, 7, 8, 9

[a] 1= beginner, 2= intermediate, 3= advanced, 4= superior

[b] 1= Arabic, 2= Chinese, 3= French, 4= Korean, 5= Japanese, 6= Russian, 7= Spanish, 8= Thai, 9= Turkish, 10= Vietnamese, 11= Hungarian, 12= German, 13= Indonesian

The second rater group included 32 American undergraduate students at MSU (see Table 3). The undergraduate raters were all native speakers of English and from a wide variety of academic programs. There were 9 males and 23 females. Their ages ranged from 18 to 22 years (*M* = 20.1, *SD* = 1.2). They had been studying at MSU between one and four years at the time of data collection (*M* = 2.8, *SD* = 0.9). The undergraduate raters all had experience taking courses taught by ITAs at MSU. The number of courses they had taken by ITAs ranged from 1 to 6 (*M* = 2.9, *SD* = 1.6). Twenty-four undergraduates reported that they talked to their ITAs once or less than once a week, seven reported twice a week and one reported three times a week. All the undergraduates had ITAs whose first languages were either Chinese or Korean. Few had Arabic, Japanese, Hindi, or Spanish ITAs. Twelve undergraduates had no foreign friends either while studying at MSU or in their upbringing and no foreign-accent familiarity. Twenty reported having a few foreign friends (mostly Arabic, Chinese, Korean, and Hindi speakers) and had limited exposure to foreign accents. None of the undergraduates had prior experience rating non-native English speakers' speech. Table 4 provides a summary of the background characteristics of both rater groups.

Table 3. Demographics of Undergraduate Rater Group

Rater ID	Gender	Age	Major	Years at MSU	ITA courses taken	Frequency talking to ITAs[a]	Foreign friends and accent familiarity
U1	F	19	Education	2	2	1	No
U2	F	19	Criminal justice	2	1	3	No
U3	F	21	Microbiology	4	6	2	Korean
U4	F	20	Missing data	2	2	2	Chinese, Arabic
U5	M	22	Computer Science	4	6	1	Chinese, Arabic
U6	F	20	Education	3	2	2	Chinese
U7	F	19	Education	2	1	2	No
U8	F	21	Education	4	3	1	No
U9	F	21	Education	3	1	3	Chinese
U10	F	20	Education	3	6	2	No
U11	M	22	Chemistry	3	1	2	No
U12	F	20	Humanities	3	5	1	No
U13	M	22	Spanish	4	5	1	No
U14	F	22	Dietetics	4	4	2	No
U15	M	19	Nursing	2	1	1	Chinese, Korean
U16	F	19	Education	2	2	2	Chinese
U17	M	19	Chemistry	3	2	3	No
U18	F	22	Spanish	4	4	4	Chinese, Korean
U19	F	20	English	3	5	1	Chinese
U20	F	19	Psychology	2	4	1	Chinese, Arabic
U21	M	20	Political Science	2	2	1	Chinese
U22	M	22	Interdisciplinary	3	4	1	Chinese, Arabic
U23	F	21	Education	4	4	3	Korean, Arabic,
U24	F	21	Education	4	4	2	No
U25	F	20	Education	3	1	2	Hindi
U26	M	20	History	3	2	1	No
U27	F	18	Communications	1	1	1	Arabic
U28	F	18	Pre-Med	1	2	3	Chinese, Korean
U29	F	19	Nursing	2	2	3	Chinese, Arabic
U30	F	20	Education	3	3	3	Chinese
U31	M	18	Criminal justice	1	1	2	Chinese
U32	F	21	Communications	3	3	2	Chinese

[a] 1= less than once a week, 2= once a week, 3= twice a week, 4= three times a week.

Table 4. Background Information by Rater Group

		ESL teachers (*n* = 13)	Undergraduates (*n* = 32)
Gender	Male	5	9
	Female	8	23
Age		*M* = 39.9 (*SD* = 9.1)	*M* = 20.1 (*SD* = 1.2)
Teaching experience		*M* = 12.5 years (*SD* = 6.1)	N/A
SPEAK rating experience		*M* = 4.5 years (*SD* = 5.5)	N/A
Years at MSU		N/A	*M* = 2.8 (*SD* = 0.9)
ITA courses taken		N/A	*M* = 2.9 (*SD* = 1.6)

Procedure

Prior to rating, I informed the raters about the purpose of the research project and the research design. I then introduced them the construct definitions of oral proficiency, accentedness, and comprehensibility, and the three rating scales employed. Once raters gave their consent to participate in the study, they completed a background questionnaire that contained questions about their demographic information (see Appendices D and E). Since the ESL teachers were all trained raters, no rater training or norming session was undertaken. The undergraduate raters were engaged in a minimal, one-on-one training, which consisted of acquainting the raters with the rating tasks and the rating rubrics. However, I did not give them any extensive practice rating or rater norming. I minimized the training in order to capture the novice raters' rating behaviors and to reflect their impressionistic judgments of foreign TAs' oral performances.

Raters evaluated examinee performances online. The examinee order was randomized across tasks and raters. I instructed the raters to rate the recordings in a quiet room that had Internet access. Raters rated from Task 1, then moved on to Task 2, and finally to Task 3. With respect to the three measures (oral proficiency, accentedness, and comprehensibility), I told the raters to listen to each of the recordings and assign scores on the examinee's oral proficiency first. Raters could listen to each recording multiple times if they considered it necessary. Immediately after they assigned an oral proficiency rating to a response, they proceeded to provide written comments regarding factors that drew their attention while they made their rating decisions. Raters could skip the entering of written comments for any particular recording if they chose to. Once raters completed the written comments (or chose to skip it), they then moved on to assign ratings on accentedness and comprehensibility for each recording. The entire ratings took between four to six hours. After raters completed their ratings, I conducted semi-structured interviews with the raters one-on-one. The interviews lasted between 30 minutes and one hour. The interview questions are listed in Appendix F.

Data Analysis

Analysis of rating data. To answer the research questions outlined previously, I analyzed the rating data using (a) different statistical analysis approaches, including descriptive and inferential statistics using SPSS (version 16) and (b) the MFRM analysis using the computer program FACETS (Version 3.67) (Linacre, 2010). To assist in the interpretation of the data analysis, an overview of the MFRM model employed in this study is given below.

The MFRM approach has been applied substantively to model rater effects in the field of language testing and assessment (Bachman et al., 1995; Eckes, 2005, 2008; Engelhard & Myford, 2003; Lumley, 2002; Lumley & McNamara, 1995; Myford & Wolfe, 2000, 2003). Evidence derived from these studies has accumulated, pointing to substantial degrees of systematic error in rater judgments that may lead to inappropriate interpretation of scores obtained from human raters (Weigle, 1999). The MFRM approach takes into account various facets, such as examinee ability, rater severity, and task difficulty, simultaneously when analyzing rating data. This joint calibration of multiple facets makes it possible to map rater severity on the same scale as examinee ability and task difficulty. These calibrations are often expressed in a common equal-interval metric, that is, the log-odds unit, or logit scale (Linacre, 1989, 1998; McNamara, 1996). As Linacre (1998) suggests, when the rating data show sufficient fit to the model, researchers can draw useful, diagnostically informative comparisons among the various facets.

In the current study, the MFRM model implemented included four facets: examinees, raters, tasks, and rater status (ESL teachers versus undergraduate raters). I carried out three separate FACETS analyses to determine whether the rater groups differed in severity when they rated the examinees' oral proficiency, accentedness, and comprehensibility.

The MFRM model for the ratings of oral proficiency is as follows:

$$\log_e(P_{nijsk}/P_{nijs(k\text{-}1)}) = B_n - D_i - S_j - R_s - F_k$$

where

P_{nijsk} = probability of examinee *n* receiving a rating of *k* on task *i* from rater *j*

$P_{nijs(k\text{-}1)}$ = probability of examinee *n* receiving a rating of *k-1* on task *i* from rater *j*

B_n = oral proficiency for examinee *n*

D_i = difficulty of task *i*

S_j = severity of rater *j*,

R_s = rater status, *s*, and

F_k = difficulty of receiving a rating of *k* relative to a rating of *k-1.*

The MFRM model for the ratings of accentedness is as follows:

$$\log_e(P_{nijsk}/P_{nijs(k\text{-}1)}) = A_n - D_i - S_j - R_s - F_k$$

where

A_n = Accentedness for examinee *n*.

The MFRM model for the ratings of comprehensibility is as follows:

$\log_e(P_{nijsk}/P_{nijs(k-1)}) = C_n - D_i - S_j - R_s - F_k$

where

C_n = Comprehensibility for examinee *n*.

For each element of each facet, the FACETS analysis provides a measure (a logit estimate of the calibration) and a standard error (information about the precision of that logit estimate). FACETS also provides a set of fit statistics that show the degree to which observed ratings match the expected ratings generated by the Rasch model (Linacre, 1989; Myford & Wolfe, 2003). Fit statistics also indicate the consistency with which each individual rater uses the rating scale across examinees, tasks, and rating criteria. Huge differences between the observed and expected ratings are expressed as standardized residuals and indicate unexpected results (Engelhard & Myford, 2003; Myford & Wolfe, 2003, 2004). The standardized residuals can be summarized over different facets and elements to provide indices of model-data fit (Linacre, 2002).

I estimated the global model fit of the FACETS runs based on the proportion of unexpected observations to the overall observations. With respect to the rater effects of interest to this study, I used rater measurement reports to examine rater severity or leniency. I also conducted a general investigation of rater consistency based on the inspection of rater fit statistics to examine rating behaviors and to identify misfitting raters.

Analysis of written comments and interviews. To answer the second research question regarding the factors that draw raters' attention while make judgements on examinees' oral proficiency, I analyzed the written comments and the interview data to understand factors that influenced rating decisions. A detailed discussion of the coding process and the analysis of the written comments is provided as follows.

The entire written comment dataset consisted of 2,151 comments provided by the 13 ESL teachers and 32 undergraduates across the three rating tasks. The ESL teachers provided 742 (34.5%) of the comment entries and the undergraduates 1,409 (65.5%). The comments varied in length, ranging from one word, such as "comprehensible" to a few sentences, such as "Accent is strong, but little interference with meaning. Speech is a bit choppy, but the performance is strong overall." The longest single comment entry consisted of 95 words.

The first step I took to analyze the rater comments was to review the data impressionistically in order to identify factors raters paid attention to. Then I read and re-read the written comments several times until I was fully familiar with the entire comments. Then I made several attempts to divide the original 2,151 comments into separate segments. Each segment included a single or several sentences, either continuous or separated by other sentences but failing within the same comment, with a single aspect of the performance as the focus (A. Brown et al., 2005; Green, 1998; Patton, 1990). I began this process by using Brown et al.'s (2005) empirically developed coding scheme for rater orientations in speaking tasks. As reviewed previously, Brown et al.'s coding scheme consists of five main coding categories: linguistic resources, phonology, fluency, content, and global assessment, and 15 subcategories. This coding scheme was a good fit for the current study because it was developed for the examination of rater orientations in speaking assessment. In addition, the scientific and transparent development process of the coding scheme reported in Brown et al. provided validity evidence of the coding process for the current study.

On the basis of a few cycles of data segmentation, I considered that Brown et al.'s coding scheme was very instrumental for this study because it could account for the majority of the data. However, I also considered that it was necessary to add one additional, main category, *nonlinguistic factors*. The initial categorization showed that 59 comments (20 from the ESL teachers and 39 from the undergraduates) addressed the issues of test-takers' test-taking behaviors (e.g., the use of test-taking strategies) and the emotional status of the examinees (e.g., confident or nervous) that could not be categorized under any of the five existing coding categories. While the number of these comments was minimal compared to the rest of the comments categorized for the five major categories, the fact that these factors were not part of the rating criteria specified in the official SPEAK rating rubric warranted further investigation. I believed that it was important to add this new coding category so as to specifically deal with nonlinguistic factors, an aspect not addressed in Brown et al.'s study.

In addition to the nonlinguistic factors, I also added two subcategories in the coding scheme. The first one pertained to the *accent* aspect of L2 and was added within the main category of phonology. The new subcategory accent only includes comments that are directly related to the examinee's foreign accent, such as "The accent is really heavy" or "I can't understand the speaker because of his accent." It should not be confused with the subcategory "pronunciation" which only includes comments related to the articulation of vowels and consonants, such as "The vowel sound 'a' is off" or "The speaker dropped the final consonant 't'." I considered accent an indispensable subcategory because it appeared that most raters mentioned the impact of accent on comprehensibility, either positively or negatively. A second new subcategory, *organization*, was added within the main category of content. My initial reviews of the written comments suggested that many comments from both rater groups were related to the organization of the examinee speech, including the organization of ideas and the overall organization of the responses. This new subcategory organization was added to examine and reflect how raters judged the organization of ideas examinees produced in their responses.

On the other hand, two subcategories *amount* and *framing* within the main category of content in Brown et al.'s coding scheme were removed. This was done because I found no relevant comment in the dataset related to these two coding categories. The revised version of the coding scheme consisted of six main categories: linguistic resources, phonology, fluency, content, global assessment, and nonlinguistic factors, and 15 corresponding subcategories. Two researchers reviewed this version of coding categories and provided feedback on the clarity and appropriateness of the category descriptions. Table 5 displays the final version of the coding scheme used in the current study.

Table 5. Coding Scheme

Main categories	Subcategories	Examples
Linguistic resources	Grammar	There were a few verb tense errors.
	Vocabulary	Very poor word choice.
	Expressions	There are some awkward expressions.
	Textualization	There is no strong use of cohesive devices.
Phonology	Pronunciation	The vowels seem to be lengthened.
	Intonation	The speech is full of intonation in odd places.
	Rhythm and stress	The stress inhibits complete comprehension.
	Accent	His accent was really heavy.
Fluency	Pauses	There were a lot of pauses in his speech.
	Repetition and repair	His repetitions of words affected the flow.
	Speech rate	She spoke too slowly.
	Global fluency	The speaker had some trouble with fluency.
Content	Task fulfillment	The task was not completed.
	Ideas	Hard to catch several ideas.
	Organization	Good organization to his response.
Global assessment	No subcategory	Well done; I could understand everything.
Non-linguistic factors	No subcategory	Perhaps his anxiety was in control.

Note: Adapted from "An examination of rater orientations and test-taker performance on English-for-academic purpose speaking task," by Brown, A., Iwashita, N., and McNamara, T. F., 2005, p.16.

To facilitate the coding process, I first segmented the entire dataset at the main category level using the finalized coding scheme. For example, the comment "There was a huge pause in the speech, and task was not completed" was divided into two segments, first of which was identified as related to the *fluency* feature of the performance and the second to *content*. Comments that addressed the same features of the performance in consecutive sentences were kept intact. For example, the comment "Final consonants are overstressed or elongated. Final consonants are even stressed so much to create an additional consonant effect." was considered addressing the issue of phonology and was not further divided. Although most of the comments were relatively short and straightforward, some were lengthier, more complex, and contained repeated references to the same category. In these cases, it was not always clear whether the repeated references were recycling of the same idea or a new idea. To make the coding process consistent, I treated all references to one of the major aspects of the performance as a single segment. I used an ellipsis (…) to indicate the linking of two separate phrases or sentences in the same segment. For example, the comment “This person paused excessively although the pauses didn't confuse me. … he only paused to correct the organization and pronunciation of his words,” was identified as one single category, fluency. Once the segmentation at the main category was completed, I categorized each segment into subcategories within corresponding main categories, following the same procedure.

To check the reliability of the coding, a second coder and I coded a random sample of 444 segments (approximately 20% of the data). Intercoder percentage agreement was calculated at the main category level (see Table 6). The overall percentage agreement achieved was 79.7%. Percentage agreement within each main category varied, with the highest agreement achieved among the phonological and fluency features, which were quite clearly identified in the rater comments. The main categories of content, global assessment, and non-linguistic factors achieved relatively low intercoder agreements. In terms of content, it was sometimes difficult for the two coders to determine whether a comment should be coded as “organization” within content or as “global assessment” because occasionally raters made comments regarding the organization of the entire responses (e.g., “The speaker was hard to follow because of her poor organization). To make the coding process consistent, we decided to code any mention of the organization of ideas under the subcategory of organization.

With regard to global assessment, we disagreed as to what to and what not to include in this category initially because several comments were about the holistic quality of the responses and at the same time provided linguistic features that could possibly be categorized under other subcategories. For example, in the comment “He is mostly effect and coherent,” I first coded it under the subcategory of textualization within the main category of linguistic resources while the second coder coded it for global assessment. After our discussion, I have come to realize that the mention of "coherent" here was not the same as cohesion, which refers to the use of cohesive devices or discourse markers, but refer to the overall comprehensibility or effectiveness of the response. We then decided to code this and other similar comments under the category of global assessment.

The last category, non-linguistic factors, received the lowest intercoder agreement rate. However, since there were only six cases being categorized and therefore the four cases that we disagreed upon somewhat made the percentage of disagreement look more serious than it actually was. Essentially, the second coder was confused as to how to categorize the mentions of test-taking strategies and test anxiety and had coded them for global assessment. After discussion, we decided that factors such as test-taking strategies or examinees' emotional reactions should be coded for non-linguistic factors and should not be confused with the judgments of the global quality of the responses. After the second coder and I discussed the 90 difficult cases one by one, we reached a 100% agreement. I then coded the entire data set both at the main category and subcategory levels.

Table 6. Intercoder Agreement by the Main Categories

	Number of codes by coder 1	Coders 1 & 2 number of codes agreement	Coders 1 & 2 % agreement
Linguistic resources	87	77	88.50%
Phonology	134	125	93.30%
Fluency	85	78	91.80%
Content	45	28	62.20%
Global assessment	87	44	50.60%
Non linguistic factors	6	2	33.30%
Total	444	354	79.70%

When coding the entire dataset, I coded each segment for one main category and one subcategory. For example, a segment coded as relevant to the main category of linguistic resources could be further coded to the subcategory of grammar or vocabulary. When tallying the frequency of codes within each main category and subcategory, I counted each code once for the main category and the subcategory. For example, a segment coded as being related to grammar within the main category of linguistic resources was counted once for the subcategory of grammar and once for the main category of linguistic resources. The entire dataset was coded into a total of 4,308 segments at the main category level, with 1,650 (38.3%) from the ESL teachers' comments, and 2,658 (61.7%) from the undergraduates'.

After data coding, I calculated the frequencies and proportions of comments made by each individual rater on each coded main category and subcategory. Then I performed a series of statistical analyses to compare the comments made by the ESL teachers and the undergraduates to determine whether the rater groups differed in the rating criteria they mentioned in the comments. Results of these analyses will be reported in the Chapter 5.

The analysis of the interview data follows the content analysis for qualitative data analysis (Miles & Huberman, 1994). The interviews were transcribed verbatim and the transcripts were subject to several iterative categorizations and analyzed through diverse analytical methods, including pattern identification, clustering of conceptual groupings, and constant comparisons. During the concept formation stage of the data analysis, I read and re-read the transcripts several times until I was closely familiar with the transcripts in their entirety. I then gradually conceptualized the underlying patterns in the data and categorized concepts gained from the data. Through constant comparison method of qualitative data analysis, coherent and related comments in the interviews were grouped as one theme (McCracken, 1988; Miles &

Huberman, 1994; Patton, 1990). The major themes identified from the interview data will be discussed in Chapter 5 to unfold the complex interaction between rater orientations and their impact on ratings.

CHAPTER 4: RESULTS OF QUANTITATIVE DATA

Descriptive Statistics

This section reports the descriptive statistics and rater reliability of the rating data. Table 7 details the raw score average of the measures of oral proficiency, accentedness, and comprehensibility for each examinee. The scores were calculated by averaging the ratings awarded by the entire pool of raters within each group.

Table 8 presents the descriptive statistics of the three measures by rater group, including the ranges of the scores, means, and standard deviations of the ratings. The descriptive statistics show that the ESL teacher group had a slightly higher mean and lower standard deviation for the oral proficiency ratings than did the undergraduate raters. Contrastively, the teacher raters had lower means and higher standard deviations for the accentedness and comprehensibility ratings than did the undergraduate raters. Lower means in accentedness suggest that the examinees were rated as having lesser degree of foreign accent; lower means in comprehensibility means that the examinees were easier to understand. Taken together, the results indicate that the ESL teachers were more lenient in their ratings of oral proficiency, accentedness, and comprehensibility. Nevertheless, there was more variation in the teachers' ratings than in the undergraduates'.

Table 7. Raw Score Average Taken over Raters within Rater Group

	Oral Proficiency		Accentedness		Comprehensibility	
Examinee ID	ESL Teachers	Undergraduates	ESL Teachers	Undergraduates	ESL Teachers	Undergraduates
1	39.5	33.8	7.0	7.6	4.9	5.6
2	38.5	36.7	7.1	7.3	5.4	5.1
3	41.0	40.2	5.9	6.5	3.9	4.5
4	42.1	39.1	5.5	6.0	4.2	4.5
5	44.6	47.0	4.4	4.9	3.0	3.3
6	42.3	42.9	5.5	6.2	3.7	4.2
7	41.0	40.6	6.2	6.5	4.3	4.4
8	43.1	42.6	6.0	6.6	3.6	4.2
9	42.6	41.7	5.5	6.5	3.7	4.3
10	43.6	45.4	4.9	4.9	3.5	3.7
11	38.7	35.1	6.7	7.4	4.6	5.3
12	43.6	46.4	4.3	5.4	3.0	3.5
13	41.8	40.1	6.5	6.9	4.2	4.6
14	47.7	46.7	3.8	5.0	2.6	3.2
15	43.1	45.3	4.1	4.9	3.2	3.7
16	42.3	40.1	5.8	6.7	3.8	4.5
17	45.6	45.2	4.6	5.6	2.9	3.7
18	46.7	44.6	4.2	5.3	2.8	3.5
19	48.5	48.9	4.0	5.2	2.4	2.9
20	44.9	48.6	4.2	4.4	2.9	3.2
21	35.6	29.8	6.4	7.1	5.4	6.0
22	40.8	39.9	5.8	6.5	3.9	4.6
23	41.0	39.6	6.1	6.7	4.0	4.6
24	43.8	41.8	4.5	6.0	3.4	4.2
25	42.1	41.7	5.2	6.1	3.5	4.3
26	45.6	46.6	4.4	5.2	2.7	3.5
27	43.6	42.2	5.7	6.6	3.3	3.9
28	52.6	52.8	2.5	3.3	1.9	2.4

Table 8. Descriptive Statistics by Rater Group

Rater group	Measures	Max. possible score	Min.	Max.	*M*	*SD*	Mean as a % of the max. possible score
ESL Teachers	Oral proficiency	60	35.6	52.6	43.1	3.4	71.8%
	Accentedness	9	2.0	7.0	5.2	1.1	57.8%
	Comprehensibility	9	1.9	5.4	3.6	0.9	60.0%
Undergraduates	Oral proficiency	60	29.8	52.8	42.3	4.9	70.5%
	Accentedness	9	3.3	7.6	6.0	1.0	66.7%
	Comprehensibility	9	2.4	6.0	4.1	0.8	45.6%

Interrater Reliability

Interrater reliability was conducted at the group level for the three measures separately, using Cronbach's Alpha. For oral proficiency, the interrater reliability was computed at .96 for the whole rater group, .88 for the ESL teachers, and .95 for the undergraduates. For accentedness, the interrater reliability was computed at .97 for the whole rater group, .92 for the ESL teachers, and .95 for the undergraduates. For comprehensibility, the interrater reliability was computed at .96 for the whole rater group, .92 for the ESL teachers, and .93 for the undergraduates. All of the interrater reliability indexes were within acceptable range (Field, 2009).

Classification of ITA Assignments

In order to compare how oral proficiency ratings assigned by the ESL teachers and the undergraduates performed in classifying the examinees into one of the three ITA assignment outcomes: nonpass, provisional pass, and clear pass, a detailed classification was carried out, using the mean oral proficiency ratings taken over each rater group. When an examinee's average oral proficiency rating was below 42.4, he or she was categorized in the nonpass category of the ITA assignment. A mean oral proficiency rating of 42.5 to 47.4, which was rounded to 45 in operational SPEAK test administration, was categorized as provisional pass. A rating of 47.5 or higher, which was rounded to 50 in operational SPEAK test administration, was categorized as a clear pass. A score of 50 means that the examinee is qualified to work as an ITA at MSU. Table 9 reports the numbers of examinees and the examinee IDs in each of the three ITA assignments by rater group.

Table 9. Classification of ITA Assignments by Rater Group

	ESL teachers		Undergraduates	
	N	Examinee ID	N	Examinee ID
Nonpass (below 42.4)	13	1, 2, 3, 4, 6, 7, 11, 13, 16, 21, 22, 23, 25	15	1, 2, 3, 4, 7, 9, 11,13,16, 21, 22, 23, 24, 25, 27
Provisional pass (between 42.5 and 47.4)	12	5, 8, 9, 10, 12, 15, 17, 18, 20, 24, 26, 27	10	5, 6, 8, 10, 12, 14, 15, 17, 18, 26
Clear pass (47.5 and above)	3	14, 19, 28	3	19, 20, 28

The percentage agreement on ITA assignments between the ESL teachers and the undergraduate raters further demonstrates the similarities, or dissimilarities, in the raters' evaluations of a qualified ITA's oral proficiency. In Table 10, the diagonal cases were those the rater groups agreed on while off-diagonal cases were those they did not agree upon. The rater groups agreed 80% both on the nonpass and the provisional pass cases. They agreed less concerning assignments to the clear pass category (66.7%). However, the low percent agreement observed here was somehow inflated because there were only three cases in the class pass category assigned by each rater group. The rater groups disagreed only on one of the three cases and thus the low percent agreementage regarding the class pass category should be interpreted with caution.

In terms of the nonpass category, the undergraduates appeared to be harsher in their evaluations. They assigned two more cases to the nonpass category (*n* = 15) than ESL teachers did (N=13). Regarding the provisional pass category, the undergraduate raters assigned fewer cases (*n* = 10) than the ESL teachers did (*n* = 12). Both rater groups assigned three examinees to the clear pass category, although the three examinees assigned to the category were not identical. A scrutiny of the clear pass cases indicates that the rater groups disagreed on the ITA assignment for examinees 14 and 20. While the ESL teachers classified examinee 14 as clear pass and examinee 20 as provisional pass, the undergraduates classified the examinee 14 as provisional pass and examinee 20 as class pass. The ESL teachers assigned a raw score of 47.7 for examinee 14 and 44.9 for examinee 20 whereas the undergraduates assigned a raw score of 46.7 for examinee 14 and 48.6 for examinee 20. The ESL teachers assigned an averaged score of 1.0 higher than the averaged score assigned by the undergraduates for examinee 14. The ESL teachers, however, assigned an averaged score of 3.7 higher than the averaged score assigned by the undergraduates for examinee 20. These raw score differences had a significant impact on the ITA assignment.

Table 10. Agreement on ITA Assignment between Rater Groups

Undergraduates	ESL teachers			% Agreement
	Nonpass	Provisional pass	Clear pass	
Nonpass	12	3	0	80%
Provisional pass	1	8	1	80%
Clear pass	0	1	2	66.7%
Overall %				78.6%

To summarize, the fact that the majority of the examinees were considered not qualified to work as ITAs indicates that, on the one hand, the examinees' oral proficiency may well fall below a rating of 50 judged by the raters, and on the other hand, regardless of rater status, raters tended to be conservative in assigning a rating of 50, if no clear indication of strong oral proficiency was manifested in the response.

Correlation Analysis

Correlations among measures within rater groups. Pearson product-moment correlation coefficients were computed to determine the relationships between (1) oral proficiency and accentedness, (2) oral proficiency and comprehensibility, and (3) accentedness and comprehensibility. The group-averaged ratings on oral proficiency, accentedness, and comprehensibility were used for the computation. Table 11 shows the correlation coefficients.

The correlations among oral proficiency, accentedness, and comprehensibility were all highly significant. The negative correlations between oral proficiency and accentedness indicate that when the examinee's oral proficiency was rated higher, he/she was perceived less foreign accented. The negative correlations between oral proficiency and comprehensibility mean that when an examinee was rated higher on oral proficiency, he/she was perceived as easier to understand. Contrastively, the positive correlations between accentedness and comprehensibility indicate that when an examinee was perceived more heavily accented, his/her speech was perceived more difficult to understand. For both rater groups, the strongest correlation was observed between oral proficiency and comprehensibility and the weakest relationship was between oral proficiency and accentedness. Nevertheless, the differences in the strengths of the correlations are marginal, indicating that the constructs of oral proficiency, accentedness, and comprehensibility are closely related.

Table 11. Pearson Correlations among Measures within Rater Groups

	Oral proficiency	Accentedness
Accentedness	-.89* (ESL Teachers) -.90* (Undergraduates)	
Comprehensibility	-.94* (ESL Teachers) -.99* (Undergraduates)	.92* (ESL Teachers) .91* (Undergraduates)

**p* < .001

Correlations among measures across rater groups. A second set of correlation analysis was performed to examine the degree of correspondence between the ratings of oral proficiency, accentedness, and comprehensibility across rater groups. The correlations were computed by correlating the group-averaged ratings of the three measures. As Table 12 shows, the correlations were all highly significant, suggesting that ratings of oral proficiency, accentedness, and comprehensibility assigned by the ESL teachers and the undergraduates were closely related. The strongest correlation was between the group-averaged comprehensibility ratings, followed by the accentedness and oral proficiency. These relationships reveal that the ESL teachers and the undergraduates ranked order the perceived comprehensibility and accentedness in a consistent manner.

Table 12. Pearson Correlations among Measures across Rater Group

Undergraduates	ESL Teachers		
	Oral Proficiency	Accentedness	Comprehensibility
Oral Proficiency	.92*	-.89*	-.95*
Accentedness	-.85*	.95*	.86*
Comprehensibility	-.94*	.91*	.96*

*p <.001

FACETS Analyses

Global model fit. The Rasch model is a prescriptive statistical method and requires the data to fit the Rasch model well. If the data fit the model, then the dataset as a whole supports a unidimensional measurement and captures one latent variable (Linacre, 1989). To estimate the overall data-model fit for each of the FACETS runs, the unexpected responses given the assumptions on the model were investigated. Linacre (2010) suggests that satisfactory model fit is achieved when about 5% or less of the absolute standardized residuals are equal or greater than 2, and about 1% or less are equal or greater than 3.

There were a total of 3,780 valid responses in each of the three measures. The percentage of responses associated with absolute standardized residuals equal or greater than 2 and equal or greater than 3 were calculated to assess data-model fit. Table 13 details the estimates. The examination of the standardized residuals indicated a satisfactory model fit for the three separate measures.

Table 13. Global Model Fit

	StdRes ≥ |3|	StdRes ≥ |2|
Oral proficiency	0.9%	4.7%
Accentedness	1.1%	4.2%
Comprehensibility	0.9%	4.8%

Variable maps. Figures 1, 2, and 3 show the variable maps of the examinee ability, rater severity, task difficulty, and rating scales for the measures of oral proficiency, accentedness, and comprehensibility. In each of the FACETS analyses, the rater and task facets were centered, but not the examinee facet. This is because the examinee abilities are conventionally measured from the local origins of all the other facets. If the average ability is high, then the average examinee ability has a positive logit measure. If the average ability is low, then the average examinee ability has a negative logit measure. Contrastively, the task difficulties and the rater severities are both measured from the center, the local origin. As such, the average rater severity has a severity of 0 logits and the average task difficulty has a difficulty of 0 logits, as can be seen at the center of each variable map.

The FACETS computer program calibrates the examinees, raters, tasks, and the rating scales so that all the facets are positioned on the same scale, creating a single frame of reference for interpreting the results from the analysis. The logits scale is an equal-interval scale, in which all logit units have the same value (Linacre, 1989), unlike the raw scores in which the distances between intervals may not be equal (Bond & Fox, 2007). In each of the three figures, the leftmost column is the logit scale. The highest values are located at the top of the variable map and the lowest values are located at the bottom.

The second column of the variable map shows the examinee measure under investigation. Each examinee was located along the logit scale according to his or her estimated value of oral proficiency, accentedness, and comprehensibility, respectively. For oral proficiency (Figure 1), the examinees were ordered with the most able at the top and the least able at the bottom. The examinees' oral proficiency measures ranged from -0.54 logits to 4.33 logits, with a total spread of 4.87 logits. For accentedness (Figure 2), the examinees were ordered with the most heavily accented at the top and the least accented at the bottom. The examinees' accentedness measures ranged from -1.18 logits to 1.59 logits, with a total spread of 2.77 logits. For comprehensibility (Figure 3), the examinees were ordered with the most difficult to understand at the top and the easiest at the bottom. The examinees' comprehensibility measures ranged from -1.83 logits to 0.45 logits, with a total spread of 2.28 logits.

The third column of the variable map compares the raters in terms of the level of severity or leniency each exercised when rating the examinees' performances. The most severe rater was located at the top of the logit scale and the most lenient at the bottom. For oral proficiency ratings, the harshest raters, (U5, U18), had a severity measure of 1.85 logits, while the most lenient rater (U25) had a severity measure of -2.03 logits. The rater severity measures in oral proficiency show a 3.88-logit spread. For accentedness ratings, the harshest rater (T11) had a severity measure of 1.68 logits, while the most lenient rater (U22) had a severity measure of -1.44 logits. The rater severity measures in accentedness show a 3.12-logit spread. For comprehensibility ratings, the harshest rater (T8) had a severity measure of 1.24 logits, while the most lenient rater (T12) had a severity measure of -1.48 logits. The rater severity measures in comprehensibility show a 2.72-logit spread.

The fourth column of the variable map shows the difficulty of the three rating tasks. The most difficult task is at the top and the least difficult at the bottom. For oral proficiency ratings, the task difficulty measures show a 0.43-logit spread. For accentedness ratings, the task difficulty measures show a 0.08-logit spread. For comprehensibility ratings, the task difficulty measures show a 0.13-logit spread. These small ranges of logit spread indicate that the three rating tasks were approximately the same in their difficulty levels.

The fifth column of the variable map displays the rating scales raters used to evaluate the speech samples. For oral proficiency, the 5-point SPEAK rating scale, ranging from 20 to 60 was used. For accentedness and comprehensibility, a 9-point rating scale was used.

Figure 1. Variable Map for Oral Proficiency Ratings

```
 Logit   Examinee              Rater                 Task  Rating
 Scale                                                     Scale
__________________________________________________________________
|   5 + More proficient    +    Severe               +Hard +(60) |
|     |                    |                         |     | --- |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|     | 28                 |                         |     |     |
|     |                    |                         |     |     |
|   4 +                    +                         +     +     |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|     |                    |                         |     |  50 |
|     | 19                 |                         |     |     |
|     |                    |                         |     |     |
|   3 + 20                 +                         +     +     |
|     | 14                 |                         |     |     |
|     | 26  5              |                         |     |     |
|     | 12  17             |                         |     |     |
|     | 10  15  18         |                         |     | --- |
|     |                    |                         |     |     |
|   2 +                    +                         +     +     |
|     | 24  27  6   8      | U18  U5                 |     |     |
|     | 25  9              |                         |     |     |
|     |                    | U27                     |     |     |
|     | 13  16  22  3   7  |                         |     |     |
|     | 23  4              | U15                     |     |  40 |
|   1 +                    + U32  U8                 +     +     |
|     |                    | U1   U3                 |     |     |
|     | 2                  | T11  U7                 |     |     |
|     | 11                 | T8   U14  U17  U22  U6  |     |     |
|     | 1                  | T10  T5   U28           |     |     |
|     |                    | T12  T2   T7   U29      | 2   | --- |
*   0 *                    * U10  U21  U31  U4   U9  * 3   *     *
|     |                    | T3   U13                | 1   |     |
|     |                    | T9   U12  U16  U23  U30 |     |     |
|     | 21                 | T1                      |     |     |
|     |                    | T13                     |     |     |
|     |                    | U2                      |     |  30 |
|  -1 +                    + T4   U20                +     +     |
|     |                    | T6                      |     |     |
|     |                    | U19  U26                |     |     |
|     |                    |                         |     |     |
|     |                    | U24                     |     |     |
|     |                    | U11                     |     | --- |
|  -2 +                    + U25                     +     +     |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|     |                    |                         |     |     |
|  -3 + Less proficient    +     Lenient             +Easy + (0) |
__________________________________________________________________
Mean      1.84                          .00            .00
SD        1.00                          .90            .22
__________________________________________________________________
```

Figure 2. Variable Map for Accentedness Ratings

```
 Logit   Examinee          Rater              Task    Rating
 Scale                                                Scale
_____________________________________________________________
|    2 + More accented  +  Severe             +  Hard   + (9) |
|     |                |                    |         |     |
|     |                |                    |         |     |
|     |                | T11                |         |     |
|     | 1              |                    |         | --- |
|     |                | U30                |         |     |
|     | 11  2          | U18                |         |     |
|     |                |                    |         |     |
|     | 21             |                    |         |     |
|     | 13             | U23                |         |  7  |
|   1 +                + U9                 +         +     |
|     | 16  23  8      | U19  U6            |         |     |
|     | 22  27  3   7  | U7                 |         | --- |
|     | 9              |                    |         |     |
|     | 6              | U28                |         |     |
|     | 25  4          | T8   U20  U4   U5  |         |  6  |
|     | 24             |                    |         |     |
|     |                | U12  U26           |         |     |
|     | 17             | T1                 |         | --- |
|     |                | U1   U16           |         |     |
*   0 * 10  12  18  26 * U10  U24  U32      * 1  2  3 *  5  *
|     | 15  19  5      | T13  U13  U21  U27 |         |     |
|     | 14             | U11  U14  U15  U17 |         | --- |
|     |                | T10  T12  T5   T7  |         |     |
|     | 20             |                    |         |     |
|     |                | U25                |         |  4  |
|     |                | T2   U2   U29  U3  |         |     |
|     |                |                    |         | --- |
|     |                |                    |         |     |
|     |                | T3   T9   U31  U8  |         |     |
|  -1 +                +                    +         +     |
|     |                |                    |         |  3  |
|     | 28             |                    |         |     |
|     |                | T6                 |         |     |
|     |                | T4   U22           |         |     |
|     |                |                    |         | --- |
|     |                |                    |         |     |
|     |                |                    |         |     |
|     |                |                    |         |     |
|     |                |                    |         |     |
|  -2 + Less accented  +   Lenient          + Easy    + (1) |
_____________________________________________________________
Mean        .47                .00              .00
SD          .63                                 .04
_____________________________________________________________
```

Figure 3. Variable Map for Comprehensibility Ratings

```
 Logit        Examinee                 Rater                       Task  Rating
 Scale                                                                   Scale

|   2 +Difficult to understand +     Severe                        +Hard + (9) |
|     |                        |                                   |     |     |
|     |                        |                                   |     |     |
|     |                        |                                   |     |     |
|     |                        |                                   |     |     |
|     |                        |                                   |     | --- |
|     |                        |                                   |     |     |
|     |                        |                                   |     |     |
|     |                        | T8                                |     |     |
|     |                        |                                   |     |  7  |
|   1 +                        +                                   +     +     |
|     |                        |                                   |     |     |
|     |                        | U30  U5                           |     | --- |
|     |                        |                                   |     |     |
|     |                        | U1   U18                          |     |     |
|     |                        | T13  U27  U32                     |     |  6  |
|     | 21                     | U15  U22  U28  U29  U6            |     |     |
|     |                        | U8                                |     | --- |
|     | 1                      | T5   U14  U17  U21  U23  U3   U4  |     |     |
|     | 11  2                  | T2   T7   U31                     | 2   |     |
*   0 *                        * U13  U26                          * 3   *  5  *
|     |                        | U11  U9                           | 1   |     |
|     | 13                     | T11  U10  U12  U2                 |     | --- |
|     | 16  22  23  3   4   7  | T1   T10                          |     |     |
|     |                        | U19  U20  U7                      |     |  4  |
|     | 24  25  6   8   9      | U25                               |     |     |
|     | 27                     | U24                               |     |     |
|     | 10                     |                                   |     | --- |
|     | 15  17                 | T6   T9   U16                     |     |     |
|     | 12  18                 | T3   T4                           |     |     |
|  -1 + 26  5                  +                                   +     +     |
|     | 14  20                 |                                   |     |  3  |
|     |                        |                                   |     |     |
|     | 19                     |                                   |     |     |
|     |                        |                                   |     |     |
|     |                        | T12                               |     | --- |
|     |                        |                                   |     |     |
|     |                        |                                   |     |     |
|     | 28                     |                                   |     |     |
|     |                        |                                   |     |     |
|  -2 + Easy to understand     +    Lenient                        +Easy + (1) |

Mean       -.56                                   .00                .00
SD          .50                                   .53                .06
```

Rater measurement report. FACETS produces an estimate (in logit) of the degree of severity each rater exercised, the error associated with this estimate, and fit statistics for detecting model-data fit for each individual rater. The rater measurement estimates are represented in Tables 14, 15, and 16, for oral proficiency, accentedness, and comprehensibility, respectively. In each table, raters are arranged from the most severe to the least severe at the bottom, as indicated by the severity measure logit.

Table 14. Differences in Severity on Oral Proficiency

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
U5	1.85	0.16	0.85
U18	1.83	0.16	0.95
U27	1.55	0.16	0.95
U15	1.13	0.16	1.41
U32	1.00	0.16	0.80
U8	0.97	0.16	0.85
U1	0.85	0.16	1.51
U3	0.79	0.16	1.26
T11	0.72	0.16	0.42
U7	0.61	0.16	0.65
U22	0.56	0.16	1.24
U6	0.53	0.16	0.79
U14	0.53	0.16	1.32
U17	0.51	0.16	0.91
T8	0.48	0.16	0.55
T5	0.30	0.16	0.38
T10	0.27	0.16	1.09
U28	0.27	0.16	1.61
T2	0.24	0.16	0.45
T7	0.24	0.16	0.51
T12	0.24	0.16	0.52
U29	0.08	0.16	1.14
U9	0.03	0.16	1.66
U31	0.03	0.16	1.05
U4	-0.02	0.16	1.47
U21	-0.05	0.16	0.76
U10	-0.08	0.16	1.15
T3	-0.13	0.16	0.66

Table 14. (cont'd)

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
U13	-0.24	0.16	1.11
U30	-0.27	0.16	1.26
T9	-0.29	0.16	0.70
U23	-0.32	0.17	0.72
U12	-0.37	0.17	0.70
U16	-0.37	0.17	1.31
T1	-0.51	0.17	0.69
T13	-0.68	0.17	0.38
U2	-0.87	0.17	1.80
U20	-1.04	0.17	0.43
T4	-1.07	0.17	0.59
T6	-1.22	0.17	0.73
U19	-1.30	0.17	1.79
U26	-1.30	0.17	0.64
U24	-1.69	0.17	1.47
U11	-1.78	0.17	1.80
U25	-2.03	0.18	2.27
Mean	0.00	0.16	1.01
SD	0.90	0.00	0.46

Note. RMSE: .16; Adj. SD: .88; Rater separation: 5.33; Reliability: .97; Fixed (all same) chi-square: 1296.6; d.f.: 44; significance: .00

Table 15. Differences in Severity on Accentedness

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
T11	1.68	0.12	0.85
U30	1.52	0.12	2.46
U18	1.43	0.11	1.61
U23	1.07	0.10	0.87
U9	0.97	0.10	1.44
U6	0.94	0.10	1.58
U19	0.87	0.10	1.13
U7	0.81	0.10	0.44
U28	0.60	0.09	2.27
U4	0.54	0.09	0.61
T8	0.50	0.09	0.63
U5	0.50	0.09	0.69
U20	0.47	0.09	0.91
U26	0.30	0.09	0.63
U12	0.27	0.09	0.50
T1	0.23	0.09	1.07
U16	0.12	0.08	2.06
U1	0.06	0.08	0.66
U24	0.02	0.08	0.82
U32	-0.01	0.08	1.41
U10	-0.05	0.08	1.54
U27	-0.07	0.08	0.65
T13	-0.01	0.08	0.59
U13	-0.12	0.08	0.72
U21	-0.13	0.08	0.50
U11	-0.16	0.08	0.83
U17	-0.21	0.08	1.37
U15	-0.21	0.08	1.80
U14	-0.25	0.08	0.76
T5	-0.28	0.08	0.69
T7	-0.29	0.08	0.68
T10	-0.32	0.08	1.21
T12	-0.33	0.08	0.80
U25	-0.46	0.08	1.20
T2	-0.56	0.08	0.40
U29	-0.57	0.08	1.96
U2	-0.59	0.08	1.33
U3	-0.61	0.08	1.04
U8	-0.86	0.08	0.41

Table 15. (cont'd)

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
U31	-0.86	0.08	0.69
T3	-0.88	0.08	1.46
T9	-0.94	0.08	0.57
T6	-1.25	0.09	0.51
T4	-1.36	0.09	1.64
U22	-1.44	0.09	0.55
Mean	0.00	0.09	1.03
SD	0.74	0.01	0.53

Note. RMSE: .09; Adj. SD: .72; Rater separation: 8.19; Reliability: .99; Fixed (all same) chi-square: 2599.4; d.f.: 44; significance: .00

Table 16. Differences in Severity on Comprehensibility

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
T8	1.24	0.08	0.48
U30	0.76	0.08	1.63
U5	0.75	0.08	0.75
U1	0.56	0.08	0.96
U18	0.55	0.08	0.80
T13	0.54	0.08	0.45
U32	0.49	0.08	1.42
U27	0.48	0.08	0.67
U6	0.45	0.08	1.12
U15	0.43	0.08	1.23
U29	0.43	0.08	1.66
U22	0.41	0.08	0.79
U28	0.41	0.08	1.23
U8	0.29	0.08	2.07
U14	0.24	0.08	0.74
T5	0.23	0.08	0.80
U4	0.21	0.08	0.93
U23	0.21	0.08	0.90
U3	0.20	0.08	1.00
U21	0.19	0.08	0.41
U17	0.19	0.08	0.84
T7	0.14	0.08	0.49
T2	0.11	0.08	0.48
U31	0.11	0.08	1.21
U26	-0.02	0.08	0.56
U13	-0.04	0.08	1.03
U11	-0.05	0.08	2.49
U9	-0.10	0.08	2.05
U10	-0.20	0.09	0.81
U2	-0.22	0.09	0.96
T11	-0.25	0.09	0.55
U12	-0.25	0.09	1.05
T10	-0.26	0.09	0.85
T1	-0.28	0.09	0.70
U7	-0.38	0.09	0.57
U19	-0.42	0.09	1.28
U20	-0.43	0.09	0.55
U25	-0.46	0.09	1.80
U24	-0.60	0.10	1.64

Table 16. (cont'd)

Rater ID	Rater severity measure (in logits)	Standard Error	Infit Mean-square
U16	-0.77	0.10	1.05
T6	-0.82	0.10	0.60
T9	-0.82	0.10	0.59
T3	-0.86	0.10	0.83
T4	-0.91	0.10	0.71
T12	-1.48	0.12	1.41
Mean	0.00	0.09	1.00
S.D.	0.53	0.01	0.48

Note. RMSE: .09; Adj. SD: .52; Rater separation: 6.04; Reliability: .97; Fixed (all same) chi-square: 1473.1; d.f.: 44; significance: .00

It is important to note that the rater severity estimates in the three tables are meaningful only within the context of each individual analysis and are not comparable across the three measures. In other words, the analysis tells us that the difference in severity between U5 and U18 (1.85 logits – 1.83 logits = 0.02 logits) on the oral proficiency rating is smaller than the difference between U18 and U27 (1.83 logits – 1.55 logits = 0.28 logits). However, the analysis does not tell us about whether U5 was more severe on ratings of oral proficiency than on accentedness (severity = 0.05 logits) or comprehensibility (severity = 0.75 logits). This is because the three FACETS analyses were run separately and thus three independent severity scales were constructed.

While it is apparent from the tables that raters differ in their severity estimates, whether these differences in severity are meaningful or not cannot be determined from the tables alone. To determine this, three statistics FACETS provided were examined. These included the rater separation index, the reliability of the rater separation index, and the fixed (all same) chi-square. These three statistics are shown at the bottom of Tables 14, 15, and 16.

The rater separation index is the ratio of the adjusted standard division (Adj. SD) of the rater severity estimates to the root mean-square estimation error (RMSE), a statistical average of the standard errors of measures (Linacre, 2010). When the raters are equally severe, the standard deviation of the rater severity estimates will be equal to or smaller than the mean estimation error of the entire data set and result in a separation index of 1 or less. Results of the analyses show that the separation index for oral proficiency is 5.33, for accentedness is 8.19, and for comprehensibility is 6.04. These results suggest that the variances among the 45 raters are substantially more than the error measurements, with the largest variance observed in the accentedness ratings, and that raters were not equally severe in each of the three separate measures.

The second statistic, the reliability of rater separation, indicates how well the analysis reliably distinguishes among the raters in terms of their levels of severity. The most desirable result, in this case, is to have a reliability of rater separation of 0.00, because this indicates that the analysis is not able to distinguish rater severity reliably; that is, raters are interchangeable. Results of the analyses, however, show that the reliability indexes were very high for all the three measures (.97 for oral proficiency and comprehensibility, and .99 for accentedness), denoting that there were statistically significant differences in the levels of severity among the raters.

The third statistic, the fixed (all same) chi-square, tests the null hypothesis that the measures of the elements in a facet (in this case, the severity of the 45 raters) are statistically the same, except for measurement errors. For oral proficiency, the result of the chi-square test indicated that the average levels of severity the entire rater group exercised were significantly different, $\chi^2(44, N = 45) = 1296.6, p < .001$. For accentedness, the average levels of severity the entire rater group exercised were also significantly different, $\chi^2(44, N = 45) = 2599.4, p < .001$.

Similarly, the average levels of severity the entire rater group exercised on comprehensibility ratings were also significantly different, χ^2 (44, N = 45) =1473.1, p < .001.

In addition to the three statistics for determining the differences in rater severity, FACETS also provides two fit statistics, infit and outfit, that can be used to monitor quality control for the assessment system and to judge rater consistency (Engelhard & Myford, 2003). The infit mean-square is a weighted mean-square residual, that it, the average difference between actual scores and the estimated scores provided by the model, whereas the outfit mean-square is the unweighted mean-square residual and is more sensitive to extreme ratings. The mean-square fit statistics indicate the amount of distortion of the measurement system and has an expected value of 1.0 (Linacre, 2010). Raters whose fit statistics are much larger than 1.0 rate inconsistently and unpredictably and their ratings exhibit more variations. Contrastively, raters whose fit statistics are far below 1.0 rate too consistently and do not well distinguish examinee performances. As recommended by Linacre (2010), the outfit problems are less of a threat to measurement than infit ones and infit statistics are preferable for reporting purposes.

The raters' infit statistics indicate the degree to which each rater is internally consistent in his or her ratings. The fourth column in Tables 14, 15, and 16 details the infit statistics for each rater in each measure. It should be noted that there is no fixed cutoff or rule for which infit statistics are acceptable for each rater; often times, such decisions are made depending upon the targeted use of the test results. Linacre (2002) suggests that the range of infit mean squares between 0.5 and 1.5 is practically useful. Specifically, the infit statistics equal or greater than 1.5 indicate underfit, or too much unpredictable ratings from the Rasch measures, whereas the infit statistics of 0.5 or less suggest overfit, or too predictable ratings from the Rasch measures. Due to the low-stakes nature of the current research study, Linacre's recommended range was considered appropriate for assessing rater consistency and thus was adopted.

Applying the abovementioned standard to Tables 14, 15, and 16, the analyses of rater-model fit showed that, for the most part, raters were internally consistent (over 73% of the raters were consistent in each of the three measures). Some raters had infit statistics falling beyond the acceptable range (12 raters in oral proficiency, 12 raters in accentedness, and another 12 raters in comprehensibility). Among these inconsistent raters, T2 appeared to show rating inconsistencies in all three measures, with infit mean-square of 0.45 for oral proficiency, 0.40 for accentedness, and 0.48 for comprehensibility. These infit statistics suggest that T2's ratings exhibited a lack of variability in comparison to the rest of the raters' ratings. This lack of variation in ratings means that T2 had restricted the range in her ratings by not using all the rating categories included in the rating scales.

In addition to T2 who showed substantial rating inconsistencies and central tendency effect in her ratings, five raters appeared to rate inconsistently in two of the three measures. Rater U28 had infit mean-square indices of 1.61 for oral proficiency and 2.27 for accentedness. These infit statistics suggest underfit, or highly unpredictable ratings. Raters T13, U9, U11, and U25 were inconsistent in their ratings of oral proficiency and comprehensibility, but not accentedness. T13 showed evidence of overfit (infit mean-squares = 0.38 for oral proficiency and 0.45 for comprehensibility). U9, U11, and U25 showed varying degrees of underfit (infit mean-squares between 1.66 and 2.49). Taken together, these results suggest that ratings of the two inconsistent teacher raters (T2 and T13) tended to be less variable, while ratings of the four inconsistent undergraduate raters (U9, U11, U25, U28) tended to be highly unpredictable.

It needs to be noted that, for the purpose of this study, all the 45 raters were included in the FACETS analyses, regardless of rater fit. Generally, in validation studies, misfitting raters are omitted one by one from the analysis in order to improve the overall model fit (Engelhard & Myford, 2003; Linacre, 1989). The reason for the inclusion of all the raters was because the main purpose of this study was to compare the rating differences between the two rater groups rather than to pinpoint inconsistent raters. Additionally, the undergraduate raters were not systematically trained and thus unpredictable ratings were expected at the outset of the study. Since the paramount Rasch model assumption of unidimensional measurement has been checked and met, I considered that there was no need to eliminate any misfitting rater from the analyses, although the FACETS results should be interpreted with caution.

Comparison of Rater Groups

To answer the first research question regarding the differences in overall severity between the ESL teachers and the American undergraduate students when they evaluated potential ITAs' oral proficiency, accentedness, and comprehensibility, I compared the average severity measures of the rater groups in each rating category. Results of the comparison are presented in Tables 17, 18, and 19.

Table 17 shows whether the ESL teachers tended to rate any more severely or leniently on average than the undergraduate raters when they evaluated the examinees' oral proficiency. The fixed (all same) chi-square tests the null hypothesis that the rater groups can be thought of as equally lenient after allowing for measurement errors. Results of the chi-square test indicate that the rater groups did not differ in the average levels of severity they exercised when evaluating the examinees' oral proficiency, $\chi^2 = (1, N = 2) = 3.2, p = .07$.

Table 17. Rater Group Measurement Report on Oral Proficiency

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model *SE*
ESL teachers	47040	1092	43.1	-0.05	0.05
Undergraduates	113770	2688	42.3	0.05	0.03
M	80405.0	1890.0	42.7	0.00	0.04
SD	47185.2	1128.5	0.5	0.07	0.01

Note. Fixed (all same) chi-square = 3.2; *df* = 1, significance = .07

Table 18 shows whether the ESL teachers tended to rate any more severely or leniently on average than the undergraduate raters when they evaluated the examinees' accentedness. Results of the chi-square test indicate that the rater groups differed significantly in the average levels of severity they exercised when evaluating the examinees' accentedness, $\chi^2 = (1, N = 2) = 67.6, p < .001$. The results suggest that the undergraduate raters as a whole tended to rate more harshly on accentedness than the ESL teachers did.

Table 18. Rater Group Measurement Report on Accentedness

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model *SE*
ESL teachers	5727	1092	5.2	-0.12	0.02
Undergraduates	16051	2688	6.0	0.12	0.02
M	10889.0	1890.0	5.6	0.00	0.02
SD	7300.2	1128.5	0.5	-0.16	0.01

Note. Fixed (all same) chi-square = 67.6; *df* = 1, significance = .00

Table 19 shows whether the ESL teachers tended to rate any more severely or leniently on average than the undergraduate raters when they evaluated the examinees' comprehensibility. Results of the chi-square test indicate that the rater groups differed significantly in the average levels of severity they exercised when evaluating the examinees' comprehensibility, $\chi^2 = (1, N = 2) = 75.4, p < .001$. The results suggest that the undergraduate raters as a whole tended to rate more harshly on comprehensibility than the ESL teachers did.

Table 19. Rater Group Measurement Report on Comprehensibility

Rater group	Observed raw score	Observed count	Observed raw score average	Average severity measure (in logits)	Model *SE*
ESL teachers	3933	1092	3.6	-0.13	0.02
Undergraduates	11090	2688	4.1	0.13	0.01
M	7511.5	1890.0	3.9	0.00	0.02
SD	5060.8	1128.5	0.4	0.18	0.01

Note. Fixed (all same) chi-square = 75.4; *df* = 1, significance = .00

In order to provide more evidence regarding the differences in overall severity between the two groups of raters when they evaluated potential ITAs' oral proficiency, accentedness, and comprehensibility, I carried out three separate Mann-Whitney *U* tests. Mann-Whitney U test is a non-parametric test used to compare whether two group means are equal or not and is more appropriate than an independent samples *t* test for small samples that are not normally distributed, which is the case in the current study.

Results of the Mann-Whitney *U* tests indicated that there was no significant difference in severity between the ESL teachers and the undergraduate raters on their ratings of oral proficiency, U(43)= 180.50, Z = -.689, *p* = .49. However, significant differences were found in the comparisons of accentedness and comprehensibility. The undergraduate raters were significantly more severe in the ratings of accentedness than the ESL teachers, U(43)= 124.50 , Z = -2.091, *p* < .05. The undergraduates were also significantly more severe in the ratings of comprehensibility than the ESL teachers, U(43) = 125.0, Z = -2.079, *p* < .05.

To summarize, the Mann-Whitney *U* tests and the FACETS analyses yielded converging results, suggesting that the undergraduate raters were significantly more severe on their ratings of accentedness and comprehensibility than were the ESL teachers. However, the rater groups did not differ in severity on their ratings of oral proficiency.

CHAPTER 5: FINDINGS AND DISCUSSIONS OF QUALITATIVE DATA

Presentations of Written Comments

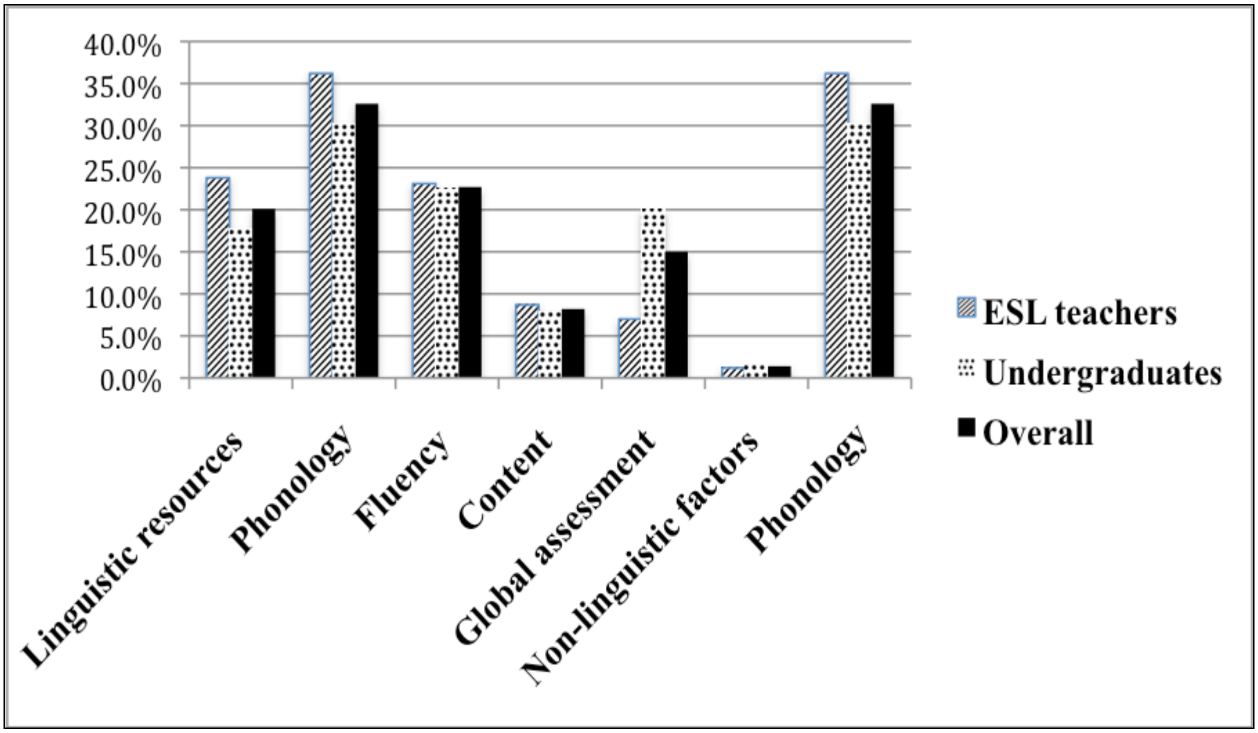
To answer the second research question regarding factors that draw raters' attention when evaluating examinees oral proficiency, the written comments were analyzed both qualitatively and quantitatively. Examples of the written comments coded for each coding category is presented in this section to illustrate the specific aspects of L2 speech raters commented on and how these factors influenced raters' judgments. The quantitative analysis of the written comments includes the tallies of the coded categories and the comparisons of the mean frequencies of coded categories across rater groups. These analyses were performed to determine whether the ESL teachers and the undergraduates differed in the rating criteria they employed.

Table 20 reports the frequency counts of the six coded main categories commented by the rater groups. Figure 4 illustrates graphically the proportions of the written comments coded for each main category. Phonology accounted for the largest group of comments, whereas the non-linguistic factors were commented on by the raters the least. The ESL teachers made larger numbers of comments on phonology, linguistic resources, and fluency, and less on content and global assessment. The undergraduates made larger numbers of comments on phonology, linguistic resources, fluency, and global assessment, and less on content. Both groups made very few comments pertaining to the non-linguistic factors.

Table 20. Frequency Counts of Written Comments across Rater Group

	Linguistic esources	Phonology	Fluency	Content	Global Assessment	Non-linguistic factors	Total
ESL teachers	393	597	381	144	115	20	1,650
Undergraduates	474	806	598	210	531	39	2,658
Overall	867	1,403	979	354	646	59	4,308

Figure 4. Proportion Distribution of Comments Coded for the Main Categories



Linguistic resources. The first main category, linguistic resources, consists of four subcategories: grammar, vocabulary, expressions, and textualization.

Grammar. Both the ESL teachers and the undergraduates made a substantial number of comments related to grammar. Although raters commented in terms of both the accuracy and complexity of grammar, accuracy received more attention. Specifically, raters referred to the examinees' use of number agreement (Example 1), verb tense (Example 2), English articles (Example 3), and referred less frequently to singular/plural marking, prepositions, and the use of compound-complex sentences. In addition to pinpointing the grammar errors the examinees made, some raters attended to the global quality of grammar accuracy and its positive or negative impact on intelligibility and comprehensibility. Several raters commented that inaccurate grammar structures hindered comprehensibility, whereas others noted that responses with inaccurate grammar structures were still highly comprehensible (Example 4). Very few comments dealt with the range of grammar structures used (Example 5).

Example 1: Number agreement is a problem. (T6)

Example 2: The speaker used present tense for the past. (T7)

Example 3: Speaker did not always use articles, which hindered understanding. (U28)

Example 4: Her English was not so good and she didn't construct her sentences very well, but she was definitely understandable. (U27)

Example 5: She used good range of structures (e.g., use of passive voice). (T6)

Vocabulary. Vocabulary was another subcategory frequently referred to by both groups of raters. Raters repeatedly commented on the accuracy or precision of word choices (Examples 6-7), and the use of sophisticated or advanced vocabulary (Example 8). A few raters pointed out the examinees' inability to use correct word form and its impact on intelligibility (Example 9).

Example 6: Explanations not always clear due to word choice. (T12)

Example 7: Very precise, impressive use of words. (U14)

Example 8: The speaker displays sophisticated vocabulary. (T6)

Example 9: The speaker chose the wrong words and word forms in a couple of key areas; that was distracting at best and confusing at worst. (T7)

Expressions. The subcategory *expressions* refers to the examinees' use of English phrases or idioms. Unlike the large number of comments on grammar and vocabulary, few raters commented on the accuracy and complexity of the expressions examinees used. Some comments were related to inaccurate or inappropriate expressions (Example 10) while others were associated with the appropriateness of the phrases examinees chose to convey certain meanings (Example 11).

Example 10: It's hard to understand this examinee because there are some awkward expressions. (T3)

Example 11: The end thought is not phrased in the appropriate manner for a suggestion, like "You can buy your souvenir here". (U20)

Textualization. Textualization refers to the use of connectives, cohesive devices, and discourse markers. Several comments were related to the examinees' abilities to use these devices to tie ideas together smoothly and clearly (Example 12) and to establish the links raters needed in order to comprehend what was being said (Example 13).

Example 12: There is no strong use of cohesive devices to link all the ideas together. (T2)

Example 13: Over-reliance on "and" as a coordinator, "one long and … and … and …," which results in somewhat unsophisticated speech and suggests weaker discourse comprehensibility. (T5)

Phonology. Raters as a whole made the largest number of comments on phonology. This main category comprises four subcategories: pronunciation (including the segmental articulation of vowels and consonants), intonation, rhythm and stress, and accent. The single subcategory, *pronunciation*, consists of more than half of the comments coded for phonology. Approximately one third of the comments were on the examinees' accent.

Pronunciation. The ESL teachers and the undergraduates both made frequent comments concerning the examinees' pronunciation problems, either globally or locally (Examples 14-15).

The most jarring errors, they wrote, included the articulation of the consonants “th”, “l”, “r” and “n”, the insertion of epenthetic vowels (Example 16), inaccurate pronunciation of consonant clusters, and the unreleased final plosives such as “p”, “t”, “k,” (Example 17). Some raters attempted to make inference on the causes of these errors and suggested that they might result from the interference of the examinees’ first languages (Examples 18-19). The impact of wrong pronunciation on comprehensibility was also addressed (Example 20).

Example 14: I gave her a 30 on the SPEAK rating because of her poor pronunciation. It was hard to understand which words she was trying to pronounce. (U15)

Example 15: Her pronunciation was especially strong on the “r” sound and like the word “lobby.” It was very distracting. (T5)

Example 16: Speaker struggles with “k” sound by adding “ah” to the end of it, like “bookahstore.” (U14)

Example 17: Pronunciation problems with dropping the final “p” and “t”, like “get.” (T8)

Example 18: The examinee might be an Arabic speaker because he has the typical Arabic RRRs sounds. (T3)

Example 19: Sounds like some Chinese speakers. Speaker struggles with the “ee” sounds, like in “street.” (U12)

Example 20: The segmental problems here are significant enough to require a patient listener. (T2)

Intonation. In terms of intonation, raters commonly referred to the accuracy and nativeness of the examinees’ intonation patterns and how they impacted comprehensibility (Example 21). Similar to the comments on pronunciation, raters made inferences on the potential impact of the examinees’ first languages on their intonation patterns (Example 22). Flat intonation patterns and the inability to divide the speech into meaningful thought groups were also commented on (Example 23-24).

Example 21: This speaker’s intonation sounds very different from what you would expect from a native speaker and makes it hard to follow. (U18)

Example 22: The rhythm and prosody are very distracting, sounds like a Chinese speaker. (T3)

Example 23: The intonation is off—monotonous and oddly inflected. (T8)

Example 24: The thought groups were not clear, which means that he needs to work on his intonation. (U1)

Rhythm and stress. Raters commented on the stress patterns within individual words (Example 25-26) as well as at the sentence level, namely, the rhythm of the speech (Example 27). Wrong stress patterns were considered distracting to the listeners and had a negative impact on intelligibility and comprehensibility (Example 28-29). The naturalness of the stress patterns was also mentioned in several comments (Example 30).

Example 25: I couldn’t catch several stressed words. (T6)

Example 26: The speaker stresses unnatural syllables (U20)

Example 27: The rhythm is a little bit off. (T3)

Example 28: The speech doesn’t generally follow the sentence stress patterns typical of English. (T2)

Example 29: Word stress patterns are distracting and make comprehensibility low. (T3)

Example 30: Speech sounds funny because the speaker stresses unnatural words. (U7)

Accent. Many raters commented on the examinees’ accent in terms of its impact on comprehensibility. Some examinees’ accent was judged to have an adverse impact on comprehensibility (Example 31), whereas others considered accented speech was still comprehensible (Example 32). Several undergraduates raised the concern of being taught by examinees who had a heavy accent (Example 33).

Example 31: Very strong accent and hard to understand her most of the time. (U2)

Example 32: Accent was strong, but did not affect comprehensibility. (T6)

Example 33: His accent was so strong, being taught by this person would be very difficult. (U5)

Fluency. Approximately one fifth of the comments were coded under the main category of fluency. Fluency encompassed the flow and smoothness of the examinees' speech. Several raters pointed out specific aspects of fluency, such as the pausing features (hesitation and fillers), the speech rate of the speakers, repetition, and self-repairs. Other times raters evaluated the speech samples based on the global quality of fluency.

Pauses. A substantial number of comments within the fluency category relates to the pausing features of the speakers. Unnatural pausing patterns of the speakers were considered to have negative impacts on comprehensibility (Example 34). The use of filled pauses, such as "uh" and "er," and stuttering and stumbling were also sources of concerns. Raters considered that excessive or unnatural use of fillers might have a detrimental effect on the flow of speech (Examples 35). A few raters made inferences about the reasons for the unnatural patterns of pauses and hesitations in the examinees' speech. They speculated that the examinees might be planning the language or organizing their ideas which resulted in unnatural pauses (Example 36). Raters tended to be sympathetic in situations where they considered that the excessive use of pauses was a result of cognitive planning or nervousness rather than an authentic reflection of the speaker's oral proficiency (Example 37).

Example 34: Her lack of pauses at the end of each utterance made me hard to follow (T1).

Example 35: The use of fillers such as “uh”, “um”, and “er” between words was a bit distracting. (U24)

Example 36: The pauses here sound to be pauses of someone thinking, not struggling to find words or anything. (T3)

Example 37: He stumbled excessively while speaking. I believe this wasn’t because he couldn’t think of a word that he wanted to use but merely because he was nervous. Although this was slightly annoying, I gave him a passing grade because I believe it was not because he had a weak grasp of the language that caused him to stumble over his words. (U3)

Repetition and repair. Repetition and repairs are specific types of disfluency that appeared to be jarring to the raters’ ears. Raters commented on the occurrences of word repetitions, false starts, and self-repairs, and their negative impacts on comprehensibility (Examples 38). Raters also discussed the challenges repetitions and repairs posed for speech processing. Some raters thought that unnatural repetitions and self-repairs could damage the fluidity of speech and made it difficult for the examinees to complete tasks in the time allotted (Example 39). Several raters commented on repetitions of individual words and phrases and felt that such occurrences were distracting to the listeners (Example 40). Lastly, the repetitions of similar grammar structures and ideas were viewed negatively (Example 41).

Example 38: Some hesitations, lots of repetitions, which lower comprehension with all the repetitions. (U20)

Example 39: Some excessive repair prevents her from completing the task or getting reasonably far enough in time. (T2)

Example 40: Lots of repetitions of words, verb phrases—very distracting. (T3)

Example 41: The structures were often repeated though there were only a few errors (e.g., come left). There were also repetitions of ideas and words with some minor errors. (T6)

Speech rate. Many raters noted the relationship between intelligibility and the speed of speech. Some examinees were judged to speak too slowly (Example 42) while others too fast (Example 43). In any case, an unnatural speech rate was judged to have a negative impact on intelligibility. However, raters did not want to penalize the examinees because of their speech rate because they felt that some examinees might be pressured by the response time and were forced to speak faster than they normally would. Raters considered that the examinees' unnatural fast speech that resulted in low intelligibility might not be an accurate representation of the examinee's speaking proficiency (Example 44). Several raters, mostly the undergraduates, showed a preference for a slower speech rate. They assumed that slower speech could improve intelligibility, particularly in cases where strong accent was present (Examples 45-46). They also suggested that clear and slow speech was important for a foreign TA (Example 47).

Example 42: If he wasn't speaking so slowly, he'd be easier to understand. (U5)

Example 43: He speakers a bit too fast. (U2)

Example 44: He may be pressured by time, I'm not sure, but his fast speech makes him really hard to understand. … I feel bad giving this one a low score because I have a feeling that he wouldn't talk like this normally. If he talked slower and more deliberately I would have understood him fairly well. (T5)

Example 45: The accent was slightly thick. However, the speaker was speaking slow enough that I understood what she was saying. (U4)

Example 46: Speaker just speaks too fast. If she slowed down some, she would be pretty good. Slowing down will also lesson the intensity of her accent. (U22)

Example 47: I could comprehend it, but the way he was saying it, in short, fast sentences was what made it hard. If he was talking like that in class, I think it would be difficult for students to understand him. It's best to speak slowly and clearly. (U25)

Content. When making references to the content of the responses, raters commented on three main aspects: task fulfillment, the ideas the examinees produced, and the organizations of the information produced.

Task fulfillment. Task fulfillment was evaluated in terms of whether the task requirements were performed competently (Example 48) and whether the responses stayed on topic (Example 49). Several raters made inferences with regard to the reasons for incomplete tasks, and pointed out the potential causes such as slow speech rate or response-time constraints (Example 50). One rater showed a preference for a score of 45 on the SPEAK rating scale for responses that showed good task fulfillment but had a few linguistic problems (Example 51).

Example 48: Task seems to be performed competently. (T6)

Example 49: The speaker was on topic although the task was not completed. (U1)

Example 50: He spoke too slowly and carefully. So he was unable to finish the task in time. (T7)

Example 51: I would have liked to have given this person a 45 if that existed because she accomplished the task very well. But her pronunciation and word choice were very distracting so I went with a score of 40. (T5)

Ideas. Raters commented on the content of the responses both in terms of the number of ideas produced (Example 52) and the way the information was delivered (Example 53). They also commented on the ideas according to their clarity (Example 54). Another point made by several raters associated with the examinees' readiness to be an ITA (Example 55). Finally, raters also discussed the difficulties in comprehending the information reported in the responses (Example 56).

Example 52: The speaker repeated the same information; very redundant. (T1)

Example 53: The speaker provides lots of information and it's very naturally stated. (U20)

Example 54: The speaker's thought groups are not always clear, but she seems to interpret information competently. (T6)

Example 55: He explained things perfectly and is organized. He would be a good TA for that reason. (U15)

Example 56: It was hard to catch several ideas. (T6)

Organization. Raters evaluated the content of the responses in terms of the organization of the responses and its impact on comprehensibility (Examples 57). Raters indicated that poor organization could hinder comprehensibility even though individual words were intelligible (Example 58). Several raters attempted to make inferences regarding the reasons for poor organization and suggested that the causes might be due to poor grammar structures or lack of background information on the topics (Examples 59-60).

Example 57: The organization of the response requires significant listener effort to decipher the message. (T2)

Example 58: The organization and flow make it hard to follow what the speaker is trying to say, but I was able to understand the individual words the speaker was saying. (U6)

Example 59: Her organization was slightly hard to follow, maybe because she mixed different verb tenses in the same sentence. She often forgot verbs too. (U3)

Example 60: His organization was bad. I guess maybe because this is an impromptu kind of test where they aren't allowed to think of any background information on the topic they are to speak about before hand, so I can't blame him. (U3)

Global Assessments. From time to time, raters made global assessments of examinees' speech (Example 61). They made frequent references to the holistic qualities of the speech samples, specifically on the comprehensibility and intelligibility of the speakers (Examples 62-63). A few raters evaluated the performances in terms of the test purpose—that is, to provide an indication of the examinees' readiness to be an ITA (Examples 64-65).

Example 61: Overall speech was very good and very easy to understand. (U14)

Example 62: Her comprehensibility could be worked on, but good overall. (U31)

Example 63: The overall intelligibility of the speech was not high. (U31)

Example 64: One thing that influenced me was that I could imagine him effectively speaking in front of a class in America. (T7)

Example 65: The speaker is not qualified for a TA because he sounds muffled when speaking English. (U29)

Non-linguistic factors. Relatively few comments were related to non-linguistic issues. These included the examinees' test-taking strategies (Examples 66), being nervous or confident during the test (Examples 67-68), and voice quality (Examples 69-70).

Example 66: Nothing linguistic was in his way—just a bad test-taking strategy. (T3)

Example 67: He sounded nervous which may have affected his fluency. (T10)

Example 68: The examinee sounded somewhat confident. (T10)

Example 69: Tone of his voice is distracting, sounds somewhat like shouting. (T5)

Example 70: His voice is so deep that you can't really understand what he is saying. (U31)

Comparison of Written Comments

Since one of the main purposes of this study was to compare the types of rating criteria raters attended to across rater groups, I tallied the coded data and computed the percentages for each rater and for each code both at the main category and subcategory levels for quantitative data analysis. Due to the imbalanced numbers of raters across rater groups, when performing statistical analysis, I used the percentages of comments each rater made for each coded category, instead of the raw frequency, to perform cross-groups comparisons. The tests of normality indicate that the distribution of the percentages calculated for each rater for each code did not meet the statistical assumptions of parametric tests. Thus, I used the nonparametric tests, the Mann-Whitney *U* tests, to compare the coded data both at the main category and the subcategory levels.

Main categories. Table 21 reports the descriptive statistics for the six coded main categories, including the mean and standard deviation for each group. A scrutiny of the table indicates that many of the coded data show high standard deviations. The variability was resulted from the variations in the proportions of comments each rater made in each coded category.

Results of the Mann-Whitney *U* Tests indicate that the two groups did not differ in the percentages of comments coded for linguistic resources, phonology, fluency, content, and non-linguistic factors. The only significant difference observed was associated with raters' global assessment of the examinees' oral proficiency, U(43)=62.5, Z= -3.65, *p* < .001, suggesting that the undergraduate raters commented significantly more frequently on the global quality of the responses than the ESL teachers did.

Table 21. Descriptive Statistics and Mann-Whitney U Tests for the Main Categories

Main categories	Overall (*N* = 45)		ESL teachers (*n* = 13)		Undergraduates (*n* = 32)		Z-value	*P*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Linguistic resources	18.1	10.7	20.6	10.9	17.1	10.7	-1.03	.29
Phonology	34.1	16.6	40.6	17.8	31.5	15.7	-1.62	.10
Fluency	21.4	11.1	23.3	9.5	20.7	11.7	-5.26	.59
Content	7.1	6.6	8.6	7.7	6.5	6.1	-7.54	.45
Global assessment	17.6	18.6	5.7	5.1	22.4	20.0	-3.64	<.001
Nonlinguistic factors	1.7	2.7	1.2	1.4	1.8	3.1	-0.02	.97

Linguistic resources. In terms of the subcategories within the main category of linguistic resources, raters as a whole made nearly equal proportion of comments to the grammar (*M* = 44.3%) and vocabulary (*M* = 44.0%) aspects of the examinee responses (see Table 22). In contrast, raters commented less frequently on the expressions and textualization. Results of the Mann-Whitney *U* Tests indicate that the rater groups differed significantly in the proportions of comments they made on the expressions the examinees produced, U(43) = -2.87, Z = 121.5, *p* <

.001. The ESL teachers appeared to comment more frequently on the expressions used by the examinees than the undergraduates did.

Table 22. Descriptive Statistics and Mann-Whitney U Tests for Linguistic Resources

Subcategories	Overall		ESL teachers		Undergraduates		Z-value	*P*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Grammar	44.3	25.5	44.9	16.9	44.1	28.6	-0.20	.84
Vocabulary	44.0	27.2	38.7	19.4	46.2	29.8	-0.47	.63
Expression	2.9	6.2	6.6	7.9	1.4	4.6	-2.87	<.001
Textualization	6.5	13.4	9.7	17.2	5.2	11.6	-1.59	.11

Phonology. The main category phonology consisted of the largest proportion of the written comments. Raters as a whole commented most frequently on the pronunciation aspect of the speakers' speech, suggesting that the examinees' articulations of vowels and consonants were salient features of L2 speech for the listeners. Table 23 displays that more than one third of the comments were related to the examinees' foreign accent (*M* = 34.5%), whereas less than 10% of the comments were related to either the intonation or rhythm and stress aspects of the speech samples.

Results of the Mann-Whitney *U* Tests indicate that the rater groups differed significantly in the proportions of comments they made for intonation, U(43) = 54.0, Z = -4.23, *p* < .001, rhythm and stress, U(43) = 87.0, Z = -3.42, *p* < .001, and accent, U(43) = 91.5, Z= -2.91, *p* <.001. The results indicate that the ESL teachers commented more frequently on the intonation and rhythm and stress aspects of the examinees' speech, while the undergraduates commented more frequently on the examinees' foreign accent.

Table 23. Descriptive Statistics and Mann-Whitney U Tests for Phonology

Subcategories	Overall		ESL Teachers		Undergraduates		Z-value	*P*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Pronunciation	52.0	23.7	53.8	15.3	51.2	26.6	-0.25	.80
Intonation	6.3	10.8	15.2	12.7	2.6	7.4	-4.23	<.001
Rhythm and stress	7.3	14.9	12.9	10.9	5.0	15.8	-3.42	<.001
Accent	34.5	26.1	17.9	17.4	41.2	26.2	-2.91	<.001

Fluency. Table 24 reports the results for the main category of fluency, the second largest group of comments raters made. As the table shows, the comments as a whole coded were concerned with, in descending order, (a) overall fluency, (b) pauses, (c) repetition and repair, and (d) speech rate. Results of the Mann-Whitney *U* Tests indicate that the rater groups did not differ in any of comments they made for any of the four subcategories.

Table 24. Descriptive Statistics and Mann-Whitney U Tests for Fluency

Subcategories	Overall		ESL Teachers		Undergraduates		Z-value	*P*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Pauses	28.8	26.4	38.4	24.8	24.9	26.4	-1.88	.06
Speech rate	13.8	15.3	17.1	16.2	12.4	14.9	-1.11	.26
Repetition and repair	15.0	14.6	13.5	9.5	15.7	16.4	-0.11	.90
Overall fluency	37.9	28.9	31.1	27.9	40.8	29.2	-0.96	.33

Content. Relatively few comments were coded for the main category of content. Table 25 shows that when raters commented on different aspects of the content, they commented most frequently on the ideas the examinees produced in their responses (*M* = 40.0%). This was followed by comments made about the organization of the responses (*M* = 28.2%) and task fulfillment (*M* = 10.1%). Results of the Mann-Whitney *U* Tests indicate that the ESL teachers commented significantly more frequently to task fulfillment than the undergraduates did, U(43)=

125.0, Z= -2.67, *p* < .001 (See Table 25). No significant difference was found on the comments coded for ideas and organization.

Table 25. Descriptive Statistics and Mann-Whitney U Tests for Content

Subcategories	Overall		ESL Teachers		Undergraduates		Z-value	*P*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Task fulfillment	10.1	21.6	21.5	26.4	5.5	17.8	-2.67	<.001
Ideas	40.0	43.2	39.7	39.3	40.1	45.3	-0.27	.78
Organization	28.2	39.5	23.4	35.5	30.2	41.4	-0.05	.95

To summarize, raters attended to six major conceptual categories when evaluating the examinees' oral proficiency: linguistic resources, phonology, fluency, content, global assessment, and non-linguistic factors. Within each major category, they paid attention to different aspects of L2 speech. At the main category level, rater groups differed significantly in the proportion of comments they made on global assessment. The undergraduate raters commented more frequently on the overall quality of the examinees' responses than the ESL teachers did. At the subcategory level, rater groups differed significantly in the proportion of comments they made on several different performance features, including (a) expressions, (b) intonation, (c) rhythm and stress, (d) accent, and (e) task fulfillment. The ESL teachers commented more frequently on expressions, intonation, rhythm and stress, and task fulfillment than the undergraduates did, whereas the undergraduates attended more often to the accent of the examinees than the ESL teachers did.

Findings and Discussions of Interview Data

The interview data was analyzed to explore the links between the ratings raters assigned and the factors that influenced their ratings. In addition to the six main rating criteria identified through the written comments presented earlier, the interview data further reveal three major factors that have important bearings on the ratings: (a) experience with accented speech, (b) accent as an important rating criterion, (c) analytic versus holistic ratings.

Experience with accented speech. The ESL teachers and the undergraduates believed that experience with accented speech could help them better understand the SPEAK examinees and lack of familiarity with foreign accents could deter students from comprehending ITA speech. All the ESL teachers reported that they were familiar with a variety of foreign accents and that their experience with foreign accents helped them understand L2 speakers, even those with heavy accents. T6 commented that "My teaching experience has helped me a lot. If I listen to a particular speech, I can usually understand and infer what the speaker's language background might be. I can easily differentiate a Japanese from a Korean, or a Chinese. Understanding their accent is usually not a problem at all." T1, whose first language was Japanese, said that she could easily understand someone with an Asian accent and felt that she might become more lenient because of that reason. "I think it becomes a bit easier for me to understand their speech because I'm so used to these students' pronunciation. I might have become very lenient," she said. Similarly, some teacher raters felt that their accent familiarity might have influenced their ratings toward the lenient side. T4, one of the most lenient teacher raters in the oral proficiency and accentedness rating categories (see Figures 1 and 2), thought that she might not have been as objective as she should have because "I'm exposed to many different accents, so I might be biased. I'm worried that my ratings might be too high because honestly I could understand them all."

Several ESL teachers felt that they should be aware of the difference in the amount of accent exposure experienced raters and naïve undergraduates had when they evaluated ITA speech. T8, who had taught in different countries and been a SPEAK rater for seven years, said that "I've taught for so many years and I probably can identify 25 different accents, like a Swedish, or a Korean. But your 17 years old couldn't." She said that she always reminded herself about her experience with accented speech whenever she rated SPEAK examinees and commented that she would not pass an examinee if she did not think that undergraduates could understand the person. Another rater, T11, who was also very conscious about the undergraduates' limited exposure to accented speech, felt that as a rater, "you have to think that whether your undergraduates can listen to this accent in the class for two hours. Like I said, I can give most of them a 50 because I understand what they're saying, but not the undergraduates. So it's a fine line and you really have to be conscious about it." Interestingly, the two raters, T8 and T11, who were most mindful about the discrepancy in accent familiarity between experienced raters and linguistically naïve undergraduates, were also the most severe teacher raters in the oral proficiency ratings (see Figure 1).

While the teacher raters all reported to have extensive exposure to accented speech, the majority of the undergraduates said that they had limited foreign accent exposure and felt that understanding foreign TAs could sometimes be a challenging task. Twelve undergraduate raters said that they did not have any foreign friend or exposure to foreign accent before they came to MSU. Several undergraduates shared their personal stories about the difficulties they had with their foreign TAs. For example, U21 said that "my math TA was an Indian and it was hard for me to understand him because of his pronunciation. I just had to study by myself." When replying to the question regarding how often he approached his ITA, he said, "I didn't approach him at all because it was not helpful coz' I couldn't understand him anyway." U8, one of the harshest undergraduate raters, complained that "having teachers who don't speak your native language is really hard, especially for a freshman, because you have to focus on two things rather than just one. I mean, you have to understand the language and also the math concepts at the same time." Another undergraduate, U25, reported that she was initially planning to major in math when she first joined MSU, but because she could not understand most of her math TAs during her first year and her classmates "couldn't understand them either." So eventually, "I switched my major", she said.

Accent as an important rating criterion. Though the ESL teachers and the undergraduates both expressed concerns about the impact of accent familiarity on comprehending ITA speech, raters' opinions about whether accent should be considered as an important rating criterion in the evaluation of ITA speech differ within groups and across groups. Within the teacher rater group, the two most severe raters, T8 and T11, emphatically expressed the importance of weighing accent highly when screening ITAs. T8 felt that "You can have somebody who has excellent vocabulary, grammar, but if the accent is causing a problem, then they should get a low score. They can have perfect grammar, but if I can't understand their accent, then they shouldn't be in the classroom." A somewhat less harsh teacher, T7, also suggested accent be one of the most important rating criteria because he thought that "if accent becomes a stumbling block for the students, then they would have to struggle more than what they should in order to understand what the teacher is saying."

However, the majority of the teacher raters thought that accent should not be an important rating criterion. Some teachers suggested that other speech features such as stress, intonation, and pauses were more important than accent. T9, who has taught many Chinese and Arabic students, reported that accent was not an important criterion for him, and "in a lot of ways, flow or fluency definitely are the most important. Then you also look at their stress patterns and their intonation. These are definitely more important." T12, who generally did not consider accent an important factor in judging SPEAK examinees, distinguished a point where he would consider accent in his rating decisions. “If the speaker has a very standard accent, wherever that accent comes from, I usually don’t take that into consideration. But if it’s a very strong accent and I believe that it will hinder undergraduates’ comprehension significantly, then I do.”

The majority of the undergraduate raters expressed their concerns regarding how heavy accent could impact their comprehension of ITA speech and indicated that accent was an important rating criterion for them. U18, one of the harshest undergraduate raters, suggested that accent was an important rating criterion for her and that if she could not “figure out what the speaker was saying, then I don’t think he should be a TA.” Another harsh undergraduate rater, U5, commented that “accent is so important because if you can’t understand their pronunciation, you get very confused.” U31 stated that accent was the most important rating criterion because “if their accent is so heavy, then I’ll have to figure out what they’re trying to say and meanwhile, I’m also trying to understand the concepts they’re teaching me.” Most of the undergraduates agreed that learning new concepts, such as math equations or statistical concepts, was a cognitively challenging task and that having to learn new materials and at the same time trying to understand an unfamiliar foreign accent would adversely influence their learning.

In contrast, few undergraduate raters acknowledged that accent was not an important rating criterion, and other fluency features, such as the pauses or speaking rate, were more important. U2 commented that “I don’t think accent is that important. I think that flow is more important. Like if they stumble, or pause a lot, I think that’s what confuses me the most when I evaluated them.” U13, who had experience tutoring ESL learners and had studied abroad in Spain, said that “I focused more on the vocabulary and the grammar rather than the accent. I felt that rating was more about how they spoke, not their accent. So if they use the right verb tense or appropriate vocabulary, then I would give a 50." Five undergraduates commented on the unnatural speaking rate the examinees had and indicated that a normal or slower speech rate was preferred. For example, U15 mentioned that "I think their speed is a big one. Sometimes I have TAs who speak very fast. Honestly, I prefer a normal pace, not too slow or too fast. But if their accent is too heavy, then it's better to speak slowly."

Analytic versus holistic ratings. One noticeable difference between the ESL teachers and the undergraduates' interview protocols pertains to their approaches to ratings. The ESL teachers as a whole appeared to commented significantly more frequently on specific linguistic features that they took into account while the undergraduates tended to adopt a more global, impressionistic approach to making their rating decisions. Most the ESL teachers pinpointed precise speech features that drew their attention and indicated how they used these features as rating criteria to finalize their ratings. For example, when commenting on the examinees' use of grammar, T1 noticed that "when the examinee made grammar errors, like the third person singular "s", that doesn't obscure meaning, so I won't mark them down. But like the wrong verb tenses, or like they use present tense for past tense, then I would give a lower grade."

As the data indicate, few undergraduates were able to use precise linguistic terms as the ESL teachers did to describe how they made their rating decisions. The majority of the undergraduate raters stated that they based their judgments primarily on whether they could understand a given speaker or they would like the person to be their TAs. U10, for example, felt that making the oral proficiency ratings was not an easy task because she had never rated L2 speakers or used a rating rubric before. She said that she mainly "graded them based on if I want them to be my TA. Like for my math class, I'm not good at math, so I wouldn't want to have a foreign TA simply because I couldn’t understand them.” Likewise, U27, who dropped a math course taught by an Asian TA in his first semester, said that “Well, I just went by how much I could really understand. As long as you can understand what they’re saying, I think that’s the most important. ”

In summary, the interview protocols indicate that raters’ experiences with accented speech had substantial impacts on one’s comprehension of ITA speech and moderate impact on how raters derived their ratings. Most teacher raters, while acknowledging that their accent familiarity had made their comprehension of the examinees easier, suggested that other linguistic features such as overall fluency were more important than foreign accent. In contrast, most undergraduate raters had very limited exposure to foreign accent and believed that accent was important in the evaluation of ITA speech. Finally, the ESL teachers and the undergraduates differ in their overall approaches to rating. Whereas the ESL teachers tended to rate more analytically, the undergraduates were more inclined to making their judgments basing on whether they felt they could understand the speakers or whether they could like the speakers to be their TAs.

CHAPTER 6: GENERAL DISCUSSION

Testing programs that administer high-stakes tests are responsible for delivering tests that are reliable, ethical, and valid. They must do so because their high-stakes tests provide the basis for score interpretations that significantly impact test takers' lives. Testing programs that administer ITA screening exams are no exception to this rule. The tests they administer are ultimately used to decide who can obtain a teaching assistantship, which will ultimately impact not only the test takers themselves, but also the lives of the test takers' family members, the ITAs' students, and the universities that hire the ITAs. Thus, the significance of the use of ITA tests cannot be underestimated.

As outlined at the beginning of the dissertation, the main purposes of this study were twofold. Firstly, I wanted to compare ESL teachers' and American undergraduates' evaluations of ITA speech. I did this by having the raters rate samples of ITA speech on three factors—oral proficiency, accentedness, and comprehensibility. I also wanted to check if both groups rated equally across those three factors, or if one group rated differentially on one or more of these dimensions. If they do evaluate ITAs' speech differently, I wanted to identify why. Secondly, to understand the cognitive processes raters undergo when rating ITAs' speech samples, I investigated to what, in the speech samples themselves, the raters attended. That is, I identified and then compared the factors in the speech samples that drew the raters' attention during the rating process. I assumed that the raters would attend to a variety of linguistic features and overall task performances, the factors toward which the SPEAK scoring rubric guided them. But did they attend to those factors differently? And, moreover, did they attend to other factors not expressed in the rubric? I applied both quantitative and qualitative methodologies to address both of these questions, both of which center on rater effects and rater orientations. And in this study in particular these questions apply across two groups of raters in an ITA testing situation.

In response to the first research question, the results suggest that the rater groups did not differ in the severity they exercised when they evaluated the examinees' oral proficiency. More precisely, results from the between-group comparison indicate that the difference in rater severity between the ESL teachers and the American undergraduates was small and did not reach statistical significance—the rater groups rated the examinees' oral proficiency in a similar fashion, and thus little difference in severity was observed. This finding is backed by the overall results of the multiple quantitative analyses, including the descriptive statistics of the raw scores, the classification of ITAs assignments, the FACETS analyses, and the Mann-Whitney *U* tests.

The result concerning the overall equality in the raters' judgments on oral proficiency contradicts Barnwell's (1989) study that found that untrained raters were harsher than teacher raters concerning their judgments of oral proficiency. On the other hand, the results are consistent with several studies reviewed earlier (Dalle & Inglis, 1989, March; Powers et al., 1999; Saif, 2002), corroborating previous findings that ratings of oral proficiency awarded by untrained, undergraduate students and ESL professionals are similar and related. Thus, this study provides additional support to the argument that the features of an individual's oral performance are potent determinants of raters' judgments of an examinee's overall speaking ability. Thus, despite disparate rating experiences (expert versus inexperienced) and contrasting linguistic backgrounds (varied versus non-varied) across the two groups in this study, the undergraduate raters were found to assign oral proficiency ratings comparable to those assigned by the ESL teachers who had much training and more linguistic experience, just as has been found in prior studies (A. Brown, 1995; Lumley & McNamara, 1995).

However, the results from this study become more complex and intriguing when the two groups' ratings on accent and comprehensibility are considered. There were significant, between-group differences between the two groups' ratings on accentedness and comprehensibility. The undergraduate raters were more severe when they judged the examinees' foreign accents. They also perceived a significantly higher level of difficulty in comprehending the examinees' speech. But these results should not be surprising. Previous work (Bailey, 1984b; Fox & Gay, 1994; Hinofotis & Bailey, 1981; Plakans, 1997; Rubin, 1992; Rubin & Smith, 1990) has indicated that American undergraduates tend to evaluate ITAs' foreign accented speech negatively. For example, Hinofotis and Bailey (1981) and Plakans (1997) both found that poor pronunciation was the most prominent failure (as judged by students) in ITAs' communicative competence. The FACETS analyses reported support such a view and extend it with respect to ratings of comprehensibility in between-group comparisons.

But why do American undergraduates tend to evaluate ITAs' foreign accented speech negatively and, concomitantly, indicate that they have a hard time comprehending such speech? Findings of the interviews lend a hand at understanding this. One possible reason for the between-group difference in severity observed in the ratings of accentedness and comprehensibility pertains to the raters' amount of exposure to, and experience with, foreign-accented speech, as suggested by the interview data. All the undergraduate raters reported that they had very limited contact with nonnative English speakers either during their upbringing or in their circles of friends, whereas the ESL teachers all indicated that they had extensive ESL/EFL teaching experience, contact with nonnative-English speakers, and were familiar with a wide variety of nonnative English accents. As the ESL teachers reported in their interviews, their extensive exposure to an array of diverse English pronunciations from learners of various

L1 backgrounds have enhanced their ability to decipher the meaning conveyed by accented, L2 speech. These results corroborate findings from a large body of previous work in speech perception and on the cognitive processing of L2 speech—work that supports the general claim that the amount of exposure to World Englishes and/or interaction with nonnative speakers can enhance the listening comprehension of those English varieties (Derwing & Munro, 1997; Derwing et al., 2002; Gass & Varonis, 1984; Kang, 2008, 2010; Munro & Derwing, 1994; Powers et al., 1999).

The second research question in this study delved into *why* raters with different backgrounds may differentially rate the speech of ITAs. In particular, with this second research question, I asked if the rater groups attended to different features (or factors) in the speech of the ITAs, and whether this differential attention could explain the observed differences in score assignments. This second research question also addresses the extent to which the ESL teachers and the undergraduates differed in the rating criteria they employed. This is an important area of investigation because all raters should rate language against the same set of criteria (Bachman, 1990; McNamara, 1996). When raters reliably use a common set of criteria against which to judge language, they are providing and operationalizing a common measurement of the test construct (Bachman, 1990). To not do so (if different raters use different judging criteria) presents theoretical problems and construct-validity issues in terms of score comparability (Messick, 1989).

By coding the written comments, I identified six main rating categories the raters reported they employed: linguistic resources, phonology, fluency, content, global assessment, and other, nonlinguistic factors. Concurring with A. Brown et al. (2005), the raters' attention to the first four rating categories (linguistic resources, phonology, fluency, and content) was broken down further. For example, within the linguistic resources category, raters made comments on the examinees' use of grammar, vocabulary, expression, and textualization. Within the phonology category, the examinees' pronunciation, intonation, rhythm and stress, and foreign accent were all sources of attention. As far as fluency is concerned, raters judged the responses based on the repetitions or self-repair patterns and the speech rate of the speakers. In terms of content, raters noted whether the examinees fulfilled the task requirements, the ideas that the examinees produced, and the organization of the responses. Nonlinguistic factors included test-taking strategies, voice quality, and examinees' emotions.

The quantitative comparisons of the written comments and the qualitative analysis of the interview protocols further helped determine the extent to which rater groups differed in the rating criteria they utilized. The results of these separate analyses converged, indicating that the ESL teachers and the undergraduates attended to several aspects of the linguistic dimensions in the examinees' speech differently, as predicted by past research (Chalhoub-Deville, 1995; Elder, 1993). Specifically, the results suggest that the teacher raters commented more frequently on a variety of linguistic features than did the undergraduates. The undergraduates, on the other hand, appeared to evaluate the examinees' oral performances more impressionistically. The interview data reveal that many undergraduates were not familiar with the rating criteria for judging the SPEAK examinees and, thus, they made their rating decisions solely through their appraisal of whether they felt a particular examinee was qualified to be an ITA, or whether they would like the speaker to be their TA—a criterion not on the rating rubric. In either case, the data appear to suggest that undergraduate raters consider their personal feelings, perhaps even their fears, and their possible future experiences as students in ITA classes in judging ITA speech. They may tend to err on the side of caution and be more severe on accent and comprehensibility, regardless of oral proficiency, in anticipation of possibly having the test taker as a teacher in the future.

Bolstering this argument further, the undergraduates provided a substantially larger proportion of comments on the examinees' accents than the ESL teachers did. And consistent with the results of the quantitative data, the rater groups differed in terms of their judgments of the examinees' accents (the undergraduates were more severe on accent). Many undergraduates commented that the examinees' accents were so heavy that they could not understand what the speakers were saying and some mentioned explicitly that they would prefer not to have an ITA with a strong accent. These findings concerning the role foreign accent plays in undergraduates' evaluations of ITA speech again concur with many previous studies (Bailey, 1984b; Bauer, 1996; Bryd & Constantinides, 1992; Derwing & Munro, 2009; Landa, 1988; Rubin, 1992; Rubin & Smith, 1990). Simply put, the undergraduates' judgments of the examinees' oral performances may have been determined by foreign accent to a large extent, and the presence of a foreign accent was viewed more negatively by the undergraduates than by the ESL teachers.

The ESL teachers appeared to comment more frequently on the accuracy and complexity of the expressions the examinees produced. This finding was similar to the results of the study by McNamara (1990) that found that grammar and expression were the most harshly rated criterion by expert raters. Although no other significant difference was found concerning the proportion of comments made regarding other aspects of the linguistic resources, the ESL teachers' comments on the examinees' use of expressions suggest that the ESL teachers considered the ITAs' ability to use accurate and appropriate expressions as important as the use of grammar, vocabulary, and discourse markers. That is, the ESL teachers appeared to value speech more in terms of its technical aspects, and less in terms of its overall accent. This result appears to carry over into the raters' evaluations of the phonological features of the speakers' speech. The ESL teachers indicated that they paid more attention to the examinees' intonation and stress patterns than the undergraduates did, demonstrating again that the experienced ESL teachers rated based on the linguistic aspects of the speech, along with the examinees' overall task fulfillments, while the undergraduates rated the speech based more on what one might call *feel*.

Further evidence of some undergraduates using "feel" to rate speech in this study stems from some undergraduates' comments on speech rate. Contrary to Kang's (2010) finding that undergraduate raters considered ITAs' speech more comprehensible when the ITAs spoke faster (i.e., higher speech rate was associated with higher comprehensibility), many undergraduates in this study commented that faster speech would impair speech intelligibility and increase comprehension difficulties, especially in cases where heavy accents were present. The effect of overall speech rate on perceived comprehensibility by native speakers appears to be very complex (Derwing, 1990; Derwing & Munro, 2001). Most language proficiency scales take a higher rate of speech as evidence of greater fluency in the language. But findings from this study corroborated with Zhao (1997), indicating that a fast speech rate along with a heavy accent in the speech can increase listeners' difficulty in understanding the speaker and is not preferred.

The data in this study suggest that the undergraduates may have lumped many features of the linguistic component under accent, features that the ESL teachers considered separately from a test taker's accent *per se*. Previous research has provided evidence showing that L2 speakers may have difficulty producing the characteristic intonation (Kang, 2008, 2010; Pickering, 2001, 2004) and stress patterns (Juffs, 1990; Kang, 2010) of English due to their L1 having less variation in pitch range or a different way of putting emphasis on words. Among the 28 speakers of this study, ten had an L1 background of Mandarin, ten Korean, and eight Arabic. The majority of these examinees were expected to have a narrow pitch range (as in Mandarin and Korean) or a tendency to stress each word regardless of its role in the discourse structure (as in Mandarin and Arabic), as suggested by previous studies (e.g., Binghadeed, 2008, for speakers of Arabic; Kang, 2010, for speakers of Korean; Pickering, 2001, for speakers of Mandarin). It appears that many ESL teachers picked up on the test takers' narrow pitch ranges and unnatural stress patterns and commented on their impact on comprehensibility. The undergraduates, however, did not comment much on intonation or stress patterns, most likely because they are linguistically less sophisticated than the ESL teachers and were less able to describe such features metalinguistically. The majority of the undergraduates may have attributed their problems in deciphering problematic intonation and stress to the test takers' accents. And this may explain, in part, why the undergraduates awarded higher accent ratings (more accented)—their target of accent was larger than the ESL teachers' target for accent. Furthermore, such an interpretation suggests that the differences in attention paid to various linguistic features may not reflect a difference in what features the raters actually attended to, but a difference in *how* they explained what features they attended to.

This study identified several nonlinguistic factors to which raters attended, including the examinees' test-taking strategies, voice quality, and evidence of confidence or nervousness in the responses. None of these factors have been thoroughly discussed in previous studies (A. Brown et al., 2005; Rubin, 1992), except Winke, Gass, and Myford (in press) who briefly discuss voice quality and nervousness in their study on rater effects. Previous research has demonstrated that nonlanguage factors, such as the speaker's ethnicity (Rubin, 1992), could affect undergraduates' judgments of L2 speech. Nevertheless, the number of comments made by both groups on these nonlinguistic factors was small, suggesting that the linguistic features of the speakers were the predominate constituents of the raters' rating orientations. Yet, these data do suggest that raters attend to more than the linguistic properties of speech, as suggested by Munro (2008) and Winke et al. (in press).

It is important to note that although the complete set of written comments could conveniently be coded into the six main coding categories and their corresponding subcategories, the decision-making processes of the raters appear to vary substantially from person to person, as evidenced by previous research (A. Brown, 2007; A. Brown et al., 2005; Meiron & Schick, 2000; Orr, 2002; Papajohn, 2002). This variation is made apparent by the high standard deviations observed in the proportions of comments raters made on the different rating subcategories. The proportion of comments coded for the subcategory of pauses within the main category of fluency showed a 13.5% difference across rater groups. This margin was one of the most pronounced quantitative differences between the rater groups although no statistically significant difference was found. It is possible that the frequency of the comments made varied so substantially from person to person that the Mann-Whitney *U* Test did not reach statistical significance. Thus, the differences observed could be attributable to just qualitative ones. That is, the wide range of factors raters commented on may stem from individual differences, which corroborate findings of several previous studies on raters orientations (Chalhoub-Deville, 1995; Cumming, 1990; Elder, 1993; Hadden, 1991), suggesting that the rating process is dynamic, complex, and interactive, and varies from individual to individual.

Implications

One pedagogical implication of this study, in addition to those suggested by previous studies, such as providing pronunciation training to L2 speakers (Derwing & Munro, 2005, 2009; Derwing et al., 1998), has to do with the training of undergraduate students with regard to how to listen to accented speech. As reported in several previous studies, many undergraduates have a general tendency to feel anxious about listening to foreign-accented speech due to their limited experience interacting with L2 speakers, or their lack of confidence in their own abilities to communicate or understand foreign TAs (Derwing et al., 2002; Rao, 1995 May). Other studies suggest that even the undergraduates' attitudes toward ITAs play a role in their comprehension of ITA speech (Rubin, 1992; Rubin & Smith, 1990). Empirical research has also verified that accented speech requires extra listener effort and more processing time (Munro & Derwing, 1995b; Schmid & Yeni-Komshian, 1999). It is not fair to say that ITAs are always at fault when there is a communication breakdown in the classroom. On the other hand, research has suggested that successful communication is shared across interlocutors (Derwing & Munro, 2009) and increased familiarity with L2 speech can improve comprehension (Gass & Varonis, 1984). Given these facts, thus, ITA programs should consider not only offering pronunciation instructions to ITAs but also make available training workshops that teach undergraduates how to listen to and process accented L2 speech. These workshops can help reduce undergraduates' anxiety while they listen to or converse with L2 speakers or their ITAs. Even through very limited training, undergraduate students can increase their ability to comprehend accented speech and enhance their willingness to talk with L2 speakers (e.g., Derwing & Munro, 2009; Derwing et al., 2002).

This study also has implications for ITA testing. Isaacs (2008) and Morley (1994) suggested that undergraduates are important stakeholders in the ITA testing context and should be included as a part of the ITA screening process. As detailed earlier, although the ratings of oral proficiency assigned by the undergraduates were comparable to those assigned by the ESL teachers, I found significant differences on the ratings of accentedness and comprehensibility across the rater groups in this study. The undergraduate raters were more severe in terms of accent and comprehensibility judgments. It can be argued that the ESL teachers' judgments of the examinees' oral performances in terms of accent and comprehensibility were more lenient than the undergraduates' because the ESL teachers paid more attention to specific, linguistic features in the speech samples, while the undergraduates tended to base their ratings more on accent (and for the undergraduates, the heavier the accent, the worse the comprehensibility of the speech) and overall *feel.* A main finding of this study is that undergraduates may not be able be act as impartial judges, even with extensive training, because they have something at stake—the possibility to be taught by ITAs who they cannot understand. Contrary to Isaacs (2008) and Morley (1994), this study's results suggest that ITA programs should *avoid* having undergraduates as official raters, but rather use them to check the threshold of what undergraduates may consider as incomprehensible speech. On the other hand, ITA testing program should not underestimate undergraduates' abilities to adapt and comprehend ITAs whose speech falls within that "grey" zone (between what undergraduate raters would call incomprehensible, but what expert ESL teachers would call comprehensible), since, as argued above, research has shown that through very limited training, undergraduate students can increase their ability to comprehend accented speech and willingness to talk with L2 speakers (e.g., Derwing & Munro, 2009; Derwing et al., 2002). Therefore, ITA testing programs should not fear such a gap. Nevertheless, the potential difference in severity between the official ITA testing raters and the undergraduates should still be constantly monitored, carefully evaluated, researched, and controlled.

Limitations

A few limitations need to be addressed. First of all, admittedly, the number of raters in this study is small, and they came from only one single university in the Midwest. It is unknown whether the findings would hold for raters from other geographical regions such as the West coast where the makeup of the student body and the wider community are much more diverse both ethnically and culturally. While it has been shown that the undergraduate raters and the ESL teachers were comparable in their evaluations of oral proficiency but differed in ratings of accentedness and comprehensibility, it is not necessarily the case that raters from different universities or regions can reach judgments congruent with those found here (e.g., Bailey, 1984b, in the southwest; Rubin, 1992, in the southeast).

Secondly, although the construct of *intelligibility* is by all means worth investigating, I only focused on how the two groups of raters were distinct in their ratings of accentedness and comprehensibility. This is due to the fact that a single reliable measure of intelligibility is not available (Isaacs, 2008) and the existing measures of intelligibility, which mainly employ the method of dictation at the word or sentence level (e.g., Munro & Derwing, 1995a; Munro, Derwing, & Sato, 2006), were inappropriate for the current study that focused on examining rater variability by using rating scales. In contrast, Munro and Derwing's (1995a) 9-point rating scale widely used for the evaluations of accentedness and comprehensibility in L2 speech literature appears to be a useful and convenient tool for the purpose of this study, thus it was employed. Future research in ITA testing might want to explore appropriate measures of intelligibility and incorporate intelligibility in the investigation of ITA speech.

Another limitation relates to the differences found in the ratings of accentedness and comprehensibility. Since I informed all raters about the purpose of this study that aimed to compare ESL teachers and undergraduate raters' perceptions of ITA speech, the raters might have been prompted to direct extra attention to the accent feature of the speech samples. Undoubtedly, the undergraduate raters, under some circumstances, might have brought their previous personal experiences with ITAs, either positive or negative, and judged the speech samples differently than they normally would if such experiences did not exist. Rubin (1992) found that even an inaccurate assumption about an L2 speaker's ethnicity could reduce a listener's comprehension. In other words, in this study, personal bias might have been a factor in some raters' judgments, although its impact might be minor given the general consistency in ratings across raters and groups. In addition, my background as an international student and a former ITA might have influenced the way my participants responded to my questions. The undergraduate students might have altered their responses to my questions as compared to what they would normally say in an informal situation. The interview data should be interpreted with this in mind.

Suggestions for Future Research

One recommendation for future research relates to the investigation of different undergraduate raters' first languages. All the undergraduates in this study were native speakers of English. It may be worth investigating whether undergraduates with different first language backgrounds would perform in the same way. The listeners in Munro, Derwing, and Morton's (2006) study, for example, were native Cantonese, Japanese, Mandarin, and English speakers and the researchers found surprising similarities across these listeners. Future research in ITA testing should investigate the effect of raters' native language backgrounds on the evaluations of potential ITAs' speech. These investigations should involve measures of rater background variables and consider longitudinal or cross-sectional studies to examine the changes, or lack thereof, in the perceptions and evaluations of ITA speech.

One empirical question regarding the impact of the amount of exposure to accented speech on comprehension deserves further investigation. It was assumed that the ESL teachers were more lenient in their accentedness ratings because they have had extensive exposure to accented L2 speech as compared to the undergraduate raters, which might have enhanced their comprehension of the examinees' speech or caused them to overlook speech features that were difficult to process for the linguistically naïve undergraduate students. However, to what extent does the amount of exposure to certain accents impact listening comprehension of these accents? Moreover, different ITA screening tests or tools, different methods of evaluations, and most importantly different aspects of L2 speech (e.g., intelligibility) should also be examined. Finally, we need to explore other rater background characteristics that play a role in the perceptions and evaluations of accented L2 speech. Answers to these questions all have implications for rater recruitment and training in ITA testing and ITA training in general and are also directions for future research to take.

CHAPTER 7: CONCLUSION

In this dissertation study, I found striking similarities across rater groups in ratings of oral proficiency but significant differences in ratings of accentedness and comprehensibility. The analyses of the written comments and the interviews revealed that the ESL teachers and the undergraduates evaluated the examinee speech through a constellation of performance features, all of which emerged to factor into raters' decision-making processes. The sheer quantity of the linguistic and nonlinguistic factors raters commented on testifies to the complexity and dynamics of human judgements in performance assessment. It also affirms the difficulty of obtaining uniform ratings across raters of different backgrounds, even among experienced ESL teachers.

Results of this study provide evidence that rater backgrounds, in this case, the experienced ESL teachers versus linguistically naïve undergraduate students, has a minimal effect on rater severity in oral proficiency ratings, and yet played an important role in raters' perceptions of accentedness and comprehensibility. Results also show substantial variations in rater orientations from individual to individual. While there is a notable number of comments sharing the same concerns about the speakers' performances, the undergraduates tended to evaluate the examinees' oral proficiency more globally while the ESL teachers appeared to rate more analytically by attending to different linguistic features of the speech samples.

This study employs a mix-method design to examine rater effects and rater orientations in an ITA testing context. Through the investigation of three separate aspects of L2 speech: oral proficiency, accentedness, and comprehensibility simultaneously, systematic between-group comparisons, and qualitative analysis of interview protocols, this study has implications for research in ITA testing, research, and pedagogy. This manifold research design is a step further from the sole comparison of ITAs' oral proficiency between ESL teachers and American undergraduates as has done in much previous research. The study has responded to the call for the inclusion of American undergraduate students in local test validation studies (Isaacs, 2008; Morley, 1994). Furthermore, the exploration of the factors that figure into raters' decision-making processes provides insights into ways that ESL teachers and American undergraduates perceive ITA speech and how the differences in perceptions might be addressed. Nonetheless, much more work is called for to further examine the wide range of possible factors that contribute to the perceptions of different aspects of L2 speech through the involvement of raters of various backgrounds and by the expansion of research scope to examine more diverse ITA speech samples from universities across various geographical regions.

APPENDICES

APPENDICES

Appendix A: State Statutes and Regulations on ITA English Proficiency

Table 26. State Statutes and Regulations on ITA English Proficiency

State	Statute or Regulation	English Proficiency Assessment
Arizona	Regulation, 1985 Board of Regents 12-407	Assess English proficiency of each TA each semester
California	Statute, 1987 Assembly Concurrent Resolution 41, Ch.103	Test of Spoken English or a similar test, demonstration with class, or a faculty evaluation
Florida	Statute, 1983 Stat. Ann Sec. 240-246	Test of Spoken English or a similar test approved by the Board of Regents
Illinois	Statute, 1986 Public Act 84-1434, Ch.122, Sec. 3-29.2	University must have a program to assess English proficiency
Kansas	Regulation, 1985, 1988 Board of Regents	Interviewed and certified by three instructional personnel: achieve a 220 on the Test of Spoken English or the Speaking Proficiency English Assessment Kit test
Kentucky	Statute, 1992 Acts Ch.407 & 1	Universities shall institute English language proficiency assessment to demonstrate ability to deliver all lectures and oral presentations
Louisiana	Statute, 1991 SB 327 Sec. 1	Universities to evaluate faculty for fluency in English
Minnesota	Statute, 1986 Ch. 401 Sec. 5	University and state board must ensure proficiency in speaking, reading, and writing
Missouri	Statute, 1986 Rev. Stat. Sec 170.012	Tested for ability to communicate orally in a classroom setting
Oklahoma	Statute, 1982 Stat. Titl. 70 OS Supp. Sec 3324-S	Each college and university to provide an annual report setting forth the procedures established to guarantee English proficiency
Oregon	Regulation, 1986 Board of Higher Education Strategic Plan	TAs should be required to provide evidence of satisfactory English-speaking and writing ability
Pennsylvania	Statute, 1990 SB 539	Appropriate criteria such as personal interview, peer, alumni, student observation and tests to determine English proficiency
Rhode Island	Regulation, 1993 Board of Governors for Higher Education	Establish appropriate policies and programs to assess and when necessary improve the oral English proficiency of all newly hired teaching personnel

Appendix A. (cont'd)

State	Statute or Regulation	English Proficiency Assessment
Tennessee	Statute, 1984 Sen. Joint Resol. No. 211	Satisfactory grade on the Test of Spoken English or a similar test approved by respective board
Texas	Statute, 1989 House Bill 638	All public universities provide a program or short course to ensure that courses be taught clearly in English language.

Note. Adapted from "How international teaching assistant programs can prevent lawsuits" by N. Oppenheim, 1997, pp. 5-6.

Appendix B: Universities That Use the SPEAK Test for ITA Screening

Table 27. Universities That Use the SPEAK Test for ITA Screening

- Arizona State University
- Auburn University
- Boston University
- Drexel University
- Duquesne University
- Florida State University
- Indiana University
- Iowa State University
- Kansas State University
- Lehigh University
- Michigan State University
- North Carolina State University
- Northwestern University
- Ohio State University
- Ohio University
- Oklahoma State University
- Penn State University
- Rutgers University
- Temple University
- Texas Tech University
- University of Alabama
- University of Arkansas
- University of California, Berkeley
- University of California, Davis
- University of California, Irvine
- University of Central Florida
- University of Connecticut
- University of Delaware
- University of Houston
- University of Illinois at Chicago
- University of Illinois at Urbana-Champaign
- University of Iowa
- University of Massachusetts Amherst
- University of Miami
- University of Missouri
- University of Nebraska-Lincoln
- University of North Texas
- University of Pennsylvania
- University of Rhode Island
- University of South Florida
- University of Texas at Arlington
- University of Texas at San Antonio
- University of Utah
- University of Virginia
- University of Washington
- University of Wisconsin-Madison
- Wayne State University
- West Virginia University
- Yale University

Appendix C. SPEAK Test Rating Scale

Table 28. SPEAK Test Rating Scale

	60	50	40	30	20
Overall performance	Communication almost always effective: task performed very competently	Communication generally effective: task performed competently	Communication somewhat effective: task performed somewhat competently	Communication generally not effective: task generally performed poorly	No effective communication: no evidence of ability to perform task
Overall features to consider	Speaker volunteers information freely, with little or no effort; and may go beyond the task by using additional appropriate functions. Native-like repair strategies Sophisticated expressions Very strong content Almost no listener effort required	Speaker volunteers information sometimes with effort; usually does not run out of time. Linguistic weaknesses may necessitate some repair strategies that may be slightly distracting Expressions sometimes awkward Generally strong content Little listener effort required	Speaker responds with effort; sometimes provides limited speech sample and sometimes runs out of time. Sometimes excessive, distracting, and ineffective repair strategies used to compensate for linguistic weaknesses (e.g., vocabulary and/or grammar) Adequate content Some listener effort required	Speaker responds with much effort; provides limited speech sample and often runs out of time. Repair strategies excessive, very distracting, and ineffective Much listener effort required Difficult to tell if task is fully performed because of linguistic weaknesses, but function can be identified	Extreme speaker effort is evident; speaker may repeat prompt, give up on task, or be silent. Attempts to perform task end in failure Only isolated words or phrases intelligible, even with much listener effort Function cannot be identified

Appendix D: ESL Teacher Rater Background Questionnaire

1. Name: ____
2. Age: ______________

3. Gender: Male _______ Female _________

4. How long have you taught English? ___________ year(s) _________ month(s)

5. Where did/ do you teach English? (check all that applies)

i.) _____ in the United States For how long? ___________

ii.) _____ outside the United States For how long? ___________

6. What are the levels of your students? (check all that applies)

beginner _________ intermediate _________

advanced _________ superior ___________

7. Are you familiar with any of the following speaker's accent? (check all that applies)

Arabic _________ Chinese __________ Korean __________

Japanese _________ Spanish __________ Thai __________

Vietnamese ________ Others ______________________________________

8. How long have you been a SPEAK rater? _________ year(s) _________ month(s)

Appendix E: Undergraduate Rater Background Questionnaire

1. Name: ____

2. Age: _____________

3. Gender: Male __________   Female ____________

4. Major at MSU :_____________________

5. Year at MSU:

freshman _______   sophomore _______   junior _______   senior ________   other ______

6. How many courses taught by international teaching assistants (ITAs) at MSU have you taken?

1 _______   2 _______   3 _______   4 _______   More than 4 ________

7. How often did you talk to your ITAs?

less than once a week __________   once a week __________

twice a week ____________   three times a week ____________

more than four times a week __________

8. Do you have friends who speak the following foreign languages (Check all the applies)?

Arabic _________ Chinese _________ Korean _________ Hindi _________

Others (Please specify) _______________________

9. Were you exposed to different foreign languages or culture in your upbringing?

Yes __________   No _________

If yes, what foreign language(s) are you most familiar with? (Please specify)

____

10. Do you have any foreign friends?

Yes __________   No ____________

If yes, what are your friends' native languages?

____

Appendix F: Follow-up Interview Questions

A. ESL teacher raters interview questions

1. Please talk about your experience rating the SPEAK test?
2. How well do you think the SPEAK test results demonstrate the examinees' speaking ability?
3. What is your opinion about the use of the SPEAK test as a screening tool for international teaching assistants at MSU?
4. How easy or difficult do you find using the SPEAK test rating scale?
5. What other criteria if any, except those specified on the rating scale, do you consider when you assign scores?
6. Besides the examinee's speaking ability, what other factors do you think are important when a department is selecting international teaching assistants? Why do you think so?

B. Undergraduate raters interview questions

1. Please describe your experience taking course(s) taught by international teaching assistant(s).
2. What factors, except speaking ability, do you think are important when screening international teaching assistants?
3. When you rated the speech samples, what rating criteria did you consider important? And why?
4. Do you have foreign friends? If yes, how often do you hang out with them? What are their native languages?
5. What is your definition of a good TA?
6. Can you talk about your upbringing? For example, are you exposed to different cultures?

REFERENCES

REFERENCES

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning, 38*(4), 561-613.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford 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*(2), 238-257.

Bailey, C. M. (1983). Foreign teaching assistants at U.S. universities: Problems in interaction and communication. *TESOL Quarterly, 17*(2), 308-310.

Bailey, C. M. (1984a). A typology of teaching assistants. In C. M. Bailey, F. Pialorsi & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in U.S. universities* (pp. 110-125). Washington D.C.: National Association for Foreign Student Affairs.

Bailey, C. M. (1984b). The "Foreign TA Problem". In C. M. Bailey, F. Pialorsi & J. Zukowski/Faust (Eds.), *Foreign teaching assistants in US universities* (pp. 3-16). Washington, DC: National Association for Foreign Student Affairs.

Bannon, P. (2005, June 24). Brilliant instructors, imperfect English*.* *The New York Times*. Retrieved from http://www.nytimes.com

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31-57.

Barkaoui, K. (2010b). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing, 27*(4), 515-535.

Barkaoui, K. (2010c). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74.

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

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

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. 85-103). London: Sage Publications.

Bejar, I. I. (1985). *A preliminary studey of raters for the Test of Spoken English (TOEFL Research Report RR-85-5)*. Princeton, NJ: Educational Testing Service. Retrived from http://www.ets.org/Media/Research/pdf/RR-85-05.pdf

Binghadeed, N. (2008). Acoustic analysis of pitch range in the production of native and nonnative speakers of English. *The Asian EFL Journal Quarterly, 10*, 96-113.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

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

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

Brown, A. (2007). An investigation of the rating process in the IELTS oral interview. In M. Milanovic & C. J. Weir (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 98-141). Cambridge: Cambridge University Press.

Brown, A., & Hill, K. (2007). Interviewer style and candidate performance in the IELTS roal interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 37-61). Cambridge: Cambridge University Press.

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purpose speaking tasks* (TOEFL Research Report RR-05-05). Princeton, NJ: Educational Testing Service. Retrived from http://www.ets.org/Media/Research/pdf/RR-05-05.pdf

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25*(587-603).

Brown, K. (1992). American college student attitudes toward non-native instructors. *Multilingua, 11*(3), 249-265.

Brown, K., Fishman, P., & Jones, N. (1990). *Legal and policy issues in the language proficiency assessment of international teaching assistants*. Houston: University of Houston Law Center.

Bryd, P., & Constantinides, J. C. (1992). The language of teaching mathematics: Implications for training ITAs. *TESOL Quarterly, 26*(1), 163-167.

Cargile, A. C., & Giles, H. (1997). Understanding language attitudes: Exploring listener affect and identify. *Language and Communication, 17*(3), 195-217.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-35.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Chiang, S.-Y. (2009). Dealing with communication problems in the instructional interactions between international teaching assistants and American college students. *Language and Education, 23*(5), 461-478.

Clarke, J., & Swinton, C. (1980). *The Test of Spoken English as a measure of communicative ability in English-medium instructional settings* (TOEFL Research Report RR 80-33). Princeton, NJ: Educational Testing Service. Retrived http://www.ets.org/Media/Research/pdf/RR-80-33.pdf

Congdon, P., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*, 762-765.

Council of Graduate Schools (2007). 2006 CGS international graduate admissions survey. Retrieved from http://www.cgsnet.org/portals/0/pdf/R_Intlenrl06_III.pdf

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Dalle, T. S., & Inglis, M. J. (1989, March). *What really affects undergraduates' evaluations of nonnative teaching assistant's teaching?* Paper presented at the meeting of Teachers of English to Speakers of Other Languages, San Antonio, TX.

Davies, C., Tyler, A., & Koran, J. (1989). Face-to-face with native speakers: An advanced training class for international teaching assisstants. *English for Specific Purposes, 8*, 139-153.

Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in language test design and delivery*. Sydney, Australia: National Center for English Language Teaching and Research, Macquarie University.

Derwing, T. M. (1990). Speech rate is no simple matter: Rate adjustment and NS-NNS communicative success. *Studies in Second Language Acquisition, 12*, 303-313.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 20*, 1-16.

Derwing, T. M., & Munro, M. J. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics, 22*(3), 324-337.

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-398.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476-490.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 383-410.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development, 23*(4), 245-259.

Dick, R. C., & Robinson, B. M. (1994). Oral English proficiency requirements for ITAs in U.S. colleges and universities: An issue in speech communication. *JACA, 2*(1), 77-86.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144.

Douglas, D., & Smith, J. (1997). *Theoretical underpinnings of the Test of Spoken English revision project* (TOEFL Monograph Series RM-97-2). Princeton, NJ: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RM-97-02.pdf

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-185.

Eckes, T. (in press). Many-facet Rasch measurement. In T. S. (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe/Language Policy Division.

Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing, 10*, 235-254.

Engelhard, G. J., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the advnaced placement English literature and composition program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). Princeton, NJ: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-03-01-Engelhard.pdf

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: MIT Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Finder, A. (2005, June 25). Unclear on American campus: What the foreign teacher said*.* *The New York Times*. Retried from http://www.nytimes.com

Flege, J. E. (1988a). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America, 84*(1), 70-79.

Flege, J. E. (1988b). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human communication and its disorders: A review-1988* (pp. 224-401). Norwood, NJ: Ablex.

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91*, 370-389.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 106*, 2973-2987.

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

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

Gorsuch, G. J. (2003, December). The educational cultures of international teaching assistants and U.S. universities. *TESOL-EJ, 7*(3). Retrieved from http://www.tesl-ej.org/wordpress/

Gravois, J. (2005, April 8). Teach impediment. *The Chronicle of Higher Education.* Retried from http://chronicle.com/section/Home/5

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

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

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hinofotis, F. B., & Bailey, C. M. (1981). American undergraduates' reactions to the communication skills of foreign teaching assistants. In J. C. Fisher, M. A. Clarke & J. Schachter (Eds.), *On TESOL '80: Building bridges: Research and practice in teaching English as a second language* (pp. 120-133). Washington, DC: TESOL.

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

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*(2), 243-269.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods, 5*(1), 64-86.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review, 64*(4), 555-580.

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

Jia, C. L., & Bergerson, A. A. (2008). Understanding the international teaching assistant training program: A case study at a northwestern research university. *International Education, 37*(2), 77-129.

Johnson, J., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485-505.

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

Kang, O. (2008). Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 181-205.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*, 301-315.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research*

Colloquium, Orlando, Florida (pp. 1-14). Cambridge: Cambridge University Press.

Kunnan, A. J. (2005). Towards a model of test evaluation: Using the Test Fairness and the Test Context Frameworks. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment* (pp. 229-251). Cambridge: Cambridge University Press.

Landa, M. (1988). Training international students as teaching assistants. In J. A. Mestenhauser & G. Marty (Eds.), *Culture, learning, and the disciplines: Theory and practice in cross-cultural orientation* (pp. 50-57). Washington, DC: National Association for Foreign Student Affairs.

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

Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions, 12*, 630-631.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2010). FACETS (Version 3.67) [Computer Software]. Chicago: WINSTEPS.com.

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.

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. New York: Routledge.

LoCastro, V., & Tapper, G. (2008). International teaching assistants and teacher identity. *Journal of Applied Linguistics, 3*(2), 185-218.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters. *Language Testing, 19*, 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Lang.

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

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*, 425-444.

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.

McCracken, G. (1988). *The long interview*. London: Sage.

McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-75.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Longman.

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

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupationtional setting. *Language Testing, 14*, 140-156.

Meiron, B. E. (1998, April). *Rating oral proficiency tests: A triangulated study of rater thought processes*. Paper presented at the meeting of Language Testing Research Colloquium, Monterey, CA.

Meiron, B. E., & Schick, L. S. (2000). Ratings, raters and test performance: An exploratory study. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 153-176). Cambridge: Cambridge University Press.

Mendelsohn, D., & Cumming, A. (1987). Professor's ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal, 5*(1), 9-26.

Merrylees, B., & McDowell, C. (2007). A survey of examiner attitudes and behavior in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 142-184). Cambridge: Cambridge University Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: ACE/Macmillan.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 92-114). Cambridge: Cambridge University Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* Thousand Oaks: Sage.

Monoson, P. K., & Thomas, C. F. (1993). Oral English proficiency policies for faculty in U.S. higher education. *Review of Higher Education, 16*, 127-140.

Morley, J. (1994). A multidimensional curriculum design for speech-pronunciation instruction. In J. Morley (Ed.), *Pronunciation pedagogy and theory* (pp. 64-91). Alexandra, VA:

Teachers of English to Speakers of Other Languages.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-218). Philadelphia, PA: John Benjamins.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing, 11*(3), 253-266.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 41*(1), 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

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

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

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

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111-131.

Munro, M. J., Derwing, T. M., & Sato, K. (2006). Salient accents, covert attitudes: Consciousness-raising for pre-service second language learners. *Prospect, 21*(1), 67-79.

Muthuswamy, N., Smith, R., & Strom, R. B. (2004, May). *"Understanding the problem": International teaching assistants and communication*. Paper presented at the meeting of International Communication Association, New Orleans, LA.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report RR-00-06). Princeton, NJ: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-00-06-Myford.pdf

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-faceted Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-faceted Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal, 62*, 266-275.

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17*, 23-44.

O'Loughlin, K. (2007). An investigation into the role of gender in the IELTS oral interview. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 63-95). Cambridge: Cambridge University Press.

Okoth, E., & Mupinga, D. M. (2007, February). *An evaluation of the international graduate teaching assistants training program*. Paper presented at the meeting of The Academy of Human Resource Development International Research Conference in the Americas, Indianapolis, IN. Retried from http://www.eric.ed.gov/PDFS/ED504334.pdf

Oppenheim, N. (1996, April). *Undergraduates learning from nonnative English-speaking teaching assistants*. Paper presented at the meeting of the American Educational Research Association, New York. Retried from http://www.eric.ed.gov/PDFS/ED504334.pdf

Oppenheim, N. (1997, March). *How international teaching assistant programs can prevent lawsuits*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL. Retried from http://www.eric.ed.gov/PDFS/ED408886.pdf

Oppenheim, N. (1998, March). *Undergraduates' assessment of international teaching assistants' communicative competence*. Paper presented at the meeting of the Teachers of English to Speakers of Other Languages. Retried from http://www.eric.ed.gov/PDFS/ED423783.pdf

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143-154.

Orth, J. L. (1983). *University undergraduate evaluational reactions to the speech of foreign teaching assistants.* University of Texas at Austin, Austin, Texas.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. Newbury Park: Sage.

Pica, T., Barnes, G. A., & Finger, A. G. (1990a). *Discourse and performance of international teaching assistants*. New York, NY: Newbury House.

Pica, T., Barnes, G. A., & Finger, A. G. (1990b). *Teaching matters: Skills and strategies for international teaching assistants*. New York: Newbury House.

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

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes, 23*(1), 19-43.

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.

Plakans, B. S. (1997). Undergraduates' experiences with and attitudes toward international teaching assistants. *TESOL Quarterly, 31*(1), 95-118.

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing, 27*(2), 235-260.

Powers, D. E., Shedl, M. A., Wilson-Leung, S., & Butler, F. A. (1999). *Validating the revised TSE® against a criterion of communicative success* (TOEFL Research Report 99-05). Princeton: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-99-05.pdf

Rao, N. (1995, May). *The Oh No! Syndrom: A Language Expectation Model of undergraduates' negative reactions toward foreign teaching assistants*. Paper presented at the meeting of the International Communication Association, Albuquerque, NM. Retried from http://www.eric.ed.gov/PDFS/ED384921.pdf

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder (Ed.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly, 21*(4), 643-672.

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511-531.

Rubin, D., & Smith, K. A. (1990). Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of nonnative English-speaking teaching assistants. *International Journal of Intercultural Relations, 14*, 337-353.

Ruderman, A. (2000, December 27). Colleges are moving to ensure English fluency in teaching assistants. *The New York Times*. Retried from http://www.nytimes.com

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

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129-152). Cambridge: Cambridge University Press.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69-90.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493.

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.

Sebastian, R., & Ryan, E. B. (1985). Speech cues and social evaluation: Markers of ethnicity, social class, and age. In H. Giles & R. N. St. Clair (Eds.), *Recent advances in language, communication and social psychology* (pp. 112-143). London: Lawrence Erlbaum.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Teicher, S. A. (2005, April 18). When you can't understand the teacher. *The Christian Science Monitor*. Retrieved from http://www.csmonitor.com/2005/0418/p11s02-ussc.html

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(1), 1-30.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*(4), 713-729.

Tyler, A., Jefferies, A., & Davies, C. E. (1988). The effect of discourse structuring devices on listener perceptions of coherence in non-native university teachers' spoken discourse. *World Englishes, 7*, 101-110.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke:

Palgrave Macmillan.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly, 26*(4), 693-711.

Winke, P., Gass, S. M., & Myford, C. M. (in press). *The relationship between raters' prior language study and the evaluation of foreign language speech samples*. Princeton: Educational Testing Service.

Xi, X. (2008). *Investigating the criterion-related validity of the TOEFL Speaking Scores for ITA screening and setting standards for ITAs* (TOEFL iBT Research Report 08-02). Princeton: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-08-02.pdf

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps?* (TOEFL iBT Research Report 09-31). Princeton, NJ: Educational Testing Service. Retried from http://www.ets.org/Media/Research/pdf/RR-09-31.pdf

Yook, E. L., & Albert, R. D. (1999). Perceptions of international teaching assistants: The interrelatedness of intercultural training, cognition, and emotion. *Communication Education, 48*, 1-17.

Yule, G., & Hofffman, P. (1990). Predicting success for international teaching assistants in a U.S. university. *TESOL Quarterly, 24*(2), 227-243.

Zhao, Y. (1997). The effects of listener' control of speech rate on second language comprehension. *Applied Linguistics, 18*, 49-68.

Zielinski, B. (2006). The intelligibility cocktail: An interaction between speaker and listener ingredients. *Prospect, 21*(1), 22-45.

Zielinski, B. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*, 69-84.