THE IMPACT OF WORLD ENGLISHES ON LANGUAGE ASSESSMENT: RATER ATTITUDE, RATING BEHAVIOR, AND CHALLENGES

BY

HUEI-LIEN HSU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Fred Davidson, Chair
Professor Jennifer Greene
Professor Rakesh Bhatt
Professor Dov Cohen

ABSTRACT

By centralizing the issue of test fairness in language proficiency assessments, this study responds to a call by researchers for developing greater social responsibility in the language testing agenda. As inquiries into language attitude and psychology indicate, there is an underlying uncertainty pertaining to the validity of test use and score interpretation based on listeners' bias against for non-standard English and negative evaluation of such speakers. Of greater relevance in oral proficiency assessment is that listeners, that is, raters, transfer such attitudes to their scoring judgments. As an attempt to address this issue, this study investigates the scoring validity of the IELTS speaking test by examining its relationship in relation to a criterion designed to measure rater attitudes towards World Englishes.

Validity arguments were formulated to guide two independent, yet related, studies based on mixed-methods approach and evaluate the claims and hypotheses set for the studies. In view of the lack of instruments measuring rater attitude towards global English in the language assessment context, the first study constructed the criterion measure, the Rater Attitude Instrument (RAI), involving 119 ESL teacher raters in the U.S. and India. As a result of the three-phase development, the RAI comprises 22 semantic differential scale items and 32 Likert scale items representing the three attitude dimensions of feeling, cognition, and behavior tendency used by psychologists. Confirmatory factor analysis supports the internal structure of the RAI with acceptable model fit indices ($\chi^2 =20.052, p =.094$, $RMSEA=0.076$, $CFI =0.954$, $TLI=0.926$). Content validity is ensured through teacher raters and content experts perspectives that continuously shaped the substance of the RAI. As the RAI demonstrates, rater attitudes towards World Englishes were generally positive and tended to focus more on speech comprehensibility; nevertheless, the majority of raters were inclined towards a preference for standard English in their scoring judgments.

In the second study, the RAI and the six IELTS descriptive tasks produced by Indian examinees were administered on-line to the 96 teacher raters and the data analyzed to evaluate the extent to which the claim that rater attitude is a biasing factor affecting their scoring judgment on IELTS descriptive tasks can be sustained. The RAI scores were analyzed by FACETS that classified the raters into positive, neutral, and negative attitude groups according to measurement logit. Next, MANOVA was performed which suggested that the ratings by the positive and negative attitude groups were significantly different, with the positive group consistently rating higher on all the four criteria of Fluency, Sentence Structure, Vocabulary, and Pronunciation. The neutral and negative attitude groups rated significantly differently on Sentence Structure and Vocabulary, with the former rating higher than the negative group. Moderate to strong correlations ranging from .272 to .569 were observed between the RAI and the IELTS descriptive task scores. Multiple regression analysis revealed that RAI scores accounted for a significant amount of variance on the IELTS descriptive tasks sub- and total-scores, ranging from 17.5% to 32.4%. Moreover, the Indian/non-Indian variable was the only rater background characteristic investigated that significantly related to the rater feeling scores that formed one of the triplet attitude constructs, though contributing to only 10% of the score variance. Lastly, the verbal protocol study provided insightful information suggesting that raters with positive attitudes generally took into account the expected performance of language learners while some negative-attitude raters used the native speaker model as the underlying criterion for judgment. The impact of the findings on validity argument, test fairness, and rater trainings are also discussed.

To my family

ACKNOWLEDGEMENTS

This six-year passage towards my PhD has been an exciting and emotional journey that trailed through so many eventful moments and critical milestones, and benefitted from the support of many unforgettable personalities. I am indebted to my advisor and dissertation director Professor Fred Davidson for his endless support, encouragement, and faith in me. During my six years as his research assistant at the Foreign Language Assessment Group (FLAG) I was privileged to work on a variety of language assessment projects that both sustained and strengthened my interest as a researcher in this challenging and interesting field. His invaluable guidance and assistance at every stage made it possible for me to complete this dissertation as planned. For this, I am forever indebted.

Professor Jennifer Greene's stimulating and thought-provoking instruction in mixed methods design was an intellectual eye-opener and will continue to inspire my future endeavors in language assessment research and applications. I am grateful to Professor Rakesh Bhatt who made my introduction to World Englishes and provided thoughtful insights that were a source of comfort in my exploration of the interface between World Englishes and language assessment. Professor Dov Cohen's valuable guidelines on the handling of data regarding a subject's psychological traits have enhanced my sensitivity and competence in quantitative data analysis.

My special thanks go to all ESL teachers and raters for their inputs on oral proficiency tests in the U.S. and India and for participating in every stage of the dissertation. I gratefully acknowledge the financial support from the TOEFL small grants for doctoral student in foreign language assessment for facilitating the completion of the dissertation, and the IELTS validation group for their full support in providing resources integral to the study. The fellowship provided to students studying aboard by the Ministry of Education in Taiwan allowed me to concentrate on the data analysis and write-up during the last two years of the doctoral program.

Friendships developed over the years in Champaign will remain a lifelong memory to cherish. Friends, both far and near, provided priceless support in so many different ways in helping towards the completion of this dissertation. FLAG team members challenged my intellectual growth, shaped critical aspects of this dissertation, and were boundless sources of mental support. I wholeheartedly thank my friends from FLAG, Kadeessa, Jiyoung, Youngshin, Sun Joo, Soyoung, Jishu, Carsten, and Heejeong. The quantitative data analysis for this dissertation benefitted greatly from the professional help of Youngshil, Andy, and Saijun. I am grateful to friends - Ming-jing, Seven, and Yan - who were always ready to provide needed transport and to baby-sit my daughter.

Finally, the endless love of my family provided the motivation for any difficult task I set out to do. In a debt that can never be repaid, I am forever grateful to my mom who came all the way from Taiwan to keep my family life organized and gave me room to concentrate on the dissertation write-up - and for trying to make this quiet little town her second home. I enjoyed being spoiled and being her baby again. My sister's lucky hands in my faculty application materials gave me the confidence while I was struggling with the dissertation and job search simultaneously. My husband's full support in everything I do and in every dream that I want to pursue has been a source of endless comfort and confidence. His background in the field of electrical and computer engineering provided fresh perspectives and interesting inputs on the assessment projects I was involved in. Hannah, our lovely daughter, has the delightful ability with simply a smile to ease my stress and sustain the momentum when the drive flagged. We were blessed that she came along during the doctoral program and having her beside me contributed immeasurably to the challenges and joys of this memorable journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Background of the Study

This dissertation study responds to a call for an investigation of rater scoring behavior in relation to the multiple varieties of English around the world. Over the past two decades, practices and theoretical debates in second language testing research has evolved to a different level as research efforts in sociolinguistics has made the re-orientation of the English language more explicit. With terminology such as World Englishes (Kachru, 1992; Smith 1992), English as Lingua Franca (Jenkins, 2006) and English as an international language (Seidlhofer, 2004), the research has documented the function, status, linguistic maturity and legitimacy of the multiple varieties of English indicating that English language can and should no longer be viewed as a homogenous entity. As an international language test provider, the Cambridge ESOL, for example, has shaped its practices and policies to allow for the fact that effective communication in the wider international context is possible despite the varieties of English (Taylor, 2002). On a theoretical level, discussions focus on which language norms should be adopted and call for a change in the approach to language assessment from both within and outside the field. A dominant view that prevails is concerned with the negative consequences of tests that are judged against a single norm and urge a communicative-oriented test that measures examinee’s ability to negotiate their ways through varieties of English. This is where the concerns are raised with regard to the validity of test use as rater judgment may vary despite rater training (Lumley & McNamara, 1995; Weigle, 1998). Language variations in a test may challenge rater judgment as a function of the broader spectrum of World Englishes and the consequence on test scores is unknown.

Issues of language norms on which oral tests are based have been mainly discussed from three perspectives: reality, ideology, and authenticity. As noted by Taylor (2002), standard American English seems to be overwhelmingly favored in well-known international language tests, such as those given by the Michigan Test Battery and Educational Testing Service. There are criticisms about the lack of validity of the tests as they do not represent the sociolinguistic identity of examinee's language use when are administered in contexts where the norms for the tests are different (Chalhoub-Deville & Wigglesworth, 2005, Davidson, 1994, 2006; Lowenberg, 1993, 2002). Arguing from a critical ideological perspective (Davidson, 1993; Spolsky,1993), the continued use of a single norm, typically either American English or British English, points to the perpetuation of American and British world-views and cultures, leading to a "neocolonialist measurement imperialism"(Davidson,1993, p. 114). As far as authenticity is concerned, scholars urge for a communicative-based assessment practice that reflects the changes spawned by the global spread of English, leading to the need for examinees to demonstrate their ability to utilize their own variety and linguistic resources for achieving successful communication in wider contexts (Canagarajah, 2006; Jenkins, 2002). In that sense, norms, regardless of the context they are based upon, appear less relevant in contemporary assessment practice. Rather, it is suggested that the criteria for assessment build upon the extent to which they effectively fulfill the communication task (Canagarah, 2006).

Based on three perspectives above, it should be noted that scholars are increasingly accepting varieties of English in their assessment practices and integrating such decisions in the production tests (i.e. speaking and writing). Nevertheless, when raters deal with communicative-oriented tests and assess the efficiency of the language use of examinees, rater recognition, perception and acceptance of varieties of English may affect their rating behavior

(Davies, Hamp-Lyons & Kemp, 2003). This concern raised by testing professionals is rationally grounded. In language-attitude research within the scope of sociolinguistic inquiry, the investigation of listeners' perceptions of a variety of accents and languages reveal a generally consistent pattern of findings where non-standard variety speakers are un-favored regardless of the listeners' ethnic backgrounds (Cargile & Bradac, 2001). Psychological research further evidences that listeners' attitudes are associated with their behavioral tendencies (Ajzen & Timko, 1986; Albarracin, Johnson, Fishbein, & Muellerleile, 2001; Fazio, Powell, & Williams, 1989; Fishbein *et al.*, 2000; Hrubes, Ajzen, & Daigle, 2001). While the listeners in the language attitude research are not oral proficiency assessment raters with power to award scores that have varying impact on examinees' lives, such as in school applications and job promotions, language assessment research should investigate the potential for negative test consequences and their impact on the validity of the oral proficiency assessment.

Recent studies in language assessment research take into account the relevance to rater attitude of varieties of English and its influence on score-making decisions (Chalhoub-Deville & Wigglesworth, 2005; Harding, 2008, Hsu, 2007; Kang 2008; Kim, 2005). Given the different measures of rater attitude, the different varieties as stimulus and inconsistent methodology, that is, using quantitative and qualitative approaches independently or mixing the two, the findings of the studies do not hold up to comparison. In addition to rater attitude, empirical research within the World Englishes context is emerging, such as examinees' performance in listening tests that incorporate multiple accents (Harding, 2008) and the ESL placement writing test to compare score differences when scoring criteria allows for the syntactical and semantic features of examinees' own varieties (Kenkel & Tucker, 1989).

Despite the relatively small number of assessment research studies focusing on issues brought by World Englishes, it is promising to see that this emerging new line of inquiry will not only bring language assessment researchers' attention to the complexity of English use in contemporary social contexts but also push language assessment further beyond traditional psychometric inquiry to a broader social practice (McNamara & Roever, 2006). As such, this dissertation situates itself on a value-laden platform to discover the value implications and intended and unintended consequences brought about by implementing and operationalizing World Englishes in second language speaking tests. Centering rater perception as research agenda that place test fairness in the core of post-Messick (1989) validity inquiry (Kunnan, 2002; 2004), this research examines the extent to which raters are prepared for linguistic diversity and not biased towards a particular variety. Even though any test cannot be completely fair as it is a chain of reasoning of the interpretation and use of test scores, test fairness is argued as being an important test quality and recent more systematic approaches to examine the fairness of test score interpretation and use are proposed (Kunnan, 2010; Xi, 2010). Examining fairness from the perspective of rater performance not only differs from the traditional approach of investigating examinee group differences, but also makes this dissertation more socially responsive by urging a re-consideration of the meaning of English proficiency by testing professionals.

Context of the Study

The context of this study is the International English Language Test System (IELTS) which is a large-scale language assessment that measures English proficiency for purposes of study or work where English is the language of communication. The British version of the Test of English as a Foreign Language (TOEFL) that was developed by the Educational Testing

Service in the U.S., the IELTS has expanded its service and some universities in the U.S. accept IELTS test scores as evidence of English proficiency for admission considerations. The University of Illinois at Urbana-Champaign is one such university. According to the IELTS website, the IELTS is offered every month in more than 125 countries, indicating its popularity among the world's English language testing system. Both its academic and general training test formats have sections on listening, speaking, reading and writing.

IELTS is chosen in this study for two main reasons. First, it is among the first few large-scale language assessments to provide an explicit statement about the acceptance of varieties of English in responding to test tasks. As its exam handbook states:

> The international test IELTS is internationally focused in its content. For example, a range of native-speaker accents (North American, Australian, New Zealand, and British) is used in the Listening test, and all standard varieties of English are accepted in candidates' responses in all parts of the test. (IELTS Information for Candidate, *http://www.ielts.org/pdf/Information_for_Candidates_booklet.pdf*)

As specified in the IELTS research notes, linguistic diversity is included in the presentation of the test in the reading and listening test components which include varietal grammar, lexis, spelling, discourse, and pronunciation. The use of varieties of English in the speaking and writing tests is to "enable candidates to function in the widest range of international contexts" (Taylor, 2002, p. 20). Therefore, both the practical and conceptual levels of the research endeavor revealed the effects of World Englishes on IELTS, which are likely to have considerable impacts on stakeholders worldwide. Students, for example, could place greater value on their own variety and not aim for native-speaker level of English proficiency in achieving effective communication. This will be further discussed in the literature review in chapter 2.

The second reason for choosing IELTS is the resource support received for the current study from the IELTS validation group where access to the actual IELTS speech samples and raters was available to the researcher as an award recipient of the IELTS Jointly Funded Research, Round 16. A detailed description of the access to the IELTS raters will be provided in the chapter on methodology.

Statement of Purpose

With test fairness forming the core of the research agenda, the aim of this study is to seek constructs of rater attitude to varieties of English and to explore if a relationship between rater attitude and scoring tendency can be established. Toward this end, a set of hypothesis is proposed and evaluated in two independent yet related studies that are guided by modern test validation approaches (Kane, Crooks, & Cohen1999; Messick, 1989; Toulmin, 2003; Weir, 2004).

Given the lack of systematic tools available to measure rater attitude within language assessment research, this study first developed a Rater Attitude Instrument (RAI) that was guided by the mixed methods design and utilized an evidentiary reasoning approach to justify the development and revision at each stage of the instrument to reveal the complexities underlying the psychological traits of raters. By evaluating all evidence and its sources, the validity of the RAI was constructed through an argument-based approach by closely linking the instrument development and validation processes.

For study 2 that investigated the attitude-behavior relationship, the major claim that is proposed is that: rater attitude towards varieties of English is a biasing factor that influences rater scoring performance on the IELTS descriptive tasks. A set of hypotheses serving as warrants (Toulmin, 2003, see chapter 2) were tested using quantitative and qualitative approaches to evaluate the extent to which the claims can be supported:

Hypothesis 1. Rater attitude towards World Englishes is not consistent and can be grouped into different attitude groups.

Hypothesis 2. The rater attitude group has a significant effect on IELTS descriptive tasks scores.

Hypothesis 3. Rater scoring performance on IELTS descriptive tasks can be predicted by attitude tendency

Hypothesis 4. Rater attitude is associated with rater background characteristics.

Hypothesis 5. Rater with like attitudes may score the IELTS descriptive tasks in a similar fashion by weighing particular salient features of Indian English more heavily than others.

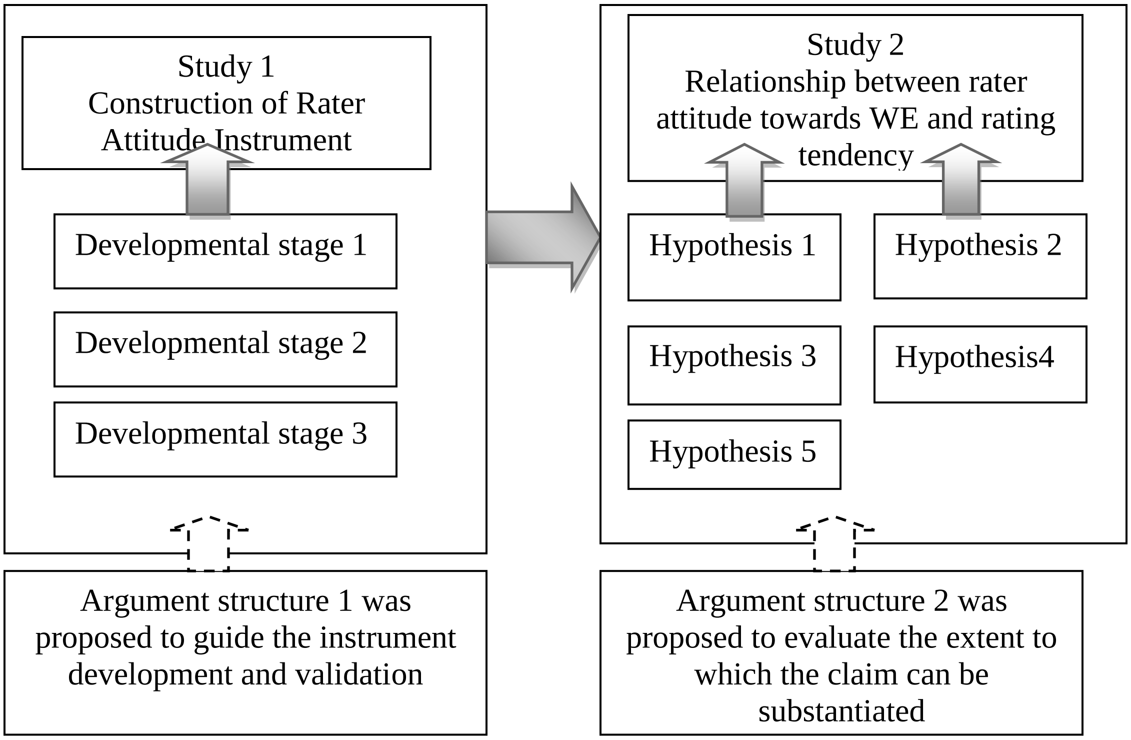

Study 1
Construction of Rater Attitude Instrument
Developmental stage 1
Developmental stage 2
Developmental stage 3
Study 2
Relationship between rater attitude towards WE and rating tendency
Hypothesis 1
Hypothesis 2
Hypothesis 3
Hypothesis4
Hypothesis 5
Argument structure 1 was proposed to guide the instrument development and validation
Argument structure 2 was proposed to evaluate the extent to which the claim can be substantiated

Figure 1. Overview of the relationship of the two studies.

The findings will be compared to the current language assessment research that focuses on rater perception in an attempt to seek comparable results. The hypotheses and claims are linked by means of Toulmin's forms of inference (2003). The details of argument structures are outlined and discussed in chapter3, 4 and 5. The dissertation is divided into two studies, as outlined in Figure 1. The bulk of the hypotheses (above) are evaluated in the second study.

Distinctive features of varieties as discussed and defined in sociolinguistic and World Englishes research include phonology, morphology, sentence structure, cultural norms and communication styles. To what extent the combination of these features in relation to rater perception takes effect within testing is unknown. Therefore, the theoretical review and the empirical study as conducted in this dissertation will be blended to generate an initial definition of constructs of World Englishes within language assessment.

Significance of the Study

This study is timely in view of the growing interest and awareness of the importance of research on World Englishes in relation to language assessment. The results of this study are of particular importance for a variety of reasons. To the researcher's knowledge, this is the first language testing research engaged in instrument development that attempts to investigate different dimension of rater perception about World Englishes. Much of the social-psychological research into language attitudes offer insights into rater attitudes and this study is expected to establish the link between current findings and studies in disciplines relevant to language assessment, such as teacher attitude toward students and L2 learner studies. Additionally, the newly constructed battery of the RAI will hopefully provide language assessment researchers a common tool in the effort to formulate a unified approach in evaluating rater attitudes toward varieties of English.

In terms of the validity argument, modern validity paradigms (Kane 1992, 2001, 2002, 2004; Kane, Crooks, & Cohen, 1999; Messick, 1989) call for a justification for the intended and unintended consequences of tests. The evidence assembled in this study will respond to this call to justify why scores are presumed to fairly and accurately reflect and measure examinee ability in relation to rater attitude, in situations where raters encounter speakers of multiple English varieties. The study will thus enhance language assessment researchers' understanding of test fairness and how test validity may be improved. Furthermore, the findings of the study would inform rater training models towards enhancing rater awareness of the World Englishes framework. Finally, the study bridges the gap among sociolinguistics, social psychology, and language testing, thereby facilitating language testers' appreciation of changes in English use in test construction and implementation.

Operational Definitions of the Terms

World Englishes

The term *World Englishes* is used as an umbrella term to refer to two lines of research. First is the legitimacy, norms and usage of the new varieties of English as established by Kachru (1985), specifically referring to nativized and institutionalized varieties in the outer circle of his concentric model. Secondly, English as Lingua Franca (ELF) has produced extensive research on the nature of English produced by speakers in the expanding circle and revealed that English produced by these speakers appears not to be random, irregular forms of English. As claimed by Seidholfer (2004), ELF has 'taken on a life of its own, independent to a considerable degree of the norms established by its native speakers, and that warrants recognition' (p.212). Although researchers are in disagreement that whether ELF can be categorized into World Englishes paradigm (see Berns 2008; Jenkins 2006), this study will treat

ELF as part of WE in view of its acknowledgement of multiplicism of English and endeavor of ELF linguistic description.

Attitude

A review of literature on the attitude , such as in the context of ESL/EFL teaching and learning, revealed the two terms, perception and attitude, were used differently. They may be used altogether as in “perception and attitude” (Reeves, 2006), interchangeably (Griego-Jones, 1994), treating them as the same construct (Bell, 2009), or placing perception in a higher order that includes attitude (Batang, retrieved from *www.asianmediacongress.org/batang.doc*). According to the definitions of the two terms from the Merriam-Webster online dictionary, “perceive” is defined as “to attain awareness” whereas “attitude” as “a feeling or emotion toward a fact or state”. This seems to imply perception is a “concept” and attitude is a specific state of mind toward an object. In other words, perception affects one’s particular attitude toward a fact. This study uses the two terms interchangeably to capture rater’s awareness of World Englishes and specific feelings about the variety that will be used as stimulus to elicit rater feeling and thoughts.

CHAPTER 2

LITERATURE REVIEW

World Englishes

The global spread of English language in nonnative contexts has gone through a process of nativization as people adopt English for use in different domains and create new and socially appropriate meanings (Kachru, 1985; 1992; 1997; 1998). Kachru's (1998) concentric circle model (Figure 2) captures the unprecedented spread of English, points to the depth of its societal penetration and its varying acquisitional patterns. Despite drawing criticisms for its oversimplification in representing the current pace of English language spread and development (Canagarajah, 2006; Higgins, 2003; McArthur, 2001; Pennycook, 2003; Rajadurai, 2008), Kachru's concentric circle model, representing one of the approaches to World Englishes (WE) (Bolton, 2005), will lead the discussion of this dissertation in view of its far-reaching influence in applied linguistics. The *inner circle* comprises countries where English is used as a mother-tongue, a primary language spoken by the majority and include the USA, the UK, Canada, Australia, and New Zealand. The *outer circle* is primarily made up of countries previously colonized by the US and the UK, such as India, Malaysia, and Liberia, and where the role of English has developed institutionalized functions. English may be bestowed importance by language policy but it could also be one of two or more codes in the linguistic repertoire of the speakers, who may be either bilingual or multilingual. Therefore, English typically exhibits an extended functional range in this circle and is used in various social, educational, administrative and literary domains. Lastly, the *expanding circle* includes countries such as China, Spain, Egypt, and Indonesia where English is learnt as a foreign language and mainly serves as a medium for international communication.

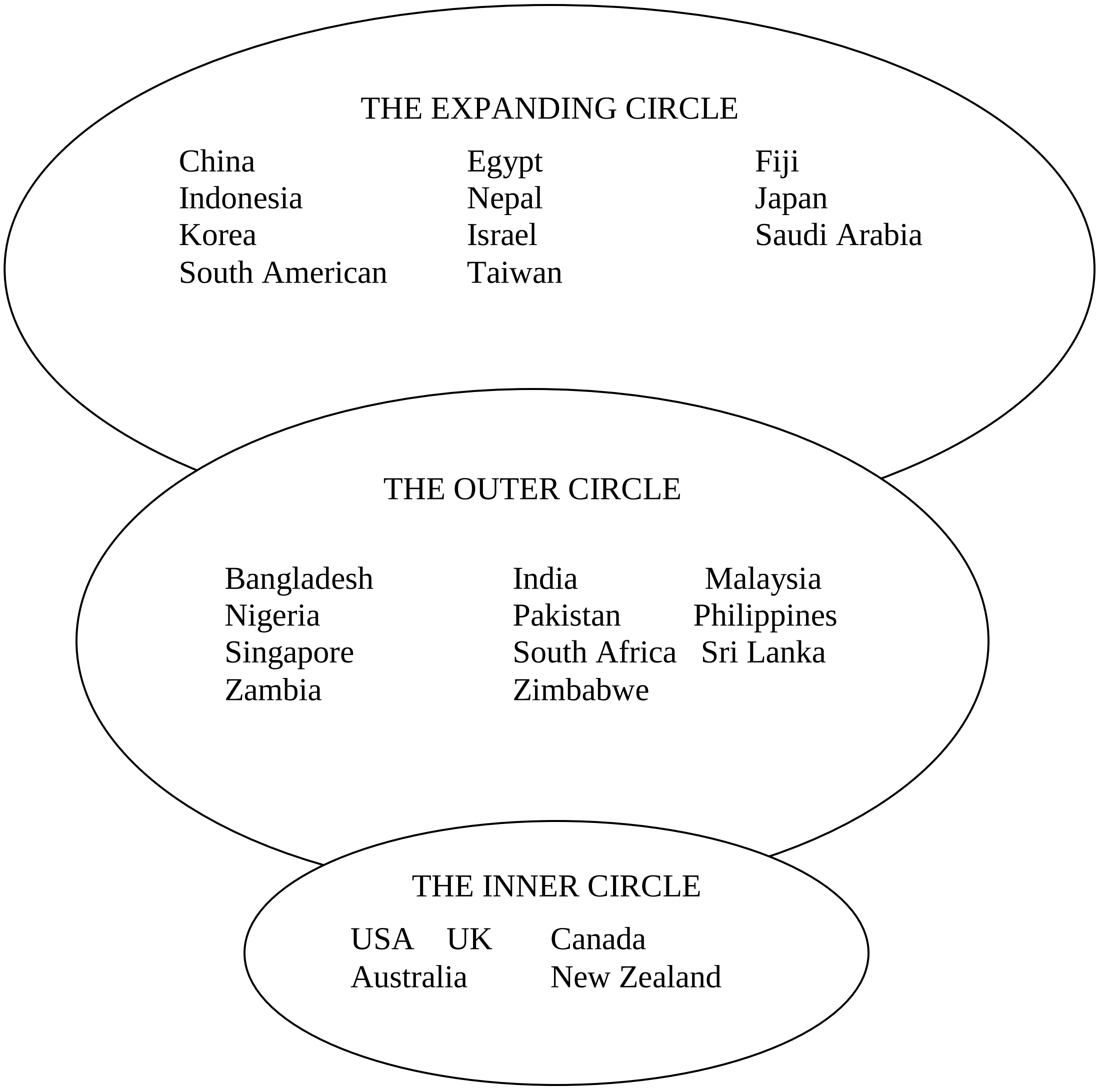


Figure 2. The concentric circle model

While the concentric circle model argues for legitimacy of the new varieties of English mainly in the outer circle and affirming their norms and usage, another line of research, English as Lingua Franca (ELF), has produced extensive research on the nature of English produced by speakers in the expanding circle particularly without involvement of inner circle speakers, revealing systematic and regular forms of English (House, 1999). ELF has ‘taken on a life of its own, independent to a considerable degree of the norms established by its native speakers, and that warrants recognition' (Seidlhofer, 2004, p.212). Although researchers are in disagreement that whether ELF can be categorized into the WE paradigm (see Berns 2008; Jenkins 2006), this study treats ELF as part of WE in view of its acknowledgement of pluricentricity of English. The nature of the ELF research is to find out salient common features of ELF use, irrespective of speakers' L1 and levels of L2 proficiency, and has focused on three levels of language: phonology, pragmatics and lexico-grammar (Seidlhofer, 2004). Of the three areas of research, a predominant emphasis falls on phonology in which "The Phonology of English as an International Language" has been generated by Jenkins (2000) to evaluate which phonological features are essential for mutual intelligibility in contexts where no inner circle speakers of English are present. The work has been termed the phonological "Lingua Franca Core" by Jenkins. For lexcogrammar, the compilation of the corpus was carried out by Seidholfer's (2002) "the Vienna-Oxford International Corpus of English (VOICE)", which captures face-to-face interactions among fluent speakers from a wide range of L1 backgrounds and covers speech on a variety of settings.

Though Kachru's WE paradigm (1985, 1988, 1992) has been recognized for its emphasis on pluralism and linguistic diversity and the power of the sociolinguistic context, the model draws criticism for its oversimplification in representing the current pace of language spread and development, the power and politics associated with it and the complexity of English use in the global context as English has expanded in its use across all the three circles that Kachru has clearly distinguished (Canagarajah, 2006; Higgins, 2003; McArthur, 2001; Pennycook, 2003; Rajadurai, 2008). In terms of English's relative status in the three circles as labeled by Kachru- inner circle's *norm-providing,* outer circle's *norm-developing,* and expanding circle's *norm-dependent*, criticism has arisen that, instead of being used by monolingual speakers in homogenous contexts, English is used more in multinational contexts by multilingual speakers (Graddol, 1999). Furthermore, the English in expanding circle does not necessarily rely on the inner circle variety as users have to adapt the English to facilitate communicative needs (Seidlhofer, 2004).

In terms of linguistic variation, each English variety absorbs and adapts some parent form into a stable and distinctive variety at all levels - phonology, lexicon, syntax, discourse, pragmatics and literary creativity to reflect local needs and facilitate communication (Mesthire & Bhatt, 2008; Y. Kachru, 2005). Clearly, linguistic variation is not unique to outer and expanding circle only; it is also commonly found in all inner circle varieties of English (Kirkpatrick, 2007). Features of different varieties have been thoroughly described and analyzed in *The Handbook of World Englishes* (Kachru, Y. Kachru, & Nelson, 2006), *Handbook of Varieties of English* (Kortmann & Schneider, 2005), and *The Oxford Guide to World Englishes* (McArthur, 2003). Despite differences at all linguistic levels and 'linguistic creativity' as termed by Bhatt (2001), common features were found in phonology (Gramley & Patzold, 2004; Mesthrie & Bhatt, 2008) and syntax (Meierkord, 2004; Mesthrie & Bhatt, 2008). In phonology, similarity includes a tendency to use full vowels rather than schwa in the first syllabus of a word such as *continue*. Other two common phonological processes include final devoicing of obstruent and consonant-cluster reduction. In syntax, syntactic simplification, transferring of local language to the new varieties and the movement of important information to the front of the utterance are common across varieties (Meierkord, 2004). One such syntactic simplification is to avoid the inflectional endings, such as *–ed* to signal past tense, third person singular and to use regular, general and unmarked forms only. These grammatical similarities are possibly caused due to a "pan-linguistic grammatical simplification process" (Crane, 1994, 358), which results in difference between varieties are “merely differences of degree, rather than differences of type” (Kirkpatrick, 2007, p.172).

One core issue discussed by the WE paradigm is uncertainty about the oft-used term *native speaker* as English speakers in the non-inner circles vary in terms of English proficiency, despite its use mainly for pedagogical, pragmatic and research reasons. Whether the concept of native speaker can apply to all situations is called into serious question as English keeps diffusing due to its spread in linguistically and culturally pluralistic societies, the functions it serves in multilingual societies and its differing roles in language planning in each English-using country. Accordingly, the traditional understandings of the ownership of English with the dichotomy of native and non-native users of English were questioned (Higgins, 2003) and appeared less relevant (Kachru, 1997). It should be noted that the field of language testing has engaged with this issue for many years (Davidson, 1993; Davies, 2003) with support from empirical studies showing native speakers’ performance in writing tests varied considerably and examinees of non-native of English could achieve almost the same level of proficiency as native speakers on the reading proficiency test when they successfully applied cognitive skills (Hamilton, Lopes, McNamara & Sheridan, 1993).

On a more theoretical ground, disputes and differences of opinions about the native speaker arise because the idea and concept is interpreted differently and has even been dismissed as a myth (Davies, 2003; Rajagopalan, 1999). Linguistically speaking, one approach sees the native speaker as the guardian of the true language, whilst sociolinguistically s/he is defined according to attitude and identity. What the contrasting views reflect is that different positions can be taken on the basis of interest in and concern for the same phenomenon.

Several researchers have commented that the use of childhood exposure to define native speaker is conceptualized within a monolingual society when the world is in fact mostly multilingual (Mesthrie & Bhatt, 2008). A child in some multilingual societies may acquire several native languages with a varying order of acquisition which is not indicative of "nativeness" in any of the languages. Speakers in today's heterogeneous societies have to interact with a wide range of people representing various power and authority positions on a regular basis via code switching "appropriate to different functions or crisscrossing of functions" (ibid, p. 37). In view of this, the dichotomy between a native and non-native speaker connotes exclusion rather than inclusion of all individuals who are users of language (Lee, 2005) and calls for a redefinition from "who you are" to "what you know" (Rampton, 1990) in terms of functionally communicative accomplishment. Alternative terms are thus used, including "language expert" (Rampton, 1990), "traditional foreigner", "the revisionist foreigner", "the other native" (Davies, 2003), "multicompetent language user" (Cook, 1999), and "competent language user" (Lee, 2005) to highlight English users as speakers in their own right in the plurilingual contexts.

Claims argued by the WE paradigm symbolize work toward greater social equality, enhancement of English users' identity awareness, and an emphasis of the functional role of English use in multilingual contexts. Nevertheless, the current practice, or acceptability of the conceptualization of WE, appears to have a long way to go before the theoretical claims could be fully embraced by scholars, as can be seen in the practice of English language teaching (ELT), which is considered the most influential English language 'gatekeeper'(Kachru 1997). Despite the WE paradigm along with the critical efforts of advocates resisting linguistic imperialism like Phillipson (1992) and Pennycook (1994) there is still strong favor for the inner circle variety and context in ELT (Jenkins, 2007), which “continues to exert a strong ideological force, particularly in influencing notions of how English should be taught and who should do the teaching” (Brutt-Griffler, 2002, p.184). The best that can be said, according to Jenkins (2006), is that WE has raised many teachers’ and teacher educators’ awareness of English language spread. Nevertheless, the insistence on inner-circle variety and its context implicates an inevitable adverse consequence as illustrated in an ELT case in Japan by Matsuda (2003). Matsuda argues that the use of only American English in classrooms suggests a disregard for the real linguistic needs of learners (i.e., learners of Japanese use English more often with outer and expanding circle users), and that it neglects to address the colonial past of the language and the power inequality associated with its history, leading learners to internalize a colonialistic view of the word and devalue their own status in international communication, and lastly it reveals an incapability to promote a right of their ownership of English.

On a practical level, the extensive spread of English and the ensuing variations in different contexts raise immediate concerns that speakers of the three circles may become mutually unintelligible, particularly in international situations. The intelligibility studies, particularly conducted by TESOL and ELF, have been predominantly investigated features of speakers’ phonology (Derwing & Munro, 2005; Munro & Derwing, 1999). For example, Jenkins’ “Lingua Franca Core” (2002; 2006) was created for the purpose of maximizing intelligibility in speaker interaction, which signifies a predominance role of pronunciation in cross-cultural communication. Interests in other linguistic features have emerged recently to broaden the scope of the intelligibility inquiry: lexis (Filed, 2005), syntax (Friederici, Kotz, Scott & Obleser, 2010), combinations of several linguistic uses (Zielinski, 2006) and communication style (Matsuura, Chiba & Fujieda, 1999; Smith & Christopher, 2001; Y. Kachru, 2008), which indicates intelligibility breakdown can be a mix of many aspects of linguistic quality.

Two works on WE have further established guidelines and clearer direction for intelligibility study: the Smith paradigm, as labeled by Kachru (2008) and Nelson (2008) and the 3x3 “World Englishes Speaker-Listener Intelligibility Matrix” (Levis, 2005, p.373). The concept of Smith paradigm has appeared in several Smith’s, Nelson’s and their collaborative work (Nelson, 1982, 1995; Smith & Nelson, 1985; Smith 1992). The most influential features of this paradigm is its explicit definition and distinction of terminology to indicate a degree of understanding on a continuum: the lowest level of intelligibility refers to word and utterance recognition, comprehensibility that is defined as their meaning to the highest level, interpretability, which is the perception or interpretation of the speaker’s intentions. Several studies have shown that communication break-down does not result from an intelligibility issue but the failure of proper interpretation of the message (Smith & Christopher, 2001; Y. Kachru, 2008). Thus, the comprehension of whole utterance may not matter as long as the basic understanding of what is going on and communicative goal is achieved, which reflects what Smith (1988) forcefully argues that “interpretability is at the core of communication and is more important than mere intelligibility or comprehensibility” (p.274). Furthermore, Levis’ (2005) 3x3 “World Englishes Speaker-Listener Intelligibility Matrix” (Figure 3) ties intelligibility to contextual factors, setting directions for intelligibility research. This matrix expanded Levis’ previous intelligibility model solely based on dichotomy between native and non-native speaker, highlighted a sociocultural context-sensitive research direction and acknowledged the dynamic of the context and different issues in intelligibility that occur in each cell.

SPEAKER	LISENTER			
		Inner Circle (IC)	Outer Circle (OC)	Expanding Circle (EC)
	Inner Circle	IC-IC *(NS-NS)*	IC-OC	IC-EC *(NS-NNS)*
	Outer Circle	OC-IC	OC-OC	OC-EC
	Expanding Circle	IC-IC *(NNS-NS)*	EC-OC	EC-EC *(NNS-NNS)*

Figure 3. 3x3 “World Englishes Speaker-Listener Intelligibility Matrix” (Levis, 2005)

Intelligibility research undertaken without involvement of inner circle speakers suggests that interaction is the processing of contextual factors including speaker, listener, and environmental factors differently than inner circle speakers (Deterding & Kirkpatrick, 2006). Analyzing group conversation by speakers from the ten Association of Southeast Asian Nation (ASEAN) countries, Deterding and Kirkpatrick found some of the shared pronunciation features by ASEAN speakers actually enhance intelligibility, such as the “avoidance of reduced vowels in unstressed syllables” and “clear bisyllabic enunciation of triphthongs” (p. 406). The former feature is found in many New Varieties of English (NVEs), such as those of the Caribbean and West Africa, which leading authors to speculate about overlap between ASEAN English varieties and NVEs around the world. The “clear bisyllabic enunciation of triphthongs” shows tendency of ASEAN English variety to insert a [w] between the diphthongs. So “our” and “hour” are pronounced as /ɑʊwə/ (ibid,p.398). On the other hand, features of pronunciation that are not shared by the speakers cause a breakdown in the communication. Thus, it was suggested by the authors that speakers travel and do business between ASEAN countries do not have to rely on external norms, but to develop skills at accommodation of local varieties.

The accommodation skills are best demonstrated by the ELF research that goes beyond merely single linguistic feature and finds intelligibility, and higher interpretability, is possible through active involvement between speakers using strategies such converging, negotiation and discourse strategies when speakers do not share the same language and resort to English for communicative purposes (Firth, 1996; Meierkord, 1996, 2002; House, 1999, 2002, 2003; Wagner & Firth, 1997). Seidlhofer (2004) summarizes the recent research in expanding circle countries and generalizes the communication of English as Lingua Franca (ELF):

> Misunderstandings are not frequent in ELF interactions; when they do occur, they tend to be resolved either by topic change or, less often, by overt negotiation using communication strategies such as rephrasing and repetition. (p.218)

Jenkins (2006) in her analysis of UCLES (now Cambridge ESOL) Certificate of Advanced English speaking exam found three communication strategies examinees employed to ensure intelligibility: *converging on one another's form, converging on a more target-like form,* and *avoiding certain forms*. Converging on one another's form is to replicate the speaking features of another, so the speech is more intelligible to their interlocutor, even though the replication of "non-standard" features. The second strategy, converging on a more target-like form, arises most likely due to examinees' perceiving L1-like form as threatening intelligibility for a specific interlocutor. In this case, Jenkins is talking about examinees' awareness to make their speech understandable to raters that assess their speaking performance given that the more native-like form may be favorable to gain higher scores. Nevertheless, when the same examinee paired with others from the same L1, the first type of accommodation strategy was used. The last accommodation strategy to ensure intelligibility found in the analysis is to avoid idiomatic language to avoid communication breakdown.

In the non-test contexts, Meierkord (2000) observes a high degree of negotiation and collaborative achievement as revealed by participants' choice of safe topics and using politeness strategies such as back-channels and routine formulaic expressions to assure maximum intelligibility. Furthermore, House (2003) demonstrates learners of English from different countries in Germany employing pragmatic strategies valued in their own culture to facilitate communications with others. The "topic management strategies" displayed by three Asian students helps maintain the conversation flow by following each participant's own agenda. The second feature is an echoing of the previous speaker's statement, which gives the previous speaker credits as a polite convention of Asian cultures. The third trait is "solidarity and consensus-orientation" (p.569), which displays an Asian style to avoid potentially troublesome remarks and maintain the pleasantness of the group talk. In all these cases, speakers bring their cultural ways of interacting to communication in English and these ways of interaction also serve to negotiate the difference and enhance comprehensibility.

Prior to wrapping up the review of WE, it is noted that Indian English will be used in this dissertation as an elicitation stimulus. Our understanding of the depth, function and range of Indian English has greatly enhanced as a result of insightful and prolific research by Indian researchers (Bhatt, 2001; D'Souza 2001; Kachru, 1985, 1988, 1992, 1997; Mesthire & Bhatt, 2008; Y. Kachru, 2005). In ELT, TESOL teachers from India have sent to expanding circle countries to teach locals English (see *http://asiancorrespondent.com/23123/indians-to-be-hired-as-english-teachers-next-year/*). Despite all these, when it comes to university admission in the US, Indian students are commonly required to take language proficiency tests, such as TOEFL or IELTS, as a proof for English proficiency. Indian teaching assistants (TAs) are also routinely tested for spoken English proficiency. Even though the TA spoken tests probably have practical needs to screen TAs whose speech is less intelligible to undergraduate students in the U.S, what behind this language requirement, on an ideological level, suggests institutions' varying degree of recognition of Indian English status.

World Englishes and Language Assessment

Recent debates on assessment of English proficiency have revolved two issues: What norm should be adopted in the test? What do we mean by English language proficiency? Two positions that represent different ideologies and approaches to test administrations are that of the standard English perspective (Elder & Davies, 2006; Elder & Harding, 2008) and WE perspective (Canagarajah; 2006; Davidson, 1994, 2006; Jenkins, 2002, 2006; Lowenberg, 2002; Spolsky, 1993). The standard English perspective argues that a single norm should be used to judge examinee English language proficiency whereas the WE view worries that a single norm ignores the linguistic richness in the current English global spread and biases against examinees who use or were brought up with different norms. Both views raise arguments centering on test fairness yet with different focus on implementation of a fair approach to testing. Some WE views arguing from critical ideological standpoints criticize that maintaining or imposing only one norm for assessing the tests as a form of continuing US or UK imperialism. Spolsky (1993) claims that English tests have long been used to perpetuate American and British world-views and cultures, leading to unease over standards based on contemporary measurement imperialism. This imperialism neglects the linguistic richness brought on by WE in order to uphold a set of norms that do not necessarily represent the diverse English varieties spoken around the world. Furthermore, Davidson (1993) comments on "the prevalent imperialism of major international tests of English" and argues the power of testing agencies and their positivism stance leads to difficulty in producing tests that are WE oriented. He says "several large English tests hold sway world-wide; tests which are clear agents of the English variety of the nation where they are produced. These tests maintain their agency through the statistical epistemology of norm-referenced measurement of language proficiency, a very difficult beast to assail" (p.119-20).

Arguments about biasing examinees that are assessed against a single norm, commonly American English, are concerned with the discrepancy between American English and examinees' variety, which threatens the scoring validity. Lowenberg (1993) provided examples on the TOEIC reading test items. He collected examples that focus a morphosyntactic, register, and style differences between inner-circle and non-inner circle varieties and analyzed problematic TOEIC items which could lead to more than one correct response thereby negatively affecting the inferences from test scores about examinee's language proficiency. His evidence shows that potential responses are commonly used in examinees' speech communities and are derived from the similar linguistic formation processes as those in inner-circle varieties. However, large-scale English proficiency tests allow only for one single norm to determine right or wrong answers irrespective of the large varieties of English used in examinee speech communities and their sociocultural language use. In other words, as argued by Davidson (2006), right answers are dependent upon a match-up of the test norm (e.g. American English) with examinee's norm (e.g. Singaporean English). If tests are used in a different setting or by different varietal groups of examinees where some other norms apply (e.g. Indian English), then items that are valid for one group may be invalid for another group when the former group shares a norm close to the test norm.

As far as unintended consequences of imposing a single norm in the test, an unintended message is sent about the superiority of test norm and thus the test may stultify language change and diversity. Examinees have to bow to ‘correct’ forms in order to obtain high test scores. A small-scale study conducted by Kenkel and Tucker (1989) analyzed placement essays written by Nigerian and Sri Lankan (outer circle) students at an American university by comparing the results, firstly, by using all American standard only, and secondly, using students’ local norms including spelling and punctuation conventions. The findings showed that around 55% of the Nigerian students had their placements changed if their local varieties were taken into consideration. The influence of exclusion of students’ varieties results in not only inappropriate placement decisions, but also carries an implication of linguistic inferiority of students’ varieties. Kenkel and Tucker aruge that when the Nigerian students who had internalized the syntax, phonology, and lexicon of English in their own variety were placed into an ESL class with others such as Saudi Arabia and Japanese from the expanding circle, it implicitly created in them the impression that their English is inferior and inadequate, which may “stimulate responses of hostility, fear, and alienation” (p. 208). Even though on a small scale, with but a total of 25 essays being investigated, the results are meaningful enough to warrant research into the importance of norm group selection and its unexpected social consequences.

Moving beyond which norm or norms to be adopted, some WE researchers take a strong communicative-oriented approach and argue language tests should go beyond assessing individual language proficiency conforming to an idealized native speaker model to an interaction based performance (Canagarajah, 2006; Jenkins, 2006). Canagarajah (2006) suggests that norms are “relative variable, heterogeneous, and changing (p.234)” and indicates that the issue of which norm to be used becomes irrelevant as proficiency means the ability to interact with other speakers between different varieties of English and different speech communities. He urges language testers

> “to shift our emphases from language as a system to language as social practice, from grammar to pragmatics, from competence to performance. Of course, these constructs are not exclusive. However, bias in language teaching and testing circles is still very much on the first construct in each pair. Defining language use as *performative* involves placing an emphasis on the second construct in each pair and considering how language diversity is actively negotiated in acts of communication under changing contextual conditions” (p.234).

On a similar standpoint, Jenkins (2006) suggests the test should be ‘communicatively motivated’ (p.48) and argues that testers need to respond by taking account of the language variability and not penalizing test takers for employing it with communicative success.

The arguments about communicative-based language test and the unfair test results deriving from assessment against a single norm make it reasonable to expect some changes in language assessment gearing toward a more sociocultural sensitive test design and practice. International large-scale tests that respond to current changes of global English spread can be seen in the delivery and policy of one such exam: the IELTS. Taylor (2002) points out that each of the four sections of IELTS reflects the English language variations to represent content and linguistic features in the context that the IELTS intends to serve, that is the UK, Australia and New Zealand. In the receptive skills (i.e. listening and reading), variations can be found in the inputs, such as spelling, lexis, grammar, pronunciation and discourse. As for the production skills (i.e. speaking and writing), the standard against which the test assesses examinees is based on the following guiding principle (ibid, p.20):

> “Candidates’ responses to tasks are acceptable in varieties of English which would enable candidates to function in the widest range of international contexts.” [taken from the examination handbook]

Despite a seemingly promising language assessment practice to embrace and acknowledge WE, particularly led by one of the international large-scale tests, the current testing practice and stated policy vary considerably across test providers. Taylor (2002) reviewed the current language tests and found some test providers restrict themselves to standard American English, such as tests offered by the Educational Testing Service and Michigan Test Battery, provide alternative test versions (e.g. British and an American version – LCCI's ELSA) or imply they are not biased against any variety of English (TOEIC). This seems to suggest an implementation of socio-linguistic sensitive tests still has a long way to go. Elder and Davies (2006) attribute the challenges of WE language assessment to the constraints of language testing and ethical responsibility of ensuring test fairness, which supports their arguments for a need of standard English in the language testing practice and design.

From the measurement perspectives, Elder and Davies (2006) argue that essential requirements in ensuring test quality and fairness bring certain constraints on test development, which have resulted in a conservative stance of language testing community to handle WE. These requirements are (1) construct validity, that is, what is to be measured given the test purpose and context; (2) fairness, referring to bias-free results regardless of individuals or examinee groups; and (3) accountability to stakeholders. In a fuller discussion on each of the requirements (Elder & Harding, 2008), the authors first forcefully argue uncertainty examinees face about what to be assessed if the norms, as a core principle of English as International Language (EIL) communication, are fluid. Questions such as, 'what is considered appropriate language use?' 'What standard is used to judge the speaking performance?', lead testing agencies to rely on well-established inner circle variety for greater certainty. Relevant to this is raters' readiness for handling multiple varieties of English in the oral proficiency tests (Davies *et al*, 2003). Score validity was questioned given two uncertain scenarios: (1) raters may be uncertain between examinee nativized variety of English and incomplete language learning; and (2) rater perception of WE and the extent to which they accept the variety may vary. With regard to the second requirement, fairness, Elder and Harding take the TOEFL listening section as an example. Instead of using multiple accents, the listening section only includes educated speakers from the US, British or New Zealand to not to disadvantage examinees who are more or less familiar with particular accents if multiple accents are used. The standard Englishes in this case chart a safe and neutral approach if among them is a variety that most of the examinees are familiar with. Elder and Harding argue the choice of the standard English is driven by fairness rather than "a view that local or non-standard varieties of English have no claim to legitimacy" (p.34.5). To further defend their stance of standard English choice, the authors explain the issue of accountability, claiming that powerful and ideological attitude held by examinees and other stakeholders may justify the preference of standard English adopted in the test. Elder and Harding provided two examples on tests developed in Indonesia and Hong Kong, respectively, based on local norm and context, stating tests from inner circle still gains favor by local stakeholders and even examinees. Based on the three constraints above, Elder and Harding made an affirmative concluding remark to question the feasibility of WE oriented language assessment:

> " . . . the field of language testing is steeped in the tradition of psychometrics, and, as language testing practitioners ourselves, we can attest to the fact that there will not be a revolution in language testing with respect to embracing EIL. Language testing is, after all, often concerned with making decisions that can affect people's lives. As with any other serious area of policy-making, changes to language testing policy must be evidence-based, and may evolve slowly in response to changes in social mores" (Elder & Harding, 2008, p.34.8).

With regard to the call for communicative-based assessment raised by WE perspectives, researchers who state a standard English view defend their positions by arguing the fact that language assessment has in fact moved away from the native speaker model toward communicative-oriented assessment valuing the interactive strategies that examinees employ for achieving communication goals (Elder & Harding, 2008; Taylor, 2006). This view is evidenced by three changes: (1) the emphasis on *can-do* statements; (2) the inclusion of communicative assessment tasks; and (3) shifts of research focus. The first change can be observed from the rating criteria for English oral proficiency assessment to judge performance against linguistic forms along with other factors such as coherence, discourse management and interactive strategies (Taylor, 2006). Frameworks, such as the influential Common European Framework of Reference (Council of Europe 2001) that includes a list of can-do statements, suggest a shift of focus from native speaker model and form to function and communication. The second change is an increase of use of pair work to assess intercultural communication skills in a number of high-stake tests, a way to reveal language testers' awareness of WE. The last change is the research focus on effects of accents on listening test scores (Harding 2008; Major, Fitzmauric, Bunta & Balasubramanian,2002) and rating performance between native and non-native speakers of English (Kim 2009).

Approaches to Fairness

The arguments of two contrasting views on norm selection are both based on the endeavor to produce a fair test result taking a wider context of global English use into consideration. This reflects the post-Messick re-conceptualization of validity, which is now treated as unitary concept and goes beyond the traditional psychometric inquiry to look broader and deeper issues of the social dimension of the test, brining context to the larger research agenda, seeking the value and consequences of score interpretation and use. The latter feature is argued as "a radical aspect of Messick's discussion as it opens the whole of language testing to a discussion in terms of values, and hence invites the kind of discussion familiar within critical applied linguistics more generally" (McNamara, 1998, p. 305). Issues of test fairness(Davies, 2003; Kunnan, 2002, 2004; Shohamy, 2001), along with other research that address the two innovative aspects of test validity as Messick urged are emerging to invigorate language assessment inquiry: implementations of socially-oriented tests measuring pragmatics (McNamra & Rover 2006), test washback (Alderson & Wall, 1993) which refers to the effect of the test in classroom context, test impact in the school and political context (Shohamy, 2001; Spolsky, 1995), and cultural aspects (Elder, 1997, 2000), policy (Fulcher, 2004, 2007). More recently this line of thought prioritizes intended effect as a guideline for test design and development (Fulcher & Davidson, 2007;Kim 2008).

From the examinees' perspective, test fairness may be one of the major concerns because they want to be assessed on a fair platform and do not want to be biased. For other stakeholders, arguments based on collected evidence about actions taken by testing agencies to increase test fairness impact professional and public attitudes and decisions about test use. Three documents, the Code of Fair Testing Practices in Education from the Joint Committee on Testing Practices (1988, 2004, hereafter, *Code*), Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999, hereafter *Standards*) and ETS standards for Quality and Fairness (2002, hereafter *ETS standards*) have brought test fairness to the forefront and tightly connected fairness to the investigation of different examinee group performance. This group difference includes examinee language, culture, age, disability and socioeconomic status. In some way, it links WE to test administration and use in that examinee linguistic backgrounds are taken into account as a part of a testing cycle, making testing professionals more socially responsible if actions are taken for meaningful investigation of group difference. Comparing the three documents, *ETS standards* is the only document that explicitly treats fairness in the entire testing procedures: design, development, administration and use of test and test scores, which is a distinctive difference as compared to *Code* and *Standards* that treat fairness as a 'after test' quality. Furthermore, the role of rater and examinees' non-native status in *ETS standards* are specifically highlighted as part of essential data to address fairness. It states that "training designed to eliminate possible rater or administrator biases" (p.19) is necessary which places fairness investigation before a testing event, in keeping with Weir's (2005) call for "a priori validity evidence" (p.17) to evaluate evidence to support inferences from test scores right from the initial test design process. Examinees' needs and linguistic difference are also part of considerations in the test development which demonstrates testing agency's acknowledgement of the huge number of non-native English examinees and to incorporate their voices into the testing practice (Llurda, 2004). Standard 4.7 requests that test developers:

> "Consider the needs of nonnative speakers of English in the development and use of products or services. For assessments, reduce threats to validity that may arise from language differences" (p.21)

With regard to conceptualization of fairness in validation, it may be best analyzed by fairness framework as proposed by Kunnan (2000; 2004). Kunnan's "test fairness" framework (2004) links validity and consequences, treats fairness in the whole testing practice rather than just test itself. It consists of five qualities: validity, absence of bias, access, administration, and social consequences. Quality close to examinees' linguistic resources is not explicitly stated but rather implied by the "absence of bias", in which the different performances resulting from group membership is suggested for further investigation.

Though the *Code, Standards, ETS standards* and Kunnan's fairness framework either explicitly or implicitly address investigations of group difference in examinee linguistic backgrounds to ensure fairness, they do not conceptualize fairness consistently and treat fairness differently in relation to validity. Xi (2010) summarized the three different views on conceptualizations of fairness in language assessment. First, fairness is an independent test quality that is separate from validity. The *Code* and the *ETS standards* represent this view.

Fairness is classified as an independent test quality given that the documents suggests test developers and encourages test users to conduct sensitivity reviews of test materials to ensure the fair test results (i.e. *Code*) and assessment products and services need to satisfy the fairness standards (i.e *ETS Standards*). The second view considers fairness as an all-encompassing test quality. Kunnan's fairness framework which subsumes and goes beyond validity is representative of the second view. Additionally, though not explicitly defined the scope of test fairness, Xi notes that McNamara and Roever's (2006) primary focus on social dimensions of language testing to investigate item bias investigation through social approaches (i.e. fairness document review) and traditional psychometric methods (e.g. Differential Item Functioning), manifests fairness as an overarching test quality. The last view treats fairness as linked directly to validity and the *Standards* best represents this view. Each type of validity evidence was discussed and its corresponding concern with fairness is further expanded in the separate section on fairness.

The fairness documents and frameworks demonstrate testing professionals' promising work toward Messick's call for more socially response testing to judge the validity or quality of the test in addition to a traditional positivist approach. Nevertheless, these fairness documents and frameworks were criticized for being conceptual and for being a lack of a systematic approach to integrate all aspects of fairness practices and investigations as well as not setting priorities for fairness investigation (Xi, 2010). Xi recently proposed a "fairness argument" (p.155) that attempted to offer guidelines for practical fairness investigation. Her work reflects recent discussions on approaches to validation that have been endorsed the work by Messick's conceptualization of unified theory of validity, Kane's argument-based approach (Kane,1992; Kane *et al*, 1999; Kane, 2001, 2002, 2004) to provide practical procedures for validation process and Toulmin's model of inference (2003) that substantiates the structure for the argument-based validation process (Bachman, 2005; Chapelle, Enright & Jameson, 2008; Kadir, 2008) and test design (Mislevy, Steingberg,& Almond,2002). In Messick's (1989) views, validity is not an inherent property of the test itself, rather it is the degree to which inferences about examinee language proficiency from the test results are justifiable. Following the modern validity precursor, Cronbach (1989), Messick (1989) urges empirical validation that requires theoretical rationales and collecting empirical data for defensible interpretation of test score and score use. A pragmatic and systematic approach to infer examinee language proficiency based on test scores are seen in Kane's four types of inference in the chain of reasoning (i.e. evaluation, generalization, extrapolation and decision), which he called " an interpretative argument". This is defined as a "chain of inferences from the observed performances to conclusions and decisions included in the interpretation" (Kane *et al*, 1999, p.6). The types of inference are extended in other studies (Chapelle *et al*, 2008) to keep seeking plausible interpretation of test score and use. Kane used Toulmin's model of argument for his approach to analyze arguments.

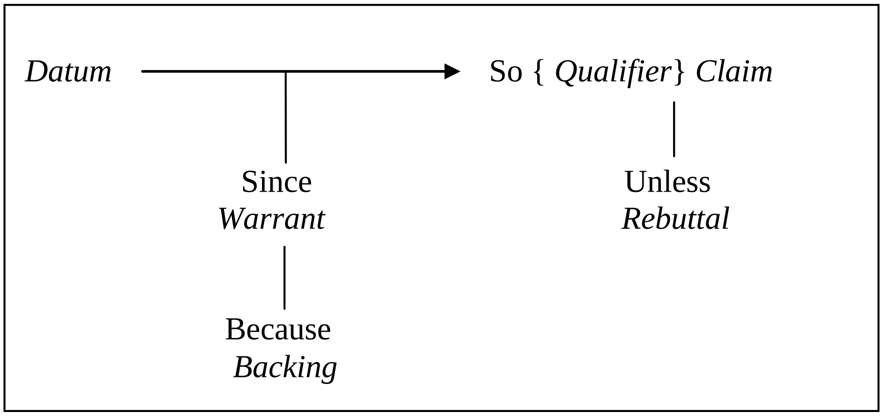


Figure 4. Toulmin's model of argument

The model presented in Figure 4, sets a *claim*, or the conclusion of the argument that is intended to justify. This is supported by *datum*, the available information or evidence. A *warrant* is required for a datum to link a claim, which yields a justification of the claim using the data. Warrants themselves need to be justified and require evidence, theoretical support, prior research or *backing*. Nevertheless, when the warrant fails to link datum to a claim, a *rebuttal* provides justification accordingly with supporting data. The rebuttal may lead to construct-irrelevant variance or construct under-representation in language testing context (Fulcher & Davidson, 2007). Finally, a *qualifier* indicates the strength of the claim and can be reported in the form of, for example, inter-rater reliability for the student speaking performance.

Returning to Xi's fairness argument (Xi, 2010), she demonstrated that a fairness argument could be built upon the validity argument by using the validation of TOEFL iBT test as an example that extends the typical inferential bridges in Kane's work (Chapelle *et al*, 2008). Xi used a series of rebuttals in each inference link as a way to compose fairness argument in an attempt to challenge the comparability of the score, score interpretation and use and consequence for different relevant groups. Two types of rebuttals that would weaken the conclusion were proposed:

> Type 1 rebuttals weaken the conclusion for all test takers and thus a lack of counter-evidence tends to reduce the force of this conclusion for the whole test-taking population. Type 2 rebuttals, on the other hand, point to the specific examinee groups to which the conclusion may not apply or to which it may not be completely tenable (p.163).

The type 2 rebuttals are specific to the fairness argument as they concern group difference, such as those stated in the fairness documents and framework reviewed above. In Xi's example where the test scores reflect the quality of language performance on relevant tasks in an academic setting, the type 2 rebuttal would be that inappropriate test content leads to group difference in scores, which subsequently affects the prediction in performance on relevant tasks in an academic setting for test takers residing in and outside the USA. As an application of Kane's argument-based approach, the fairness argument prioritizes fairness investigations in order to provide backing to refute the rebuttals and thus enhance the overall validity argument.

Similar to Kane's argument-based approach, Xi's fairness argument seems to mainly focus on the test results when the test administration is complete and does not take test development into investigation, limiting in scope of fairness investigation. Thus, efforts, actions or training taken to enhance rater consistent performance were not part of fairness investigation. Furthermore, Kunnan (2010) comments that the fairness argument nests fairness within validity argument and so diffuses the focus of a fairness research agenda and reduces the role of fairness to only depend on the validation arguments rather than to have its own agenda. Nevertheless, Xi's fairness argument should be given credit given its attempt to provide practical guidance on fairness investigation that has long been on a conceptual level as shown in the fairness documents and frameworks; it may also motivate more research for developing practical approaches to fairness investigation.

Language Attitude Study

Attitude, Attitude Formation and Attitude-Behavior Relationship

Concerns raised by language testing professionals about rater perception and acceptance of WE transferring a potential negative impact on scores (Davies *et al*, 2003) and thus weakening the inferences about examinee English language proficiency can be approached via language attitude study. Language attitude studies have revealed that non-native speakers are viewed unfavorably by listeners, regardless of whether they share the same variety with the speakers or not, and this unfavorable attitude lead to negative evaluations of speaker competency.

The structure of attitude, from psychological perspectives, has been identified to be comprised of one or a combination of the following three components: affective, cognitive and behavioral (Ajzen & Fishbein, 1980; Albarracin, D. Johnson, B.T, & Zanna, M.P, 2005; Cargile, Giles, Ryan & Bradac, 1994; Trafimow & Sheeran, 1998). Affect refers to the feelings, cognitions to an individual's belief structure and behavior is the tendency to behave in a certain way. In terms of 'language attitude', it refers to "consciously-held ways or beliefs about a specific language or to an orientation (positive or negative) towards a specific language that influences the individual's evaluation of that language and its speaker" (Cluver, 2000: 315). It has been assumed that language attitudes are long-term phenomena that tend to become more specific over generations and tend to be unchangeable if they exist for a long time. Language attitude is affected by complex factors, such as experience and education; nevertheless, Cargile *et al.* (1994) argue that many empirical studies treat language attitude as simple responses to language stimuli and advocate conceptualizing language attitudes as a process, not "a singular, static phenomenon. Rather, they affect, and are affected by, numerous elements in a virtually endless, recursive fashion" (p.215). They proposed a "social process model of language attitude" (Figure 5) in an attempt to capture multiple factors involved in the application of language attitudes in specific situations. According to the social process model, factors affecting the formation of language attitude include characteristics of speakers, listeners and contextual factors. In terms of individual influence, certain attitudes are evoked as an interaction between speaker linguistic behavior and characteristics, such as physical features, and listener characteristics, including listener goals, emotional state, expertise and social identity. In Figure 5, two-way arrows are drawn between speaker and listener, which is meant to

> " indicate that speaker language does not inevitably trigger certain attitudes within the hearer, but rather hearers are actively involved in the process of selecting and attending

> to those language behaviors that meet their needs. Language can indeed lead to particular attitudes, but hearers can also choose those language behaviors around which they construct their attitudes and evaluation" (p.218).

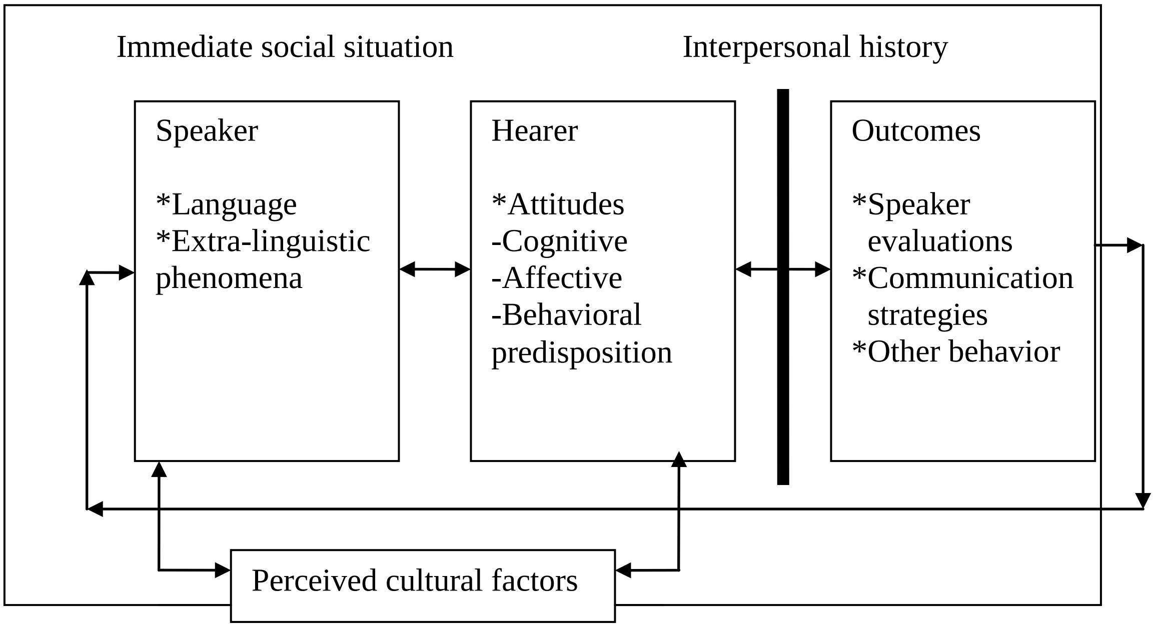

Immediate social situation
Interpersonal history
Speaker
*Language
*Extra-linguistic phenomena
Hearer
*Attitudes
-Cognitive
-Affective
-Behavioral predisposition
Outcomes
*Speaker evaluations
*Communication strategies
*Other behavior
Perceived cultural factors

Figure 5. Social process model (Cargile *et al.* 1994)

With regard to contextual factors, expression of attitudes is mediated by individual's perceived interpersonal history, immediate social situation and perceived cultural factors. Dawning on uncertainty reduction theory that increasing interactions between strangers would reduce levels of uncertainty about the other (Berger & Bradac, 1982), the social process model argues that listener attitude is most likely to affect his/her behaviors in contexts of low familiarity. In terms of immediate social situation, language forms may be evaluated positively (e.g. slow speech to aid in comprehension) in one situation but negatively when the situation changes (e.g. slow speech at a cocktail party). Taking a wider context into account, the formation of attitude is mediated by listener's perceived cultural view, which refers to historical relations between listener (e.g. American) and speaker (e.g. Chinese) groups and their relative sociostructural strengths.

The influence of attitude toward a behavior and its relationship between actual behavior has been established by psychologists (Ajzen & Timko, 1986; Albarracin *et al*., 2001; Fazio *et al.*, 1989; Fishbein *et al*., 2000; Hrubes *et al*, 2001). One frequently cited theory to support the relationship between attitudes and behavior is the Theory of Reasoned Action proposed by Ajzen and Fishbein (1980). The Theory of Reasoned Action suggests that behavior is directly influenced by intentions that results from attitude toward a specific behavior, subjective norms associated with that behavior, and an individual's perceived behavioral control over that behavior. Empirical studies in support of the attitude-behavior relationship demonstrate the negative attitude towards the nonnative speakers and subsequent unfavorable evaluations of speaker himself or his competency. Rubin's (1992) study revealed that, with standard American English as the only speech stimulus, participants responded that they were listening to nonstandard speech when they were faced with a photograph showing an ethnically Asian instructor. More seriously, listening comprehension appeared to be undermined simply by identifying (visually) the instructor as Asian. In terms of actual behavior, participants' perceived TAs' accents to be foreign undermine those their evaluation of TAs. Similarly, in a study of English native speakers' attitudes toward Korean, Lindemann (2002) requested participants to complete a map task, pairing up a NS of English and a Korean. Attitude held by NSs of English towards Korean appeared to affect both their perceived and actual success of interaction. Lindemann found that the attitude-comprehension relationship "is mediated by the native speaker's choice of strategies" (p.419). It was found that NS of English with negative attitude used "avoidance" and "problematizing strategies" (p.433). Avoidance strategies were to refuse providing feedback to one's partner when there were difficulties in understanding. NS of English who employed problematizing strategies did not give credits to or omit acceptance of

NNS partner’s contributions to the communication. Lindemann (2002) attributed this attitude-comprehension relationship to an ideology that views “the non-native speakers as a subordinate group” (p.439) in the US society.

Three approaches relevant to the study of language attitudes have been used: content analysis of societal treatment, direct measurement and indirect measurement (Sebastian, 1982). In the content analysis of societal treatment, reviews of laws and policies regarding language use in the public domain are to examine language maintenance and shift. This type of analysis forms the basis for descriptions of standard language and language change. The direct measurement technique evaluates language attitudes by use of questionnaires, either in written form or in individual interview. This method tends to elicit response in participants’ beliefs. In the indirect method, the participants are not aware of their attitude being examined. The measurement techniques include speaker evaluation studies, such as matched-guised studies, in which participants are asked to evaluate different varieties of a language spoken by the same speaker. This method observes the socio-psychological perspective on language attitudes.

Attitude Study within Language Assessment Research

Within language assessment research, recent empirical studies that concerned the impact of WE began to center attitude as research agenda to explore rater or examinee attitude towards varieties of English, rating performance between different nationality groups and attitude-rating behavior relationship. A recent dissertation conducted by Kim (2005) appears most relevant to the current dissertation. Kim (2005) examined raters’ backgrounds and their attitude towards WE in language testing and how these variables interacted with ratings awarded to six Korean students’ speech performance on the Test of Spoken English (TSE) picture description task using holistic and analytic scales. Four groups of teacher raters were recruited in the study: NS of English in the US, NS of English in Korea, NNS of English in Hong Kong, and NNS of English in Korea. Rater attitudes toward WE were assessed by a questionnaire. According to rater response, raters were categorized into three groups: positive, neutral and negative. The results showed no significant interaction between raters' language backgrounds and their attitudes toward WE despite the fact that NS in the US had more positive attitudes towards WE, as compared to other rater groups. In terms of the effect of rater attitude on their rating judgment, it was found that raters achieved quite similar rating performance on the holistic ratings. Nevertheless, raters' different attitudes toward WE significantly affected the analytic ratings on grammar, rate of speech, and task fulfillment. Raters who were labeled as 'positive' showed more leniency in rating of three criteria as noted above.

Following Rubin (1992), Kang (2008) compares college student rater attitude toward two ethnic groups, Asian and Caucasian, and found no significant effect on raters' attitudes toward ethnicity. In terms of rater background characteristics in relation to rater evaluation of international teaching assistants (ITAs), NS/NNS of English status, language teaching experience and number of NNS contact significantly related to the prediction of variances in ratings on comprehensibility, accentedness and language proficiency. NNS raters consistently rated lower than NS on all language proficiency criteria, including pronunciation, grammar, vocabulary, speech rate, communication skills, expression of ideas, word choices and overall proficiency. She used an intervention to increase interaction between undergraduates and ITAs and improved the former's attitude-behavior relationship. She found "informal and pleasant contact with interpersonal intimacy and equality can bring a positive change in undergraduate attitudes toward ITAs and consequently influence undergraduates' perceptions of ITA speech performances. . . "(p.200).

Relevant to attitude study, Harding (2008) first looked into the use of speakers with L2 accents of English on an academic listening test. It was found that a shared-L1 or familiarity effect was not pervasive, but may exist when certain conditions were present on a listening test relating to task demands, speaker pronunciation and the linguistic demands of the text. Findings also showed that examinees overall held reasonably positive attitudes towards L2 accented speakers, and that there was no clear relationship between attitudes towards speakers and subsequent performance on a listening test featuring that speaker. Aiming to examine rater perception of examinee speaking proficiency on TOEFL speaking tasks, Chalhoub-Deville and Wigglesworth (2005) found 124 raters from different inner-circle speaking countries, including Australia, Canada, the UK, and the US, had no significant difference in evaluating ESL examinees' speaking performance. Even though the authors said to investigate raters' "shared perception" (p.383) of examinee speaking proficiency, they did not look at the perception or attitude as conceptualized in other studies reviewed above. The authors argued that similar ratings among raters seem to imply similar perception shared by raters; however, that conclusion must be seen in light of other studies which report that raters may arrive similar ratings for different reasons or focusing on different aspects of language use (Brown, 2000; Brown, Iwashita & McNamara, 2005; Orr, 2002). Overall, language assessment research relevant to attitude study indicates growing interests in placing rater attitude or psychological traits as a potential variable that affects the arguments about the examinee language proficiency. Nevertheless, the study findings may not be generalizable as the studies were conducted in different contexts and the instruments used to investigate rater or examinee attitude were different. The proliferating research conducted in language attitude study better informs language testing professionals of the attitude that NS and NNS listeners hold towards different varieties of English and be aware of the powerful impact brought by one's attitude on subsequent behavioral tendency.

Empirical Language Attitude Study

It is important to stress that the studies reviewed next focus on listeners, ones that are not empowered to award scores and affect speakers' lives, as are raters in the test settings. Nevertheless, the findings provide valuable insights for language testing professionals to consider rater psychological traits when handing varieties of English in the oral proficiency tests. The terms, NS and NNS, shown below are used as originally appeared in the studies and used here again for the purpose of discussion only.

Language attitude research has predominantly focused on accented speech and generated generally consistent patterns of results on listener attitude toward language varieties. Despite different cultures and contexts where the studies were conducted, accented speech is negatively rated and listeners have tendency to favor standard varieties, as measured by a variety of scales. Reviewing a wide range of literatures, Giles and Billings (2004) report that when NSs serve as listeners, speakers of standard variety are typically upgraded on status-related traits, such as confidence, intelligence and ambition. This appears to be the case when the listeners are either speakers of standard or non-standard varieties. In contrast, non-standard speech tends to be evaluated more highly in terms of 'solidarity' when compared to varieties of standard speech. Speakers of non-standard varieties are generally rated highly on dimensions such as honesty and friendliness, particularly when the listeners are learners/speakers of a non-standard variety themselves.

Studies that target non-native speakers as listeners use L2 learners as subjects in the educational settings and yield a set of diverse results. McKenzie (2008) employed a verbalguise study to investigate the attitudes of 558 Japanese university students towards six varieties of English speech: Glasgow standard English, heavily-accented Japanese English, Southern US English, moderately-accented Japanese English, Mid-west US English and Glasgow vernacular. A semantic differential scale (see definition in Chapter Three) was used where items loaded onto two components: competence and social attractiveness. It was found that Japanese listeners particularly favored standard and non-standard varieties of UK and US English in terms of 'competence' and rated the Japanese speaker of heavily-accented English highest on the social attractiveness trait. The results conform to study findings in which native speakers of English were listeners. As the author noted, the complex yet contradictory attitudinal reactions among Japanese learners suggests that the strong Japanese accent may indicate an 'in-group" identity and its pedagogical implication in terms of selection of models of English should be viewed as 'points of reference' rather than 'norms of use' (Quay 2004).

In terms of comparison between NS and NNS listener attitude, Barona (2008) used accented speech produced by speakers of Korean, Spanish and Arabic. Listeners were Chinese, Korean, Portuguese and Spanish respectively from general public in Northern New Jersey. It was reported that all listener groups rated lower about speakers' 'competence', 'integrity', and 'social attractiveness', as compared to the ratings by the NS listeners, indicating NNS listeners' negative feeling towards the accented speech. Furthermore, the three non-native accents were evaluated differently: Korean-accented speech was rated highest on speaker's competency and integrity, followed by Spanish and Arabic respectively whereas Spanish-accented speech was rated the highest in terms of 'social attractiveness', followed by Korean and Arabic-accents. The author suggested the higher acceptability of Spanish and Korean accented speech is attributed to the increasing foreign born Hispanic and Korean population in northern New Jersey, which seems to tie the exposure to accented speech to listener positive attitude.

Nevertheless, when comparing different varieties within a country used by its citizens, the results differed from the general findings and indicated the standard variety does not necessarily gain preference by citizens as far as communication in the wider context is concerned. Kioko and Muthwii (2003) investigated the majority view concerning English used in Kenya and what variety is preferred by Kenyan for use in the formal domains, such as school, law courts and media. The authors looked into Kenyan speakers' attitudes towards three varieties of English: native speaker English (i.e. inner-circle variety), standard Kenyan English, and ethnically marked Kenyan English (i.e. "a variety of English that exhibits salient linguistic features associated with the ethnic language of a speaker") (p. 135). As opposed to general findings that standard variety (e.g. British, American English) symbolizes power, status, and success, the analysis of the questionnaire showed that Kenyan speakers related standard Kenyan English to successful professionals. Furthermore, the result showed that Kenyan prefer a variety less displaying features of Kenya's ethnic languages, called "non-ethnically marked Kenyan English" (p.135). This led the author to claim that "much of the actual identity of the language(s) used is a product of the interaction of the ethnicity factor, the rural-urban dichotomy, and the attitudes that Kenyans have toward the languages within their repertoire" (p.142). Many Kenyans respondents prefer linguistic neutrality when using the English language to fit in a wider world than their own ethnic ones. The linguistic neutrality brings about more unity than the other varieties when interacting with countryman from other tribes.

Moving beyond studying accent in isolation, studies attempted to establish a connection between attitude and listener social identity emerged recently. Guided by social identity theory that individuals exhibit a preference for the variety of language associated with their in-group, Bresnahan, Ohashi, Nebashi, Liu and Shearman (2002) recruited 311 university students identified themselves as white to look into the relationship between attitude and strength of social/ethnic identity. As predicted, the ‘in-group’ is more favored in various occasions. The study showed that friends, the ‘in-group’, were viewed more positively for affect and attitude compared to teaching assistants regardless of accent. Additionally, participants exhibiting strong ethnic identity preferred American English while those with weak ethnic identity were more accepting of foreign accent.

Lindemann’s (2003) study takes a larger concept of language ideology to investigate listeners’ expectation of speaker’s L1 background to relate listeners’ identifications of the speakers’ ethnicity to salient social groups for listeners. Thirty-nine undergraduate NSs of English evaluated the speech produced by ten NSs of Korean and seven NSs of English. Listeners were asked to identify the speaker’s ethnicity and native-speaker status in an attempt to relate the speakers’ ethnicity to salient social groups for listeners. The results showed that listeners usually misidentified the Korean voices as Chinese, Japanese, or “non-east Asians” particularly as Indians. Listeners also indicated negative attitudes to these groups as shown on the low ratings on speakers’ language-focused traits. The author thus suggests that the listeners appear to “identify a generalized ‘foreign faultiness’ rather than a relationship between specific features and speaker traits” (p.359).

Rater Variability

Rater is an important factor for the validity of the performance assessment, that is, oral and written tests. In performance assessment, McNamara (1997) indicated that ‘rating is a result of a host of factors interacting with each other ‘(p.453). He interpreted the rating as a product of an interaction among rater, task, examinee, testing performance, rating criteria and interlocutor. Rater has been the focus of research agendas in that rater characteristics and the way they use and interpret the rating scales play an influential role on the rating results. Models of speaking test performance as put forward by McNamara (1996), Skehan (2001) and Fulcher (2003) keep expanding the scope of speaking test performance and broadening our understanding of the complexity of rater variables that impact the scores. Fulcher’s model explicitly highlights rater characteristics and the importance of rater training to control for the effect of rater variation so as not to jeopardize the fairness of conclusions that we make about individual examinees. Studies reviewed below highlighted different aspects of rater variability and their effects on test scores, all of which will be used as variables in current dissertation to study the relationship between the variables, rater attitude towards WE and test scores.

Effects of rater educational and professional experience. Research is inconclusive regarding how raters’ professional experience affects their ratings of examinee oral proficiency. In Cumming’s (1990) investigation, expert raters, as compared with novice raters, may be less influenced by surface language structures and more capable of examining content, language use and rhetorical organization concurrently. Expert raters tend to “have a much fuller mental representation of ‘the problem’ of evaluating student compositions” . . . whereas novice raters tend to “edit student texts extensively” (Cumming, 1990, p.43). Studies by Barnwell (1989), Chalhoub-Deville (1995), and Hadden (1991) all found that classroom teachers and nonteaching native speakers differ in their assessments of learner’ second language oral ability. It was found that the naive native speaker raters were relatively stricter than the trained rater group (Barnwell 1989), nevertheless teachers were more critical of students’ linguistic abilities than were nonteachers (Hadden 1991). Chalhoub-Deville’s (1995) investigation found contrasting findings where the three rater groups (teachers of Arabic, nonteaching Arabs living in the USA and nonteaching Arabs living in Lebanon) vary in their expectations and evaluations of students' speech performances. The teaching rater group focuses more on communicative aspects of the language whereas the nonteaching groups appear to emphasize students' grammar-pronunciation features.

Effects of residency. Kim (2006) found that residency is a factor that contributes to score difference. Four groups of raters: NS in the US, NS in Korea, NNS in Hong Kong and NNS in Korea, rated Korean students' speech samples on the TSE picture-description task using holistic and analytic scores as measures. The results suggest that even though no significant difference was found on the holistic ratings, raters awarded scores significantly different on analytic ratings such as grammar and organization. The group of NSs in Korea provided lower mean scores on grammar than the other three rater groups. In terms of organization, the group of NSs in the US gave higher mean scores on organization than the other three rater groups.

Effects of rater nationality and native languages. Different results were found with regard to how raters' nationality and native language affect their ratings of examinee oral proficiency. Brown's (1995) results pertaining to the Japanese Test for Tour Guides showed that there is little evidence that native speakers are more suitable than nonnative speakers, or that raters with teaching backgrounds are more suitable than those with an industry background. Similarly, Shi's (2001) investigation into Chinese students' English writing revealed that scores were not significantly different between native speaker and non-native speaker teachers despite the finding that the two groups of raters attend to different aspect of writing. Chalhoub-Deville and Wiggleswoth (2005) investigated whether there was a shared perception of speaking proficiency among raters from different English speaking countries: Australia, Canada, the United Kingdom, and the United States, when rating speech samples of international English language students. They found that the UK raters were the harshest and the US raters were the most lenient.

Effects of gender. O'Loughlin (2002) examined the effect of raters' gender on test scores. He argued that the IELTS (which has a person serving both as an interviewer and as a rater) raises the question of whether a gender affects the rating decision of the examinee's oral proficiency level. Sixteen students' (8 males and 8 females) had a practice IELTS interview on two different occasions, once with a female and once with a male interviewer. The 32 interviews were tape-recorded and reevaluated by 4 raters (2 males and 2 females) and then analyzed using multifaceted Rasch bias analyses. O'Loughlin found that gender did not have a significant impact on the IELTS ratings.

Methodologies to investigate rater orientation and decision making

Recent studies began to use verbal protocol to investigate rater orientation and decision making in second language speaking test use and elicited detailed and valuable information on rater decision making process that quantitative studies of test scores cannot necessarily explore. One advantage of verbal protocol reports is that subjects are likely to remember the original behavior if presented with the same stimulus (Ericsson & Simon, 1984). In a study about the IELTS speaking test, Brown (2000) used stimulated verbal recall (DiPardo, 1994) and found raters interpreted the rating criteria differently and included rater's own criteria that were not specified in the rating criteria. She argues that rater variability cannot be avoided and their "individuality and their internal variability" should be allowed and probably a need to "look for other ways to ensure fairness for candidates" (p.81). Similarly, Orr (2002) used retrospective verbal reports to investigate the First Certificate in English (FCE) speaking test and found a wide range of rater variability, even on rater's interpretations of "the model of communicative language ability on which the rating scales are based "(p.153). He calls for further examination into rating scales and investigates a need to make any adjustment. Brown *et al* (2005) used both retrospective verbal reports and a discourse analysis of spoken language produced in the Test of Spoken English found that rater orientation conformed to the actual discourse features of a performance. In support of findings from previous studies, they indicate that ome features of a performance appears more salient to different raters.

CHAPTER 3

METHODOLOGY

Study Design

This dissertation comprised two validation studies: first, the development of Rater Attitude Instrument (RAI) and secondly, an examination of rater scoring tendencies in relation to rater attitudes towards WE. The validation built upon an argument-based approach (see chapter 2) closely linking two processes: the validation process and the investigation of two issues of interest, namely the development of the RAI and seeking rater scoring tendencies based on their attitude towards WE. Using Toulmin's form of inference (2003), each validation study was guided by a claim and supported by warrants, backing, data, and discussions of counter data.

The first validation study develop and validated a battery of instruments, RAI, to obtain a deeper and broader understanding and a better interpretation of the complexity of rater beliefs, affectives, intentions and scoring tendencies. Though some rater/listener attitude studies within language assessment context (Harding, 2008; Kang, 2008) adopted the Speech Evaluation Scale developed by Zahn and Hopper (1985), the extent to which such a scale reflects the same set of attitude evaluation dimensions of oral/writing proficiency raters is uncertain and may fail to serve the needs and contexts of language assessment research. Language attitude studies indicate that in cases where the variety of instruments used makes it difficult to draw solid conclusions (Lindemann, 2005), the development of appropriate instruments is called for and the RAI is timely in proposing a uniform approach for language testers and researchers to engage in rater attitude inquiry within the WE paradigm.

After the RAI was constructed, study 2 was conducted to elicit rater responses to the

RAI and to rate six IELTS descriptive tasks. The results were analyzed and provided guidelines for the selection of raters in the verbal protocol study that sought to identify a rating pattern existing within similar perceptions. A meta-analysis was performed to synthesize all data and provide an interpretation of the rater attitude towards WE and its association with their rating in the IELTS descriptive tasks. Figure 6 presents an overview of the study design.

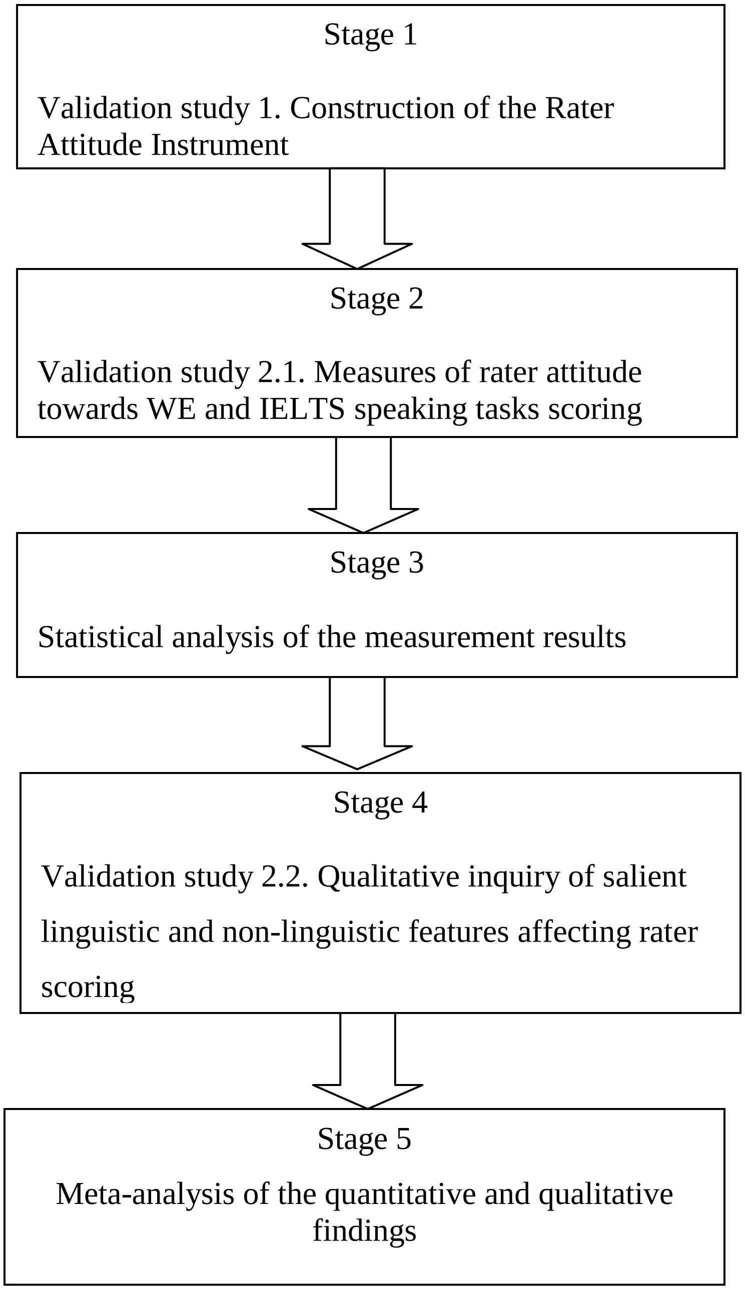

Stage 1
Validation study 1. Construction of the Rater Attitude Instrument
Stage 2
Validation study 2.1. Measures of rater attitude towards WE and IELTS speaking tasks scoring
Stage 3
Statistical analysis of the measurement results
Stage 4
Validation study 2.2. Qualitative inquiry of salient linguistic and non-linguistic features affecting rater scoring
Stage 5
Meta-analysis of the quantitative and qualitative findings

Figure 6. Overview of the study design

Forms of Inference for the Two Studies

Study 1 Procedures

An overview of the description and data collected in each phase during the RAI construction is presented in Table 1. As an initial step, the RAI explored rater attitudes towards WE by means of in-depth interviews with raters of a commercial oral proficiency test and a varietal speaker evaluation study aimed at obtaining rater feelings of the speakers of multiple varieties. The former was part of an early research project in partial fulfillment of the researcher's PhD degree requirements at the College of Education at the University of Illinois at Urbana-Champaign. Informed by the findings of the two empirical studies, a total of 82 items were constructed. In Phase 2, a new group of 20 raters responded to the online RAI and

Table 1

Construction of Rater Attitude Instrument in Three Phases

Phase	Description	Timeline
Phase 1	• Attitude attributes obtained from interviews and the varietal speaker evaluation study • The draft of RAI completed	• Summer and Fall, 2007 • Summer 2010
Phase 2	• Raters of oral proficiency test at three universities recruited. • Raters responded to the RAI delivered online • The RAI modified	• Fall 2010
Phase 3	• IELTS raters and ESL/EFL teachers in the US and India recruited. • Raters responded to the RAI and rated six IELTS descriptive tasks on-line.	• Summer 2011

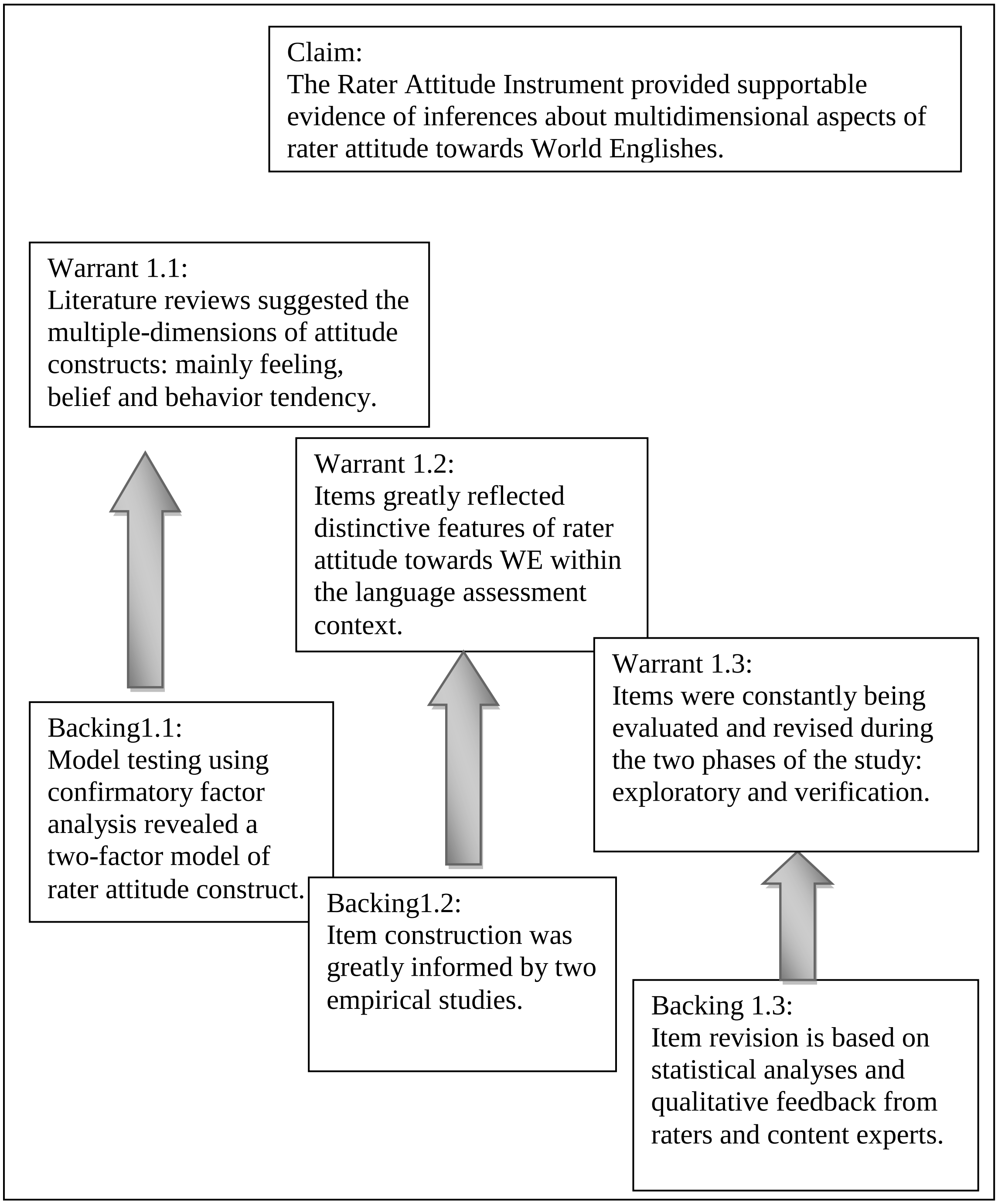
provided feedback on the clarity and appropriateness of each item. Results were analyzed by exploratory factor analysis and item-total statistics to determine the internal structure of the RAI and to ascertain the need to revise the wordings and even remove items that yielded low alpha values or that raters considered as less relevant. In the last phase of verification of the RAI, 96 raters rated six IELTS descriptive tasks each, with very little guidance, and responded to the modified RAI.

Both of the rating tasks were conducted on-line. Findings were analyzed by confirmatory factor analysis to verify the multi-factor model of the rater attitudes towards WE.

Study 1 Forms of Inference

For the two studies, the forms of inference to support the claims were presented with their own stand-alone methods of data collection and analysis.

Figure ***7*** presents the argument structure for study 1. The claim for study 1 was that the RAI provided supportable evidence of inferences about multidimensional aspects of rater attitudes towards WE. To seek warrants to support this claim, the literature was first reviewed which informed a three-dimension construct of rater perception (*Warrant 1*). To evaluate if the three-factor model can be tested psychometrically upon completion of instrument construction, a confirmatory factor analysis was conducted using the Statistical Package for the Social Sciences for Windows Release 15.0 (SPSS Inc., 2007) and AMOS Version 7.0. Preliminary descriptive statistics were calculated and assumptions regarding univariate and multivariate normality were inspected. The CFA models were tested using a common model-fitting procedure: Maximum Likelihood estimation (DeCoster, 1998). Two items with low square multiple correlations were removed to greatly improve the model fit indices.



Claim:
The Rater Attitude Instrument provided supportable evidence of inferences about multidimensional aspects of rater attitude towards World Englishes.

Warrant 1.1:
Literature reviews suggested the multiple-dimensions of attitude constructs: mainly feeling, belief and behavior tendency.

Warrant 1.2:
Items greatly reflected distinctive features of rater attitude towards WE within the language assessment context.

Warrant 1.3:
Items were constantly being evaluated and revised during the two phases of the study: exploratory and verification.

Backing1.1:
Model testing using confirmatory factor analysis revealed a two-factor model of rater attitude construct.

Backing1.2:
Item construction was greatly informed by two empirical studies.

Backing 1.3:
Item revision is based on statistical analyses and qualitative feedback from raters and content experts.

Figure 7. Forms of inference for study 1: validation and construction of the RAI.

To assess model adequacy, several indices common to social science research and provided by the AMOS software were used. All indices were evaluated together to determine the adequacy of hypothesized models. The following cutoff values are recommended by Hu and Bentler (1999):

χ^2: This is an absolute fit index which indicates the degree of fit between the proposed model and the observed data. The smaller the χ^2 the better the model fit. Chi-square is generally not used as a sole index of model fit in practice due to its sensitivity to sample size.

Comparative fit index (CFI): Proposed by Bentler (1990), CFI is used to avoid underestimation of fit caused by small samples. CFI ranges between 0 and 1 and values at or above 0.95 indicate a good fit.

Root Mean Square Error of Approximation (RMSEA): This is related to residuals in the model. It is a measure of fit introduced by Steiger and Lind (1980) and is relatively insensitive to sample size. RMSEA values close to .06 or below are considered acceptable.

Tucker-Lewis Index (TLI): Proposed by Tucker and Lewis (1973), this compares the fit of the proposed model to that of a null model. It can be treated in a similar fashion as CFI, but is less sensitive to sample size. Value of 0.95 is the cutoff for a good model fit.

The second warrant to support the claim was to demonstrate that items were designed within the language assessment context, as compared to those in the general context (e.g., Zahn & Hopper, 1985), to reflect distinctive attitude attributes and statements as revealed by raters in the two empirical studies (i.e. interview and speaker evaluation study) (*Backing 1.2*). The interview data was analyzed by the portraiture approach developed by Witz, Hart, & Thomas. (2001). This approach highlights the importance of going beyond general and observable characteristics of the participants and attempts to uncover their subjective universe and outlook. According to Witz (2006),

> "Portraits… give all kinds of impressions and hints of subtler and deeper aspects, such as the developments of these participants' self. . . their aim is not to present a fuller understanding of the person as a whole, but only enough of such an understanding so that one can see how [issues of interest] is part of the person as a whole and part of her life" (p. 3).

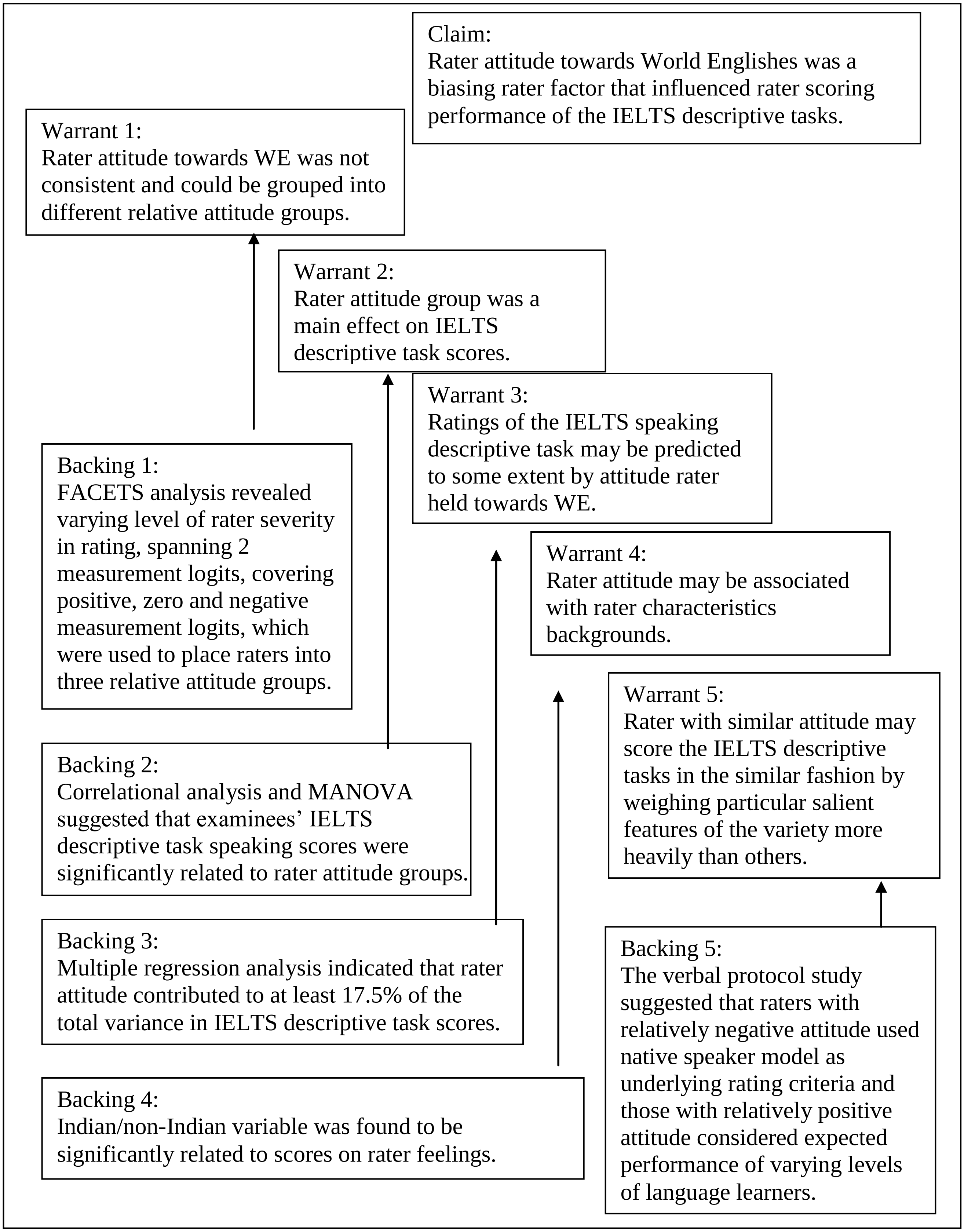
The ultimate objective of using portraiture analysis is to discover raters' inner selves and consciousness through empathizing with their feelings and complexities to determine their motivations at oral proficiency test ratings. The findings of the portraits created in the study helped construct item statements that greatly reflected rater concerns and attitudes towards WE in relation to oral proficiency assessments. Lastly, items that withstood from the rigorous evaluation of appropriateness and quality were retained (*Warrant 1. 3*). This was evidenced by item analysis, including item-total statistics, exploratory and confirmatory factor analysis and qualitative feedback from raters (*Backing1. 3*).

Study 2 Procedures

To investigate the extent to which the attitudes towards WE was accounted for by their scoring performance, raters provided online ratings on the RAI and IELTS descriptive tasks. Results were analyzed using appropriate statistical analysis methods. Based on the findings of two rating tasks, eight raters were contacted for a verbal protocol study that looked into salient features of the variety that affected scoring tendencies of raters with similar attitudes towards WE.

Study 2 Forms of Inference

Figure 8 presents the form of inference for the second study. The claim for study 2 was that rater attitude towards WE was a biasing factor that influenced their scoring performance on



Claim:
Rater attitude towards World Englishes was a biasing rater factor that influenced rater scoring performance of the IELTS descriptive tasks.

Warrant 1:
Rater attitude towards WE was not consistent and could be grouped into different relative attitude groups.

Warrant 2:
Rater attitude group was a main effect on IELTS descriptive task scores.

Warrant 3:
Ratings of the IELTS speaking descriptive task may be predicted to some extent by attitude rater held towards WE.

Warrant 4:
Rater attitude may be associated with rater characteristics backgrounds.

Warrant 5:
Rater with similar attitude may score the IELTS descriptive tasks in the similar fashion by weighing particular salient features of the variety more heavily than others.

Backing 1:
FACETS analysis revealed varying level of rater severity in rating, spanning 2 measurement logits, covering positive, zero and negative measurement logits, which were used to place raters into three relative attitude groups.

Backing 2:
Correlational analysis and MANOVA suggested that examinees' IELTS descriptive task speaking scores were significantly related to rater attitude groups.

Backing 3:
Multiple regression analysis indicated that rater attitude contributed to at least 17.5% of the total variance in IELTS descriptive task scores.

Backing 4:
Indian/non-Indian variable was found to be significantly related to scores on rater feelings.

Backing 5:
The verbal protocol study suggested that raters with relatively negative attitude used native speaker model as underlying rating criteria and those with relatively positive attitude considered expected performance of varying levels of language learners.

Figure 8. Forms of inference for study 2: rater attitude and scoring tendency

the IELTS descriptive tasks. This was first examined by seeking the extent to which rater attitude was similar on the level of severity (*Warrant 2. 1*). FACETS analysis (Linacre 1989) was performed to examine rater severity levels and the difficulty level of each RAI component (*Backing 2.1*). Rater scoring performance on the IELTS descriptive tasks was inspected to seek any association with rater attitude towards WE (*Warrant 2.2*). A one-factor multivariate analysis of variance (MANOVA) was performed to explore how the variability in the analytic ratings (i.e. on Fluency, Pronunciation, Vocabulary, Sentence Structure) of the IELTS descriptive tasks can be explained by the effects of rater attitude group (*Backing 2.2*). Then, the hypothesis that rater scoring performance on IELTS descriptive tasks may be predicted by the RAI scores and rater characteristic backgrounds was tested (*Warrant 2.3*). Correlational analysis was used to determine the direction and magnitude of the two criteria (i.e. IELTS descriptive task scores) and predictor variables, that is, RAI scores (*Backing 2.31*). To examine how much of the variance of IELTS descriptive task ratings, either total or sub score(i.e. Fluency, Pronunciation, Sentence Structure and Vocabulary), was accounted for by the attitude total, part scores and rater background variables, regression analyses using stepwise methods were performed (*Backing 2.32*). Next, rater attitude is hypothesized to be predicted to some extent by some of the rater characteristic backgrounds *(Warrant 2.4*). Correlational analysis and regression analysis were used in support of this warrant *(Backing 2.4*). Finally, a contention that raters with similar attitudes may score the IELTS descriptive tasks in a similar fashion by weighing particular salient features of a variety more heavily than others was tested (*Warrant 2.5*). A verbal protocol study was then conducted with eight selected raters of varying perceptions to WE as revealed in the responses to the RAI and varying levels of severity in rating IELTS descriptive tasks (*Backing 2.5*).

MANOVA was used in testing hypothesis 2.2 instead of multiple ANOVAs (analysis of variance) given its advantages over ANOVA (Tabachnick & Fidell, 1996). First, due to its test on multiple dependent variables (i.e. four analytic ratings) simultaneously, the effects of independent variables (i.e. attitude group difference) were evaluated at the same time. This decreases the risk of type I errors (a null hypothesis is rejected when it is true) as may be the case when conducting multiple ANOVAs independently. Furthermore, MANOVA can be more powerful than individual ANOVAs given that all dependent variables are considered together and group differences are maximized.

Participants

Raters_Phase 1

In Phase 1, the purpose was to explore rater views of WE and their potential effects on ratings. Three volunteer raters of American English who had rated the Berlitz Proficiency Interview (BPI) for at least half a year when the study was conducted were recruited from Berlitz Inc. The BPI is a phone-based speaking test developed between Berlitz Inc. and the UIUC Foreign Language Assessment Group of which the researcher was a member. Two of the raters were interviewed twice and the third only once as he had fixed ideas with regard to WE and associated issues arising from WE with raters' incapability to make scoring judgments. Each interview was conducted over the phone and lasted approximately 40 minutes.

Raters_Phase 2

The RAI was pilot tested in Phase 2. Twenty raters of oral proficiency tests administered by one of the three universities/organizations, that is, the University of Illinois at Urbana-Champaign (UIUC), Purdue University and the Michigan English Language Assessment Battery (MELAB), participated in this phase of the study. All twenty raters had more than six months of rating experience. Ten of them were from the UIUC, six from Purdue University, and the other four were MELAB raters. Their demographic characteristics are listed in Table 2.

Table 2

Demographic Profile of Raters

Variable	N	Percent %
Gender		
Female	15	75
Male	5	25
Nationality		
American	11	55
Non-American	9	45
Native language		
English	13	65
Non-English	7	35
Year of rating experience		
More than half an year	5	25
1-3 years	5	25
4-6 years	2	10
More than 6 years	8	40
Major of highest degree		
TESOL	9	45
English	2	20
Others	9	45
Affiliated institution		
UIUC	10	50
Purdue University	6	30
MELAB	4	20

Raters_Phase 3

IELTS raters in the U.S. and India along with ESL/EFL teachers with teaching experience of at least half a year in the U.S. and India respectively at the time of the study were recruited for the main study in Phase 3. Due to difficulty in reaching the target number of 100 IELTS raters, the decision was taken to include ESL/EFL teachers as these two groups have similar backgrounds. According to the IELTS website, all IELTS raters must possess relevant

TESOL qualifications and at least three years of ESL/EFL teaching experience. The recruitment of IELTS raters and ESL/EFL teachers was carried out concurrently. The former were contacted mainly through the IELTS world-wide rater manager. As for ESL/EFL teachers, approximately 150 invitation emails were sent to members of TESOL organizations and directors of the ESL programs offered privately or affiliated with the universities in New York City, San Francisco, and India. The selection of New York City and San Francisco was to facilitate access to the teachers in the follow-up qualitative study. The invitation email was approved by the Institutional Review Board at the University of Illinois at Urbana-Champaign and included a brief introduction to the study and the remuneration (See Appendix A). The director or coordinator then forwarded the email to eligible ESL/EFL instructors to inform them of the need for their participation in this study. Teachers who were interested in participating responded via email. The researcher then sent them the instructions and link to the study. The total sample yielded 96 ESL/EFL teacher participants, among which were 23 IELTS raters. Of these, 13 were Indian and 83 were non-Indian, with Americans predominating. The demographic distribution of non-Indian raters was 68 American, 4 Chinese, 2 Korean, 2 Japanese, and one each of Brazilian, Russian, Greek, Malaysian, Filipino, Pakistani, and Nigerian nationality. In terms of highest educational qualification attained, the majority (75%) possessed a Master’s degree in Teaching English to Speakers of Other Languages, including Linguistics. Table 3 presents the demographics of participants for this phase.

Judges

Judges participated in phase 1 of the RAI construction for the purpose of elicitation of feeling attributes in an attempt to facilitate the construction process. Forty undergraduate students in EDPSY 220 at the UIUC campus were recruited to complete this task. All completed consent forms and background questionnaires in accordance with ethics procedures. The students first language (L1) distribution was as follows: 34 spoke English as their first language, 2 spoke Korean, and 1 each spoke Polish, Russian, Mandarin Chinese, and Indian. This task was named "Varietal Speaker Evaluation". It involved a rigorous process of evaluating speakers and their

Table 3

Sample Demographics (N=96)

Sample characteristics	N	%
Country of current residency		
US	78	81
India	12	13
Others	6	6
Nationality		
Non-Indian	83	83
Indian	13	13
Gender		
Female	73	76
Male	22	22
Missing data	1	1
Native language		
English	73	76
Others	22	23
Missing data	1	1
Year of teaching experience		
Less than 1 year	6	6
1-3 years	18	19
4-6 years	19	20
More than 6 years	52	54
Highest level of education		
Bachelor's	14	15
Master's	72	75
Doctoral	8	8
Missing data	2	2
Major of highest degree		
TESOL (including Linguistics)	72	75
Education	9	9
Others	10	10
Missing data	5	5

voice of four outer-circle varieties (i.e., India, Pakistan, Sri Lanka, and Singapore) and four expanding circle varieties (i.e., Taiwan, Vietnam, Turkey, and Korea).

Raters_Verbal Protocol Study

Informed by the analysis of quantitative data, eight raters from different combinations of tendency of WE attitude and rating severity on the IELTS descriptive tasks were selected. Additional details regarding the limitations of rater selection and alternatives to compensate for the limitations appear in chapter 5 on Rater Attitude and Rating Behavior. Table 4 reports raters' background information where of the eight rater interviewees, one is Indian and the rest American. The attempt to balance raters from various nationalities was not achieved due to the limited number of Indian raters who met the selection criteria above. Gender distribution displayed a balanced representation with each gender accounting for half of the interviewees. In

Table 4

Information on Rater Interviewees

Rater Code	Nationality	Residency	Gender	Years of ESL/EFL Teaching Experience	Years of rating experience on any commercial oral test
01	British	U.S.	Male	11	0
04	American	U.S.	Female	22	0
23	Brazilian	U.S.	Male	3	0
27	American	U.S.	Female	6	0
48	American	U.S.	Male	19	8
53	American	U.S.	Male	16	5
54	American	U.S.	Female	12	0
77	Indian	India	Female	20	6

terms of the length of teaching experience, the majority of them had taught ESL/EFL for more than 10 years. As for experience in rating commercial oral proficiency tests, only three had rated the IELTS at the time the study was conducted.

Content experts

Each phase in the development of the RAI was reviewed by content experts to ensure its content validity (Grant & Davis, 1997) and in strengthening the inference argument as supported by multiple evidences. The Standards for Educational and Psychological Testing (American Educational Research Association, 1985) recommends three criteria for content experts involved in the content review process, namely experience, qualifications, and relevant training. With this in mind, four specialist PhD content reviewers, two each in second language assessment and sociolinguistics worked independently with the researcher to examine the representativeness, comprehensiveness, and clarity of the RAI. Representativeness refers to the degree to which each item reflects the issues of second language assessment in relation to WE; comprehensiveness of the entire RAI was to identify items which they perceived to be congruent with perspectives that conceptualized the attitude constructs and finally, each reviewer evaluated the clarity of items and wording to ensure no poorly written items.

Speech Samples

Descriptive tasks extracted from the IELTS speaking section, part 2, were used in the main study. The selection criterion of speech samples needed to conform to the following criteria: Indian examinees and those scores representing a range of IELTS score bands (i.e., bands 4, 5, 6, 7, 8, and 9). The reason for selecting the Indian variety is that it has been a major research agenda in the WE research (see chapter 2) and the use of only one variety is to control the research variables. The researchers at the IELTS validation group in the U.K. helped select six samples that met the above requirements, copied the samples to a CD and sent it to the dissertation researcher. Each sample was edited on Audacity software to remove the first and third part of the speaking test, which is a short monologue and two-way dialogue respectively. As interlocutors have been shown to influence examinee performance and scores (Brown, 1995), only the second part of the IELTS speaking test that required examinees to provide description on particular topics was used for the study. This part of the test lasted 90 seconds.

Mixed Methods Design and Analysis

Rationale for the Mixed Methods Study

This study adopts the operational definition by Tashakkori and Teddlie (1998) of mixed methods (MM) research that comprises a combination of qualitative and quantitative approaches into the research methodology of a single or multi-phased study (pp.17-18).

In line with the Social Process Model of Cargile *et al*. (1994) which depicts the process of language attitude formation as comprising multidimensional components rather than being unidimensional, the first study that developed and validated the RAI sought to capture the complexity of constructs of rater attitude towards WE by employing multiple methods with results from one method helping to develop or plan the next method. This entailed the sequential use of the following study procedures: in-depth one-on-one interviews, construction of an item pool, descriptive statistics, and factor analysis. The purpose of the MM design for the first study is *development,* one of the five purposes for MM studies as advanced by Greene, Caracelli and Graham *(*1989). For purposes of *expansion*, the second validation study that explored the relationship between raters' WE attitude and rating tendency employed different statistical methods to test each of the hypotheses and captured linguistic and non-linguistic features that influenced raters' scoring decisions in relation to the attitude they held towards WE.

This involved the use of quantitative analysis of rater attitude scores, the IELTS descriptive task scores and results of the qualitative verbal protocol study (see chapter 5).

Dimension of Differences in Mixed Methods Design

According to Caracelli and Greene (1997) and fuller descriptions in Greene (2007, pp.22-23), the salient and critical dimensions of MM design form either *component* or *integrated* designs. Design is determined by implementing methods independently or interactively, weighting equally or unequally and sequencing concurrently or sequentially. Component designs are commonly found in practice with methods implemented independently and mixing during data interpretation and conclusion, whereas the more sophisticated integrated designs intentionally mix paradigms and methods at different stages of the study. Greene (2007) provides a typology that includes four integrated design types: *iteration, blending, nesting or embedding, and mixing for reasons of substance or values* (p.125). Iteration designs have the methods implemented sequentially with varying degrees of weight given to each method; blending designs may implement methods concurrently to explore the different facets of the same phenomenon; nesting/embedding involves the integration of a supplementary method into a set of primary methods. A salient feature of this approach is " the secondary method follows or adheres to key parameters of the primary method-for example, sampling or designed controls-rather than following the parameters usually associated with this secondary method" (p.127). An example of such a design in language assessment research was found in the study by Xi (2005) that used quantitative-dominated methods to look into effects of visual chunks and planning on speaking performance on the graph description task. Xi's study quantified the entire qualitative data that served as a secondary method and analyzed the qualitative data applying quantitative analysis techniques. The last type is mixing for reasons of substance or values, commonly labeled “transformative mixed methods design” (Greene, 2007, p.129), and mixing for the purpose of ideological concerns.

Table 5

Interactive Mixed Methods Design for Study 1: Construction of Rater Attitude Instrument

Stage	Study focus	Mixed Methods Purpose	Methods	Weight of Methods	Sequence of Implementation	Type of Integration
1	Explore constructs of rater WE attitude	Development	Interview	Unequal: qual+ Quan	Sequential	Iteration
2	Construct item pools		Semantic differential scale & Likert scale			
3	Revise the Instrument					
4	Implement the instrument					
5	Verify the constructs of rater WE attitude		Quantitative analysis			

The different dimensions in MM design for the two main studies in this dissertation are presented in Table 5 and 6 .Table 5 presents the MM design for study 1 that was dedicated to the construction of the RAI. The method in the first stage (i.e., interviews) informed the development of the second method (i.e. semantic differential scale and Likert scale) to measure constructs of rater attitude towards WE as emerged in the first method. Therefore, the sequence of method implementation and type of method integration are sequential and iteration respectively. In terms of weight given to each method, the exploration conducted by the one-on-one interviews began the process of instrument construction; nevertheless, the following stages of instrument construction relied heavily on the psychometric analysis of the instrument, which led to an unequal status of methods, with quantitative methods forming the majority of the weightage.

Table 6

Interactive Mixed Methods Design for Study 2: Relationship between Rater Attitude towards World Englishes and Rating Tendency

Stage	Study focus	Mixed Methods Purpose	Methods	Weight of Method	Sequence of Implementation	Type of Integration
1	Seek variations in rater WE attitude	Expansion	Rater Attitude Instrument	Equal: Quan+Qual	Concurrent	Iteration & Embedding
	Explore relationship between rater attitude and demographic background info		Rater Attitude Instrument			
	Search relationship between rater attitude and rating tendency of IELTS speech sample		IELTS scores & Rater Attitude Instrument			
2	Identify linguistic and non-linguistic features influential to rater scoring decision in relation to rater attitude towards WE		Verbal protocol study		Sequential	

Study 2 attempted to use different methods to learn about the different phenomena brought about by the association derived from rater attitude towards WE to rater scoring tendency within the same study. As illustrated in Table 6, all methods in stage 1 utilized instrument scores and ratings given to the IELTS descriptive tasks to concurrently assess the extent to which raters' scoring decisions can be accounted for by their attitudes to WE. Based on the results in stage 1, selected raters were chosen for a verbal protocol study in an attempt to identify linguistic and non-linguistic features of Indian English that were influential in their scoring decisions. Therefore, the integration approach was iteration while embedding in the sense that the verbal protocol study was a nesting of a secondary method, as part of study's primary methodology. In terms of sequence of method implementation, while it was concurrent within stage 1, it was sequential within the entire study 2 to strengthen the linkage between stage 1 and 2 together.

Data Analysis in Mixed Methods

The mixing of data for component MM design mainly occurs during data interpretation and inferencing whereas the highlight of mixing for interactive design is through the analysis stage (Greene, 2007). It is through this stage of the mixing that the difference among the data set and conflicting results may emerge to lead to further critical thinking of issues being investigated. Greene claims:

> Convergence, consistency, and corroboration are overrated in social inquiry. The interactive mixed methods analyst looks just as keenly for *instances of divergence and dissonance*, as these may represent important nodes for further and highly generative analytic work (p.144, emphasis added).

Following this, this dissertation expects both convergent and divergent findings of rater attitude towards WE and generates unanticipated insights and understandings of its effects on rating performance. The approach of data analysis for two studies is summarized in Figure 9 and Figure 10. For study 1, the substance of the instrument was predominantly informed by interview data and results of the attitude elicitation session (see chapter 4). Next, keyword

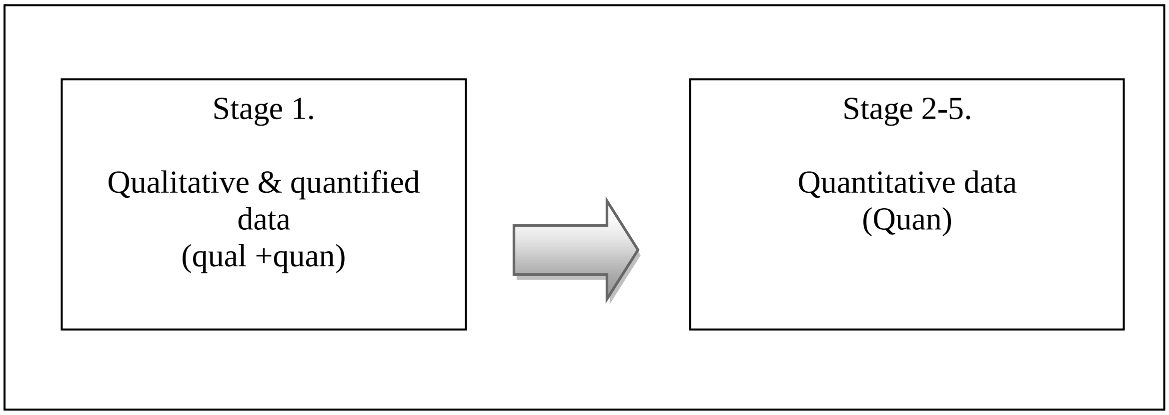


Figure 9. Data analysis for study 1.

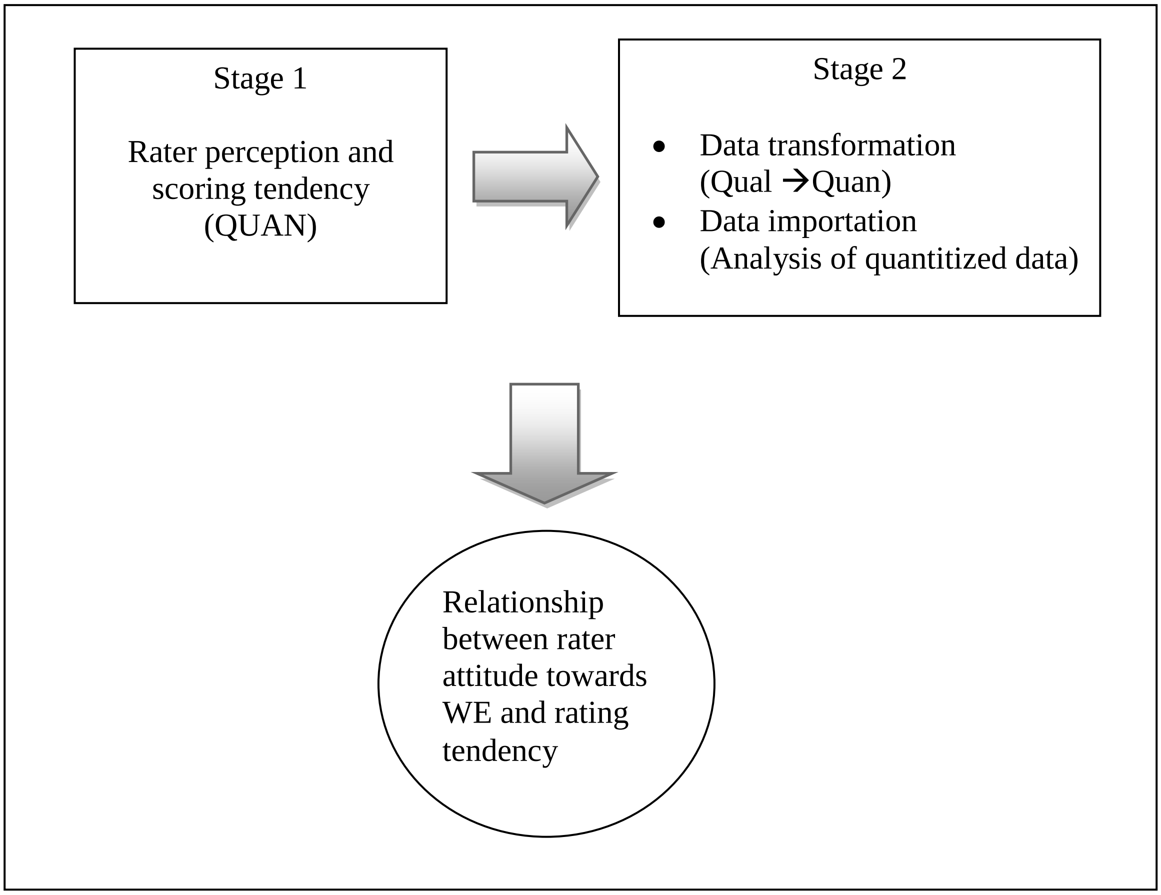

Stage 1
Rater perception and scoring tendency
(QUAN)
Stage 2
• Data transformation
(Qual → Quan)
• Data importation
(Analysis of quantitized data)
Relationship between rater attitude towards WE and rating tendency

Figure 10. Strategies for data analysis in study 2

analysis was applied to the attitude elicitation session and results were quantified. The results of the two analyses significantly contributed to the development of the instrument in the following stages.

For study 2, each procedure was accompanied by a strategic label in the parenthesis as used in Greene (2007). More strategies of MM design were applied in study 2 compared to study 1. As displayed in Figure 10, stage 1 mainly involved the investigation of multiple dimensions of rater attitude towards WE and its effect on scoring tendency by means of quantitative analysis. Informed by the findings from stage 1, stage 2 was about the verbal protocol study that generated qualitative data and was transformed by quantitizing the verbal protocol reports focusing on linguistic (e.g., syntax) and non-linguistic features, such as the level of English education) (*Data transformation*). Next, all the features formed a new quantitative data set and was further analyzed to seek commonality and expectations *(Data importation)*. Then the results were compared to rater interviewees' respective rating performance using FACETS analysis (see chapter 5), as informed by findings in stage 1, in the form of matrix in an attempt to visually represent an attitude-rating relationship.

CHAPTER 4

CONSTRUCTION AND VALIDATION OF RATER ATTITUDE INSTRUMENT

Considerations of Instrument Development

The Rater Attitude Instrument (RAI) is intended to measure attitude rater hold towards WE. The items attempt to measure tripartite constructs of the attitude: belief, feeling and rating tendency, rather than experience. According to the widely cited framework of scale development proposed by DeVellis (2003), eight comprehensive procedures were recommended for advancing the validity of scale within social science inquiry:

1. Determine clearly what it is you want to measure
2. Generate an item pool
3. Determine the format for measurement
4. Have the initial item pool reviewed by content experts
5. Consider inclusion of validation items
6. Administer items to a development sample
7. Evaluate the items
8. Optimize scale length

The foremost step for scale construction is to clearly determine the attitude object, which should be specific to the behavior with reference to the target, action, and context. In this dissertation, WE as perceived by rater (i.e. target) with ESL/EFL teaching experience for at least half an year (i.e. action) as related to oral proficiency assessment (i.e. context) is the primary focus rather than WE as perceived by others (e.g. students and teachers) outside the language assessment context, such as in the ESL/EFL classroom setting. Next, DeVellis (2003) suggests creating an item pool before determining the scaling methods for measurement.

Nevertheless, it may be problematic to write items without knowing the format to be used as items presented in different scaling methods should be written in a way to reflect the distinctive function and characteristics of particular scaling method. Thus, the commonly discussed and used scale formats for measuring language attitude were reviewed. They include four types, each with different item designs: Likert scales, semantic differential scales, Guttman scales, and Thurstone scales.

Mueller (1986) summarized the functions of each format. The Likert scales are most commonly used scaling techniques in psychology and social science. This scale type is a summative tool, which is composed of a set of items measuring constructs of interest and sums all item scores to typically obtain a single score. Declarative statements should be strongly worded to elicit more variations in the responses. Typically Likert scale is assessed on a 5-point item response format.

The semantic differential scales use bipolar adjectives or adjective phrases as endpoints on a 7- or 9-point continuum between these two adjectives. The strength of the semantic differential method is that it is short, relatively easy to administer, and highly reliable as shown in some test-retest reliabilities having internal-consistency coefficients of around .90 (Schibeci, 1982, as cited in Mueller, 1986). Furthermore, the scores from the semantic differential scales typically correlate very highly with those from the Likert and Thurstone attitude scales (Mueller, 1986). Figure 11 presents the example of the semantic differential scale.

	1	2	3	4	5	6	7	
Rich								Poor
Unfriendly								Friendly

Figure 11. Example of sematic differential scale

Hierarchical number	Item number	Statement
1	5	Occasionally engaged in work (1 or more days out of 10)
2	3	Sporadically engaged in work (3 or more days out of 10)
3	6	Sometimes engaged in work (5 or more days out of 10)
4	1	Usually engaged in work (7 or more days out of 10)
5	4	Regularly engaged in work (9 or more days out of 10)
6	2	Consistently engaged in work (every day)

			Statement			
Respondent	5	3	6	1	4	2
E	Yes	Yes	Yes	Yes	Yes	Yes
B	Yes	Yes	Yes	Yes	Yes	No
A	Yes	Yes	Yes	Yes	No	No
F	Yes	Yes	Yes	No	No	No
C	Yes	Yes	No	No	No	No
D	Yes	No	No	No	No	No

Figure 12. Example of an abbreviated Guttman scale and response matrix Adapted from Christ & Boice (2009).

A Guttman scale includes items strongly linked to a single factor. Items are arranged in a hierarchical order so that, for example, agreeing to item 6 implies endorsement to item 1-5. Thus, responses to Guttman scale look similar to a matrix, as presented in Figure 12. The benefit of a Guttman scale is economical in the sense that only a subset of items are administered and responses to other items can be inferred from previously established response patterns, which makes the Guttman scale cumulative. The disadvantage of this type is that the scale development takes time, requires more piloting, and applies arbitrary standards to determine the relative relationship between items (Christ & Boice, 2009).

A Thurstone scale is consisted of a series of statements. Unlike Likert scales, a Thurstone scale gives each statement a value or weight determined by item developers during item construction. The statements chosen for study are evenly spread in intensity from least favorable to most favorable. This scale type is referred to as *an equal-appearing intervals scale*

(Aiken, 1996). This type of scale is easy for respondents to answer but the assigned item values may vary if a different group of item developers were hired. An example of Thurstone scales is shown in Figure 13. Note that on an actual scale items are not arranged in order of value and the values are not listed.

Below you will find a number of statements expressing different descriptions of the target student's behavior.

Put a check mark if you agree with the statement for the target student.
Put a cross if you disagree with the statement for the target student.

Try to indicate either agreement or disagreement for each statement. If you simply cannot decide about a statement you may mark it with a question mark.

Scale value	Item number		
(2.5)	10	__________	Student frequently talks with peers during instruction.
(5.4)	7	__________	Student giggles and talks with peers occasionally during instruction.
(6.3)	5	__________	When placed in a small discussion group, student talks the majority of the time about the topic.

Figure 13. Example of an abridged Thurstone scale for student academic engagement

Adapted from Peterson, R.C., & Thurstone, L. (1933). Motion Pictures and the Social Attitude of Children. New York: Macmillan; as appears in Aiken, 1996.

Given the difference in item design between scaling methods, the next step after attitude object to be measured was determined was to select appropriate methods to elicit rater attitude, which was greatly informed by findings of the preliminary pilot study with raters of a commercial oral proficiency assessment and extensive literature reviews on language attitude study and language assessment (see chapter 2). Techniques used in language attitude study generally provide a stimulus to arouse listener feelings. Thus, to capture the immediate feeling on speaker accompanying by his/er voice, the semantic differential scale seems to be a good choice. On the other hand, many insightful opinions on WE were elicited during the preliminary pilot study, suggesting that rich statements are needed to best reflect raters' beliefs and insights in WE and its effects on rating performance, as opposed to adjective or adjective phrases used in the semantic differential scales. Therefore, in order to capture raters' unpolished feeling upon hearing a voice, deeper beliefs in WE and potential rating tendency, it was decided to use both semantic differential scale and Likert scale to best serve the needs of each elicitation purpose. After the format for measurement was determined, DeVellis's (2003) framework on scale development guided the following procedures for scale construction and will be presented in the sections below, including item review by content experts, selection of item inclusion, administration of items to the target group, evaluations of items and finally optimization of scale length.

Content Validity: Construction Phase 1

Evidence-Driven Instrument Design

The purpose of the construction phase 1 was to elicit attitude attributes specific to language assessment context. It was decided not to generate items exclusively by the researcher given that language attitude scales have not been previously developed within language assessment inquiry, such as the speaker evaluation instrument (Zahn & Hopper, 1985) used in the recent attitude studies in language assessment research, including Harding (2009) and Kang (2008). It is unknown whether the existing scales contain items pertinent to the language assessment inquiry or not. The development of the RAI began with two preliminary studies to explore rater inner voice and views on WE: first, an in-depth interview with raters of oral proficiency assessment from a global language test provider, Berlitz Inc, and secondly, a study that elicited immediate feelings and emotions toward varieties of English in EDPSY220 at the

University of Illinois at Urbana-Champaign. The findings of the two studies contributed to a development of a set of 60 Likert scale items and 25 semantic differential scale items.

Preliminary Study 1: In-Depth Interviews

An in-depth interview was conducted in Spring 2008 with three raters of a phone-based oral proficiency assessment provided by Berlitz Inc. The interview was analyzed by a portraiture approach developed by Witz *et al* (2001, see chapter 3). Table 7 presents the summary of each portrait. All rater names are pseudonyms.

Portrait of Luke

- *Bringing multicultural experience in Miami to facilitate in rating*

Living in Miami, Luke is a Spanish and English bilingual who is exposed to the mixed language form of Spanglish all the time. Although it may be confusing to other Americans, this has never been a problem for him. Luke's comfortable coexistence with the world of non-standard English was shaken somewhat when he embarked on his career as a rater. He was aware that his familiarity with one form of varietal English could inadvertently or unconsciously cause him to overlook or ignore certain factors when judging examinees' oral test performance. As he sat back and mulled on whether his Miami experiences would complement or obstruct his efforts to treat examinees as objectively as he should, Luke came to realize that there was something missing in his rating capability.

- *Realizing rating differs from his daily contact with people in Miami*

Luke considers the ability to communicate oneself was more important than adhering strictly to rules of grammar and language structure or following a certain style of spoken English. He was thus surprised to find out that this wasn't completely true or good enough in achieving fair and objective ratings. The rater training was a tremendous wake-up call

Table 7

Summary of the Three Portraits

Luke	Kyle	Nash
• **Bringing multicultural experience in Miami to facilitate in rating** “ . . . You hear a lot, especially Spanglish, it’s kind of common place. . . It gives me a certain amount of comfort . . . it not bother me”.	• **Treating language as an instrument to facilitate people engagement** “. . . Studying history and the different forms of language have always been interesting and fascinating topics for me!”	• **Being proud of his job performance as a rater** “. . . They told me I was doing a great job. . . I don’t wanna change . . .”
• **Realizing rating differs from his daily contact with people in Miami** “ . . . In terms of the call center, you think of people in India, the Philippines, and these are great jobs for these people. . . We were told [during the rater training] to have a more objective approach and not to let the subjectivity affects us”.	• **Being knowledgeable that language is for engagement in the world context** o Knowledgeable about English language development o Knowledgeable about Chinese & German language evolvement	• **Treating American English as the only standard** “ . . . There is no room area for skeptical like I was thinking that is not making sense to me, so therefore I wouldn’t grade them well”. o Issues of varieties of English equals to raters’ lack of rating experience.
• **Aiming to be objective and looking for patterns to determine unfamiliar phrases as part of a variety or incomplete linguistic forms** **“. . .** if I ask Indian speakers, *‘how do you like this?’,* he said, *‘it’s too good’*. Then I asked them to describe something that I know they were finding good as well and see if they use the same pattern.”	• **Rating is a fulfillment of people contact** o Fluency is considered being able to communicate o Varieties of English is not that vital but the ability to get meaning across	

for him. “I think I become tougher”. Luke had been greatly influenced by the rater training and the change was noticeably apparent. His ability to provide fair and objective evaluations of tests had been significantly improved; his ratings were no longer based on an ordinary Miami speaker’s perspective but that of a rater who critically and carefully evaluated his examinees ability to handle workplace communication proficiently.

- *Aiming to be objective and looking for patterns to determine unfamiliar phrases as part of a variety or incomplete linguistic forms*

Whenever Luke heard unfamiliar phrases or structures spoken by the examinees, Luke did not immediately consider or judge them as errors deriving from partial second language acquisition process. Instead, he manipulated the interview questions to see if the same patterns would be repeated in the same context. If the patterns continued, meaning that they were not spoken randomly, Luke would decide that they were not mistakes but rather a steady, systematic, and regular speech form representing a part of the examinees’ variety. Such pattern-searching greatly assisted Luke in overcoming any doubts he had about whether examinee were making potential speech errors or whether the responses were a genuine and legitimate component of a variety. Figure 14 presents Luke’s orientation to oral proficiency assessment.

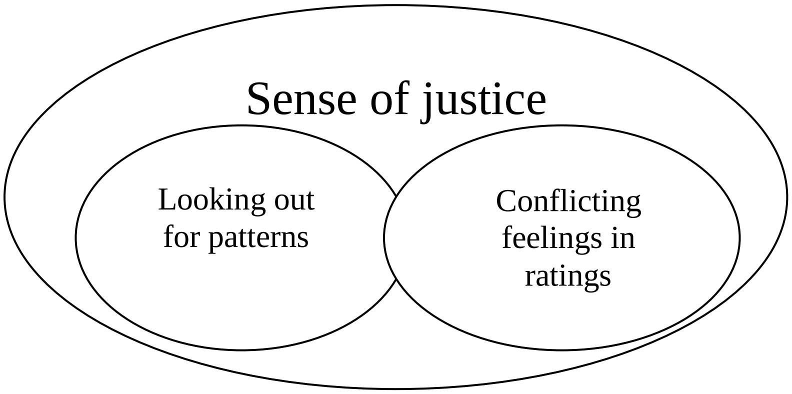


Figure 14. Luke’s state of mind as a rater

Portrait of Kyle

- *Language as an instrument to facilitate people engagement*

Anyone talking with Kyle would notice his exuberance for learning about different people and languages, widening his social contacts, and interest in traveling all of which are aimed at understanding the real lives of the people he meets. He particularly looks forward to meeting people who speak different types or versions of English because this gives him the opportunity to experience and handle diversity. He believed that language as a medium of communication achieves its role best when it conveys the message being expressed. It is not necessary that everything said should be perfect in terms of grammar or structure; what is important is that it was effective. So phrases and tenses can be changed or modified to suit the circumstance. In this way Kyle was able to interact better with people no matter what their language proficiency skills were and learn more about them, their culture and their history.

- *Language for engagement in the world context*

Beyond Kyle's love for language, his interest in interacting with people and traveling enabled him to appreciate the evolution of the English languages as it spread around the globe. With an academic background in European history and linguistics, Kyle is knowledgeable about the global spread of English and the legitimate status of World Englishes. He believes the language has always been dynamic enough to accommodate adaptation by users to facilitate social intercourse, and economic and cultural interactions. He indicated that although local versions of languages may generally be in existence they cannot be considered new or different languages. Instead, they function or exist under a central system and, despite having local innovations or characteristics, still maintain most of the unified structures or features of the original language.

- *Rating is a fulfillment of people contact*

Kyle's appreciation of the importance and beauty of languages had a significant impact on the way he approached his task as a rater. While rating, he allows for enough latitude and scope for variations to enable examinees to demonstrate the range of their communication aptitude as well as the extent to which they are able to utilize language to get across to people. For example, Kyle understood "I'm a fresher" is part of the Indian English repertoire and accepted its usage because it was a common expression used by most of the college graduates he had interviewed. Kyle's mental paradigm is shown in Figure 15.

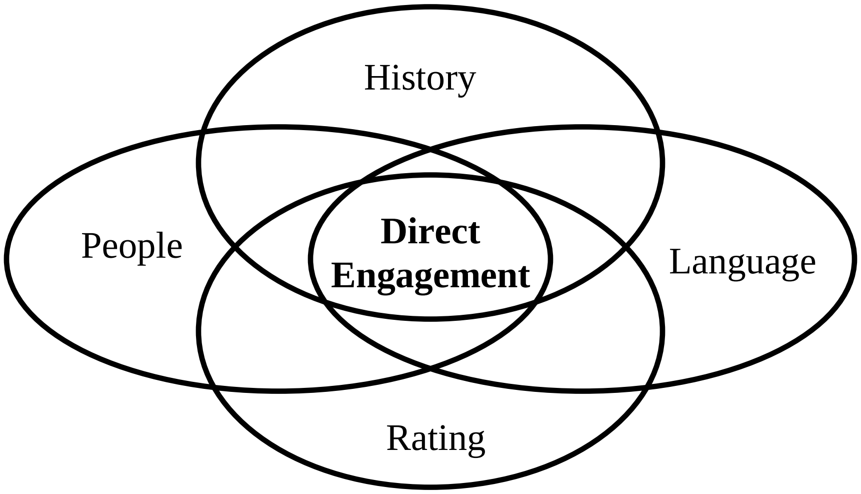


Figure 15. Kyle's mental paradigm

Portrait of Nash

- *Being proud of his job performance as a rater*

Nash's performance as a rater was well received by the companies that required prospective employees to take the Berlitz test (i.e. BPI) and he was soon considered a benchmark for other raters. With a Master's degree in Psychology, Nash believed his interviewing style and approach developed over the years as a student doing psychological research was adequate and was understandably pleased with the recognition he received for the quality of his job performance. He felt no reason to modify or alter a rating style which he had used all along.

- *Treating American English as the only standard*

For Nash, measuring language ability is a very clear-cut process since it is based on very distinct and specific criteria, namely, grammar, fluency, linguistic range, and phonological control; and for raters there shouldn't be any ambiguity in that because they were apprised of these criteria from the outset. It was obvious that Nash had very clear and fixed ideas about going about his rating tasks. For example, mistakes in pronunciation may derive from incorrect stresses or mispronounced phonemes, compared to the way they were pronounced in American English, or the different way of pronouncing the words due to the influence of varietal English. Since what was being measured was very clear, he felt that the issue of English variability should not even arise and such varietal English was not acceptable. Figure 16 shows Nash's approach to rating.

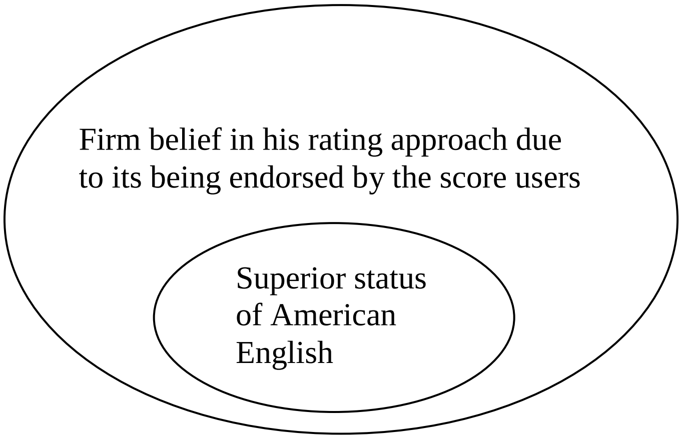

Firm belief in his rating approach due to its being endorsed by the score users
Superior status of American English

Figure 16. Nash's Approach to Rating

Cross-Case Findings

The cross-case analysis showed that rater attitude towards WE was greatly shaped by many factors, including educational background, hometown environment, personal hobbies, job achievement and exposure to different varieties. Luke's and Kyle's attitudes toward the variety of English were not only liberal but they also recognized the variations as linguistically correct and legitimate. Both raters exhibited a positive tone toward English variations and were convinced that successful cross-cultural or regional communications could be achieved without stringent adherence to standard American English usage. They were open-minded enough to embrace the differences between their standard of English and that of others and did not consider American English as the benchmark or superior to other forms of English in enabling effective communication. Irrespective of rater training, Luke and Kyle with a positive attitude toward WE had their own unique rating strategies and transferred their real life experience to the rating setting. Their rating tendency focused more on successful communication than on distinctive linguistic features. On the other hand, though Nash was aware of the different versions of English used around the world, he was unable to accept such variability as real or actual forms of interactions among people that have to be taken into account in any assessments. He seemed stricter on his rating behavior and not accepting of the differences between varieties. Nash viewed standard English as superior to other forms of English in promoting effective communication.

Responding to language variations, raters expressed an uncertainty to identify unfamiliar phrases or structures as part of an examinee's variety or a result of a second language acquisition process, raising the critical issue of the extent to which raters' uncertainty in distinguishing between the two factors matter in terms of the scores they awarded (Davies *et al*, 2003). Comparing the three raters, Luke seemed to be most concerned with the issue and developed his own strategy to manipulate the interview questions in order to look for patterns of the speech and determine whether a examinee speaks his/er own variety. On the other hand, Kyle viewed the variety as forming only a small part of the language phenomena requiring attention and was not really concerned about the forms of the language as long as the communication goal was achieved. In Nash's case, variety was dismissed as an unacceptable speaking attribute in oral tests. The raters' views differed significantly and the way they handled features as a potential variety dependent entirely on the individual rater's styles and levels of acceptance or tolerance. Even though this small-scale qualitative study did not compare interview transcriptions with scores they awarded, the three raters apparently revealed distinctive rating tendency. Though focusing on different issues of concerns, the current findings conform to other studies looking into rater orientation in rating, which suggest raters have their own interpretations of what constitute L2 speaking competency (Brown, 2000; Brown *et al*, 2005; Orr, 2002).

Several themes emerged as reiterative reviews of the interview transcriptions. Even though not explicitly asked about perception of the varieties, rater belief of the varieties could be inferred from their views and stances as expressed in the interview. First, standard English seems to be an underlying criterion for some raters to judge examinees' oral proficiency performance. It was apparent a case for Nash. For Luke, his transformation of being a professional rater led him to rate more harshly and considered that examinees should not simply be understood by him but others with little exposure to varieties of English. Thus, standard English was implied in his talk a good criterion to base the rating on in that standard English should be most intelligible to most of the listeners. Second, raters' acknowledgment of WE seemed to more relate to his hometown environment, personality and education background than to his general knowledge of global English spread. Third, the role of WE in the ESL/EFL learning context was implicated by raters' endorsement of good English education examinees received at school. These three points in part reflect the "perceived cultural factor", an element that influences language attitude formation as proposed by the Social Process Model (Cargile *et al*, 1994). The "perceived cultural factor" includes static and dynamic aspects of attitude formation: the former refers to a "more static dimension and it describes the extent to which norms for correct usage have been codified, adopted, and promoted for a particular language variety" and the latter includes "status, demographic strength and institutional support" (p.226). Thus, the theoretical support and empirical data suggest a need to create items addressing the issue of standard English in the oral proficiency test along with items concerning more general views and expectations on the use of WE in educational settings and wider communication contexts.

In terms of rating tendency, it was found that raters' scoring decision was considerably influenced by his personal backgrounds, such as academic concentration and hometown environment. This conforms to the theory of uncertainty reduction that associates familiarity with positive speaker evaluation (Berger & Bradac, 1982 as cited in Cargile *et al* 1994). Thus, the conceptualization of attitude construct within the dimensions of belief and behavior tendency is outlined in Table 8 below.

Preliminary Study 2: Varietal Speaker Evaluation

In addition to the attitude attributes identified in the interview processes, an attempt was made to further elicit rater feeling attributes to facilitate the instrument construction. Forty undergraduate students in EDPSY 220 at the University of Illinois at Urbana-Champaign were recruited to complete this task. This task is named "Varietal Speaker Evaluation". It involved a rigorous process of evaluating speakers of four outer-circle varieties from India, Pakistan, Sri Lanka, and Singapore and the other four expanding-circle varieties from Taiwan, Vietnam, Turkey, and Korea. Each speaker gave a direction instruction about a map. The EDPSY220 students were asked to respond to four tasks through which the attitude attributes were elicited. Following Munro and Derwing (Derwing & Munro, 1997; Munro, Derwing, & Morton, 2006), the speaker evaluation required listeners to complete four tasks: (1) an orthographic transcription, (2) a comprehensibility rating, (3) an accent rating, and (4) an accent identification (see Appendix B). Judges were asked to listen to and transcribe each speech sample. They then marked the comprehensibility of the speech on a 9-point scale, with 9 the most difficult to comprehend and 1 the easiest. They followed the same procedures for all 8 varietal speakers. To evaluate accentedness, the judges listened to the 8 speeches again in a

Table 8

Conceptualization of Rater Attitude Constructs on Belief and Behavior Tendency Dimensions

Attitude construct	Conceptualizations of the construct
Belief	What standard English should be adopted in the oral proficiency test?
	Do raters acknowledge current status of WE worldwide?
	What are raters' views on the role of WE in the ESL/EFL learning context?
	What are rater expectations of examinees' cultural strength and language use in the oral proficiency test?
Rating tendency	What is the rater scoring tendency when encountering unfamiliar expressions in the oral test?
	To what extent do raters familiarize themselves with examinees' variety?

randomized order, using a 9-point scale with 9 representing the most strong accent and 1 no accent. Then judges provided up to three adjectives to describe what they thought of the speakers when hearing them speak English by completing the sentence “The speaker sounds. . . ”.

A key-word analysis of the adjectives was calculated and adjective pairs with a higher distribution frequency were then integrated and classified. The judges provided 125 adjectives from which pairs that were overlapping in meaning were removed. It yielded 18 adjective-and-antonym pairs selected in the construction for the semantic differential scale.

Item Construction

The tripartite attitude constructs were measured by two scale methods: the semantic differential scale assessed rater affective dimensions and the Likert scale measured rater belief and behavior tendency.

Rater Feeling

The 18 adjective-and-antonym pairs obtained from the varietal speaker evaluation were set on a 7-point scale to assess individual rater’s intensity and direction of each affective component. Following the common practice of the semantic differential scale, the positive and negative adjectives were randomly placed, that is to say, the positive adjectives were not always placed on the right side of the scale. Furthermore, the review of interview transcription also revealed four adjective pairs used by the raters to relate their feeling to a variety:

Interesting/boring; difficult/easy; natural/ unnatural; comfortable/uncomfortable

Thus, together with the interview study and varietal speaker evaluation, a total of 22 pairs of adjectives were generated. All pairs of adjectives were presented in Table 9.

Table 9

Adjective-and-Antonym Pairs

intelligent/unintelligent	nervous/relaxed (clam)	confident/unsure
certain/uncertain of grammar	articulate/unclear	fluent/not fluent
sure/hesitant	knowledgeable/uneducated	quick/slow
thoughtful/inconsiderate	timid/happy	enthusiastic/indifferent
kind/unkind	friendly/unfriendly	informative/unhelpful
Interesting/boring	difficult/easy	natural/ unnatural
comfortable/uncomfortable	rushed/easy	quiet/loud
choppy/ weak		

The sample of the semantic differential scale is presented in Table 10. For the full version of the scale, see Appendix C.

Table 10

Rater Feeling as Measured by the Semantic Differential Scale

The speaker sounds . . .	1	2	3	4	5	6	7	
Articulate								Unclear
Inexperienced								Experienced
Intelligent								Unintelligent
Slow								Quick
Knowledgeable								Uneducated
Unkind								Kind
Fluent								Not fluent
Good-natured								Hostile
Considerate								Inconsiderate

Rater Belief

Two subscales of items measuring rater belief were developed: perceived cultural factors and expectation of Indian English.

Perceived culture factor. Eight statements concerning raters' perceived cultural features of the varieties were constructed. The questions referred to the extent to which variety of English should be allowed or supported in second language assessment, institutional and cross-cultural settings. Another five statements measured raters' knowledge about WE spread, including raters' attitude towards WE status, the recognition of demographic strength (the number and distribution of WE speakers) and acknowledge of WE as a subject to be taught or a medium used in ESL/EFL learning contexts.

Expectation of Indian English. This category indirectly measured rater belief by elicitations of rater expectation of Indian English. According to language expectancy theory (Burgoon and Miller, 1985, as cited in the Cargile *et al* 1994), the discrepancy between expected and actual language use leads to negative evaluations of the speaker. Eleven items that measured raters' expectation of Indian English were included.

Rater Behavioral Tendency

Raters' behavioral tendencies were measured from the two perspectives: rating tendency and familiarity with WE.

Rating tendency. Twenty-one items directly asked raters' rating tendencies in relation to different aspects of variety as spoken by examinees during the test, such as the language use by the examinees, strategies used to achieve communication goal and raters' role as active listeners. Items inquiring about raters' actual behaviors are omitted as they are influenced by many things besides attitude and therefore were not always accurate indices of attitudes (Mueller, 1986).

Interpersonal history. Fifteen items in this category were designed to include the amount of raters' exposure to varieties, familiarity with varieties and general knowledge about WE.

Rater Biasing Factors

Apart from rater attitude constructs, five rater biasing factors reviewed in chapter 2 were included: rater educational and professional experience (Chalhoub-Deville, 1995), residency (Kim 2005); Chalhoub-Deville & Wigglesworth, 2005), rater nationality and native language (Brown, 1995), and gender (McKenzie, 2008; O'Loughlin; 2002), some of which contributed to the extraneous variables that affected scores awarded. Hence, this section seeks to identify which biasing factor is associated with rater attitude and ultimately takes effect on rater scoring decisions.

The final draft of the RAI consisted of 60 Likert items and 22 semantic differential scale items. The former included 42 positive and 18 negative statements. It was intended to generate a larger item pool than actually needed as some of the items were expected to be deleted after exploration phase of the RAI construction. See Appendix C for the full version of the instrument.

Content Review

Upon completion of the item writing, items were reviewed by the study researcher, two doctoral students specializing in second language assessment and two researchers with a background in second language acquisition and sociolinguistics respectively for clarity, representativeness and comprehensiveness of the items and whether items leads to response bias.

Construct Validity: Construction Phase 2

The construct validity was tested in the two phases: exploratory and verification. The results of each component of the RAI (i.e. part A, measure of rater feeling, part B, measure of rater belief and part C, rating tendency) were presented respectively.

The RAI was delivered on-line. Twenty raters (see chapter 3) indicated their feelings about eight IELTS Indian examinees on the 7-point semantic differential scale. They then proceeded to the 5-Likert questionnaire measuring rater belief and rating tendency. Each scale was accompanied by a comment section for raters to provide further feedback, such as clarity of the items and the appropriateness of study flow. The time length for the entire study was approximately an hour. Each rater received $20 remuneration upon completion of the study.

Next, data were analyzed to determine the suitability of each item, the scale and to remove undesirable items, if any.

Measure 1: Rater Feeling

Descriptive Statistics and Internal Consistency

Each rater provided ratings on the 25 semantic differential scale items for each of the 6 Indian speakers, which yielded a total of 180 observations. The item means, standard deviations, internal consistency, univariate normality and correlation matrix were computed and examined. Table 11 presents the mean and standard deviation for the data set. Item “knowledge”, had the highest mean score of 5.7, whereas item “aggressive” had the lowest mean score of 3.14. Of the 25 items, only three items had a mean lower than 4. The mean for the data set is 4.84, which gave initial observation that raters’ feeling of Indian speakers was quite positive.

Table 11

Means and Standard Deviations for Feeling Attributes

Pair	Mean	Standard Deviation (SD)	Skewness	Kurtosis
Clear	4.97	1.499	-.396	-.689
Sure	4.68	1.972	-.319	-1.479
Enthusiastic	4.69	1.498	-.555	-.234
Fluent	5.54	1.446	-.984	.244
Confident	5.24	1.683	-.702	-.741
Calm	4.95	1.676	-.273	-1.181
Intelligent	5.81	1.173	-.982	.906
Thoughtful	5.52	1.125	-.324	-.673
Happy	4.44	1.249	.183	-.483
Quick	4.96	1.297	-.381	-.559
Knowledgeable	5.70	1.098	-.544	-.662
Kind	5.00	1.036	.477	-.829
Friendly	4.96	1.145	-.046	-.376
Informative	5.18	1.242	-.756	.591
Easy	3.67	1.452	.219	-.472
Quiet	3.97	1.045	-.019	.752
Strong	4.26	1.198	.050	.017
Organized	4.91	1.434	-.756	-.054
Experienced	4.73	1.449	-.350	-.340
Good-natured	5.36	1.123	-.188	-.730
Pleasant	5.17	1.229	-.451	.108
Considerate	5.07	1.074	.069	-.705
Talkative	4.59	1.330	-.131	-.786
Aggressive	3.14	1.461	.042	-.768
GoodPro	4.58	1.474	-.197	-1.007

Cronbach’s Alpha was calculated to assess internal consistency. An alpha of .880 was obtained, which indicates a high level of internal consistency (de Vaus, 2002; George & Mallery, 2003). To test the assumption of univariate normality, skewness and kurtosis were checked. A more liberal recommendation on the acceptable levels as proposed by Kline (2005) was used: cutoff of -3 to +3 for skewness and -10 to +10 for kurtosis respectively. The skewness of the 25 semantic differential scale items ranged from -.984 to .477, and the values for kurtosis ranged from -1.479 to .906, indicating the responses were normally distributed and well within the liberal recommendation.

The examination of correlation matrix for item consistency revealed several items may be problematic. Items that are too highly correlated (i.e. r >.80) suggesting multicolinearity whereas items not correlated sufficiently (i.e. r <.30) indicated not much shared common variance could be generated which may yield as many factors as items (Pett, Lackey, & Sullivan, 2003). The correlation matrix showed that no items had multicolinearity problems but seven items had low correlation with most of other items: enthusiastic, thoughtful, friendly, easy, quiet, good-natured, and aggressive. These items were not removed and thus were evaluated again against other criteria when running factor analysis in order to verify whether those low correlations were spurious or (alternatively) helped to clarify the factor structure. Table 12 presents the abridged correlation matrix for item 1 to 8. For the complete correlation matrix for feeling attribute, see Appendix D.

Exploratory Factor Analysis

A principal component analysis (PCA) was performed in an attempt to obtain preliminary information regarding the potential number of dimensions of the scale, i.e. the latent factors representing the items in the scale. Ratings on each of the 25 semantic differential scale items for each of the 6 Indian speakers (i.e. a total of 150 observations) were assessed for suitability. All data was collated in an SPSS file. Bartlett’s test of sphericity was significant (*p* =0.0000), and the Kaiswer-Meyer-Oklin, index for comparing the magnitude of the observed correlation coefficients to the magnitude of the partial correlation coefficients was 0.856, well

Table 12

The Correlation Matrix for Item 1 to Item 8 of the Rater Feeling Attributes

	Clear	Sure	Enthusiastic	Fluent	Confident	Calm	Intelligent	Thoughtful
Clear	1.000							
Sure	.455**	1.000						
Enthusiastic	.125	.283**	1.000					
Fluent	.537**	.659**	.090	1.000				
Confident	.361**	.652**	.272**	.531**	1.000			
Calm	.385**	.465**	-.069	.465**	.405**	1.000		
Intelligent	.489**	.455**	.123	.538**	.543**	.532**	1.000	
Thoughtful	.270**	.253**	.205*	.246**	.358**	.272**	.545**	1.000

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

exceeding the recommended value of 0.6 (Pett, *et al*, 2003). Based on these initial findings, factor analysis was deemed appropriate to analyze the data.

Factor extraction and rotation

The PCA using oblimin rotation method was conducted. The choice of oblimin rotation method is based on the assumption that items or factors of rater feeling are most likely correlated to some degree (cf. Pett *et al*, 2003). Five components with eigenvalues greater than one were extracted. The scree plot was examined which revealed a four- or three-factor model may represent the data adequately given a marked change in slope after three factors. The PCA was conducted a second time to force extractions of only four and three components respectively. Criteria that determined the acceptable number of the factor included: (1) items load substantively (>.30) on only one factor, (2) items load at approximately zero (+0.10 to -0.10) on some other factor (Tabachnick & Fidell, 1989) and (3) interpretability. That is to say, the ultimate decision about the number of factors to extract was based on simple structure and the interpretative clarity of the loadings. As a result, the three factor model reached the satisfactory results and was selected. Figure 17 shows the Scree plot for the three factor model.

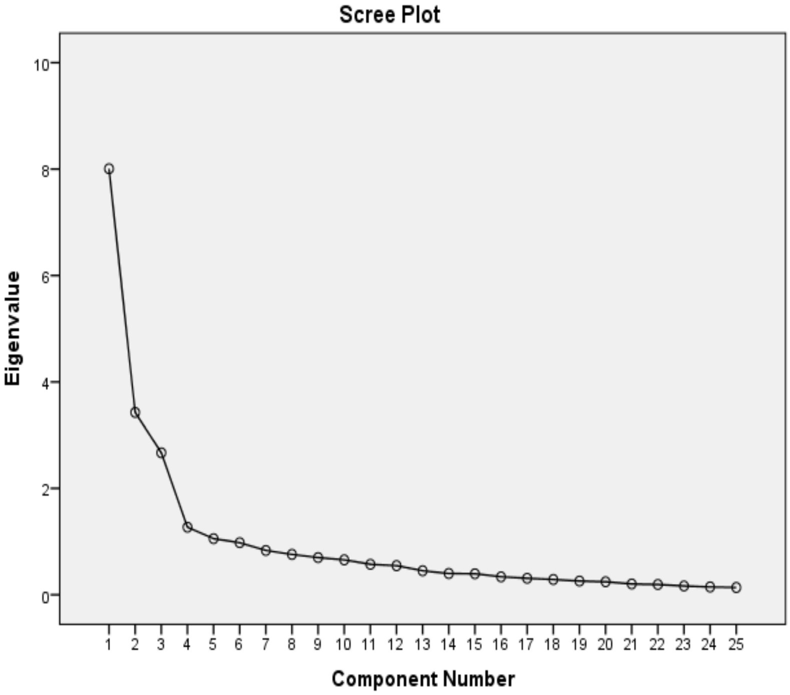


Figure 17. Scree plot for the three factor model.

Next, each item was evaluated for possible removal so as to maximize the explained variance. Item communality that measures the proportion of variance of a particular item that is explained by all the factors jointly is used as a guideline for item deletion (Worthington & Whittaker, 2006). Item communalities greater than 0.8 is considered high (Velicer & Fava, 1998). Nevertheless, in social science data, more common magnitudes are low to moderate communalities of 0.40 to 0.70 (Costello & Osborne, 2005). Thus, it was decided to remove any item that has communality of less than 0.50 because these items are not highly correlated with one or more of the factors in the solution. As a result, six items were removed: enthusiastic, thoughtful, happy, easy, quiet and strong. PCA was performed again to evaluate whether all the item communality was improved and above 0.50. Three more items, calm, aggressive and organized, that had a communality smaller than 0.50 were deleted. Another PCA was performed to evaluate whether all the item communality was improved and above .50. Table 13 presents the communality in the final model.

Table 13

Communality in PCA

Pair	Communality	Pair	Communality
Clear	.748	Good-natured	.760
Sure	.719	Considerate	.760
Fluent	.718	Talkative	.649
Intelligent	.613	GoodPro	.757
Quick	.644	Kind	.764
Knowledgeable	.599	Informative	.581
Experienced	.593		

Table 14

Results of Principal Component Analysis

	Factor loadings		
	1	2	3
Clear	**.853**	.091	-.138
GoodPro	**.840**	-0.18	-.026
Intelligent	**.752**	.268	-.372
Fluent	**.731**	-0.34	-.575
Knowledgeable	**.670**	.312	-.519
Good-natured	.191	**.868**	-.252
Kind	.011	**.865**	-.134
Considerate	.136	**.865**	-.321
Talkative	.214	.271	**-.807**
Quick	.105	.100	**-.782**
Sure	.612	.027	**-.710**
Experienced	.336	.430	**-.710**
Informative	.159	.504	**-.668**
Eigenvalues	4.886	2.406	1.585
% of variance accounted for	37.587	18.505	12.192
Cronbach's Alpha	.839	.851	.798

Factor independence and reliability

Reliability was calculated for the remaining thirteen items. Cronbach's Alpha for the total semantic differential scale is .846. Each factor also demonstrated an acceptable degree of internal consistency, with Cronbach's Alpha above .80. The final three-factor model accounted for 68.284% of the total variance.

Table **14** reports the results of the PCA factor analysis.

Correlations between factors are presented in Table 15. Correlations between the factors are relatively low, ranging from -.251 to .096, supporting the finding of factor analysis that semantic differential scale measured relatively distinct dimensions of rater feelings.

Table 15

Correlations between Factors

Factor	1	2	3
1	1.000	.096	-.291
2	.096	1.000	-.251
3	-.291	-.251	1.000

Closer examinations of the items and factors revealed several items may be re-classified to different factors due to interpretability. Currently, Factor 1 includes *Clear, Good Pronunciation, Fluent, Intelligent and Knowledgeable*. It seems to indicate a speaker's sound quality along with his/er intellectual level of the speech play similar weight in raters' evaluation. On the other hand, as

Table **14** presents, item, *Sure*, with a loading of .612, may be grouped into factor one too. As for the other two items in Factor 1, *Intelligent* and *Knowledgeable*, they may be grouped into factor 3 implying a speaker's confidence level. Thus, before labeling the factor and justifying factor interpretation, different combinations of the items should be further factor analyzed to determine the best model for interpretation. This will leave to the next phase of analysis when confirmatory factor analysis is performed. Two models, based on the current item distribution and easy interpretability respectively, will be tested out. The items distributions in the current and proposed models are summarized in Table 16. Item in italics in Model 2 are the proposed changes.

Table 16

Two Models for Confirmatory Factor Analysis

Model 1 (current PCA results)		
Factor	1	Clear, Good Pronunciation, Fluent, Intelligent, Knowledgeable
	2	Good natured, Kind, Considerate
	3	Talkative, Quick, Sure, Experienced, Informative
Model 2 (alternative model based on interpretability)		
Factor	1	Clear, Good Pronunciation, Fluent, *Sure*
	2	Good natured, Kind, Considerate
	3	Talkative, Quick, Experienced, Informative, *Intelligent*, *Knowledgeable*

Measure 2: Rater Belief and Rating Tendency

Analysis of the data for measure 2 included examining the reliability of Likert scale items based on Classical Test Theory (CTT). Negatively worded items were reverse coded prior to the analysis so that higher scores indicated a more positive belief or rating tendency. Cronbach’s Alpha was calculated to examine internal consistency. That is, to examine whether the scale items are all measuring the same underlying attributes. Table 17 shows the results of reliability analysis of the 61 Likert scale items. The reliability estimates for the variables range from .260 to .557 with Cronbach’s Alpha of .609 for the overall measure 2. As it is recommended that a minimum Cronbach’s Alpha of .70 is needed to demonstrate a good internal consistency (de Vaus, 2002; George & Mallery, 2003), all items were re-examined if they went below the desirable value. Alphas if-item-deleted along with the qualitative input provided by the raters were examined. Twenty-one problematic items across sections were revised or removed to improve the clarity of the questionnaire. This resulted in 35 items remained in the revised scale. Cronbach's Alpha for the entire questionnaire was improved to .738. See Table 17 for the reliability estimates of modified scale. As illustrates in Table 17, the Cronbach's Alpha for each variable was also improved, even though only the section, Expectation of Indian English, met the .70 cutoff value. The other three sections along with their new Cronbach's Alpha were as follows: Rating Tendency (.597), Perceived Cultural Factor (.590) and Interpersonal History (.457).

Table 17

Reliability Estimates of Measure 2 in Exploratory Phase

Variables	Number of items	Cronbach's Alpha	Revised number of items	New Cronbach's Alpha
Rating tendency	21	.557	9	.597
Interpersonal history	15	.260	6	.457
Perceived cultural factor	13	.515	12	.590
Expectation of Indian English	12	.422	6	.726
Overall	61	.609	35	.738

Construct Validity: Construction Phase 2

The modified RAI was administered to 96 ESL teachers in the U.S. and India, 23 of which were IELTS raters at the time of the study. See chapter 3 for rater background descriptions. Raters were asked to respond to the RAI delivered online as well as provided ratings to the six IELTS descriptive task samples. The RAI's psychometric structure is further verified and the results are reported in this section. The ratings of the IELTS speaking samples along with the RAI scores were used for further analysis as will be described in the next chapter.

Procedure

A URL address to access the study materials was sent out to the raters through Netfiles, an online service tool available to students and faculty at the University of Illinois at Urbana-Champaign. The study materials included the followings:

1. The modified RAI
2. Six IELTS descriptive task samples
3. Instructions for the study (see Step1 to 3 below)
4. A consent form (see Appendix E)

Participants were instructed to read the instructions and sign the consent form before proceeding to the study according to the following procedures:

Step 1. RAI Part 1: IELTS descriptive tasks

1.1 Listen to an IELTS descriptive task

1.2 Rate the IELTS descriptive task according to the four criteria: Fluency, Pronunciation, Sentence Structure, and Vocabulary. No prior training on the use of the criteria is given. Each criterion is measured on the 10-point scale, ranging from 0-9, with 0 represents the lowest and 9 the highest oral proficiency level.

1.3 Repeat the steps above for the remaining five IELTS descriptive tasks.

Step 2. RAI Part 2: Rater Belief and Rating Tendency

Respond to the questionnaire of rater belief and rating tendency, in a total of four sections comprising 32 questions

Step 3. RAI Part 3: Rater Feeling

3.1 Listen to the IELTS descriptive task

3.2 Indicate how you feel about the speaker by responding to the seven-point semantic differential scale

3.3 Repeat the steps above for the remaining five speech samples

An email reminder was sent to the raters two weeks after they received the link to the study. The time length for the entire study was approximately one hour. Each rater was compensated $15 for his/er participation.

Measure 1: Rater Feeling

SPSS 17.0 for Windows (2009) was used to analyze demographics and compute Cronbach's Alpha. Internal consistency reliability of the full RAI and each subscale was examined using Cronbach's Alpha. All statistical analyses were interpreted with an Alpha level of .05.

A confirmatory factor analysis (CFA) was conducted using AMOS Version 7.0 to determine the plausibility of the three-factor structure generated by EFA in the previous phase. Multiple fix indices were used for evaluating the goodness-of-fit of the model. The indices used include: chi-square, the comparative fit index (CFI), root mean square error of approximation (RMSEA) and Tucker Lewis Index (TLI). Values >.95 were indicative of good model fit for CFI and TLI; RMSEA close to .06 or less indicate good fit (Hu & Bentler, 1999). Squared multiple correlation that explains the variance accounted for by the factor was also examined.

Two a priori models as identified in the previous phase (see Table 16) were examined using CFA. Model 1 was generated by exploratory factor analysis. Model 2 was a proposed alternative model based on interpretability. Modification indices (MI) and examination of residuals are used to improve the model fit. In order to examine the MI, the data needs to be completed without missing values. The current set of data contains 36 missing values. The expectation maximizing (EM) algorithm (Dempster, Laird, & Rubin, 1977 as cited in Beadnell, Baker, Gillmore, Morrison, Huang, & Stielstra, 2008) was performed to impute missing data as recommended by Schafer and Grahan (2002) to minimize bias when only a small amount of missing data occurred. As a result, the full 596 sample size was retained for further CFAs.

Descriptive Statistics and Internal Consistency Estimate

Ninety-six raters each rated 6 IELTS descriptive tasks on the 13 semantic differential scale items, which yielded a total of 576 observations. The item means, standard deviations, correlation matrix were computed and examined. Table 18 presents the mean, standard deviation, skewness and kurtosis for the data set. Item "*Good natured*", had the highest mean score of 5.54, whereas item "*Good Pronunciation*" had the lowest mean score of 4.70. Of the 13 items, only five items had a mean lower than 5. The initial screening of the mean for the data set provided some implications for raters' feeling tendency. Raters as a whole generally had positive feeling of the Indian speakers. To test if the variables used demonstrate multivariate normality as assumed by the CFA, results were assessed through the inspection of univariate normality index values, with skewness indexes smaller than absolute cutoff value of 3 and kurtosis indexes smaller than absolute cutoff value of 10 indicative of liberal normality (Kline, 2005). Except for one item, *Educated*, that has lowest kurtosis value (-1.061), skewness and kurtosis indices for all items were between -1 to +1, and again, well within the liberal range. The assumption of normality was met. In terms of internal consistency, Chronbach's Alpha was .904 for the semantic differential scale, well above the recommended .70 cutoff for good internal consistency reliability (de Vaus, 2002; George & Mallery, 2003).

Table 18

Means and Standard Deviations for Feeling Attributes

Pair	Mean	Standard Deviation (SD)	Skewness	Kurtosis
Clear	4.97	1.644	-.418	-.969
Experienced	4.96	1.485	-.468	-.488
Intelligent	5.23	1.373	-.448	-.490
Quick	4.87	1.233	-.233	-.302
Educated	5.30	1.271	-.281	-1.061
Kind	5.34	1.121	-.657	.020
Fluent	5.12	1.637	-.605	-.449
Good-natured	5.54	1.114	-.490	.357
Considerate	5.38	1.178	-.309	.028
Talkative	4.94	1.365	-.647	-.012
Good Pro	4.70	1.629	-.503	-.908
Sure	5.05	1.513	-.402	-.356
Informative	5.29	1.428	-.298	-.540

Confirmatory Factor Analysis

The two 3-factor a priori EFA models were evaluated by CFA. According to the fit indices, both models did not fit the data adequately. An examination of the squared multiple correlation explaining the variances accounted for by each of the thirteen items revealed that two of the items in each priori model may be problematic due to low variance. These items were *Talkative* and *Quick*. The items then were removed in each model and CFA was re-run. Table 19 provides a summary of CFA goodness-of-fit indices by analysis for the two models. The fit indices for Model 1 and 2 show that χ^2 statistics was significant for both models, suggesting an inadequate fit of the models to the data (*Model 1*: χ^2 =325.900, df=41, *p=000;*

Model 2: χ^2 =198.208, df=41, *p*= 000). Other fit indices were examined to determine the best model. As shown in Table 19, the fit indices for Model 2 yielded the better model fit and met the cutoff criteria for acceptable levels. The Comparative Fit Index (CFI; Bentler, 1990) value was 0.959, the TLI was 0.945, and the RMSEA of 0.082 was within the recommended range of model fit (Byrne, 2001). The chi-square difference between the two models is 127.692, indicating a significant improvement (*p* <.001) in model fit. Thus, the results suggest Model 2 better fits the data and will be used for further analysis.

Table 19

Summary of CFA Goodness-of-Fit Indices for the Two Priori Models

Model	χ^2	*df*	*p*	RMSEA	CFI	TLI
1	325.900	41	.000	0.110	0.926	0.901
2	198.208	41	.000	0.082	0.959	0.945

Note. RMSEA=Root Mean-Square Error of Approximation. CFI= Comparative Fit Index. TLI= Tucker-Lewis Index

Figure ***18*** illustrates Model 2, the three-factor correlated model for rater feeling with the standardized solutions obtained from the AMOS output. The factor loadings were moderately high ranging from .624 to .937. The largest and lowest coefficients (i.e. *Considerate* and *Kind*) were presented by the two indicators concerning speaker's kind-heartedness. There was moderate correlation between the three factors. The highest correlation was noted between the Speech Competency and Level of Confidence (r =.899) followed by Level of Confidence and Kind-heartedness at a moderate .481 while the lowest was between Speech Competency and Kind-heartedness (r = .284). Although the correlation between Speech Competency and Level of Confidence was quite high compared to the other factors, the three-factor correlated model 2 was thought to appropriately fit the data as hypothesized.

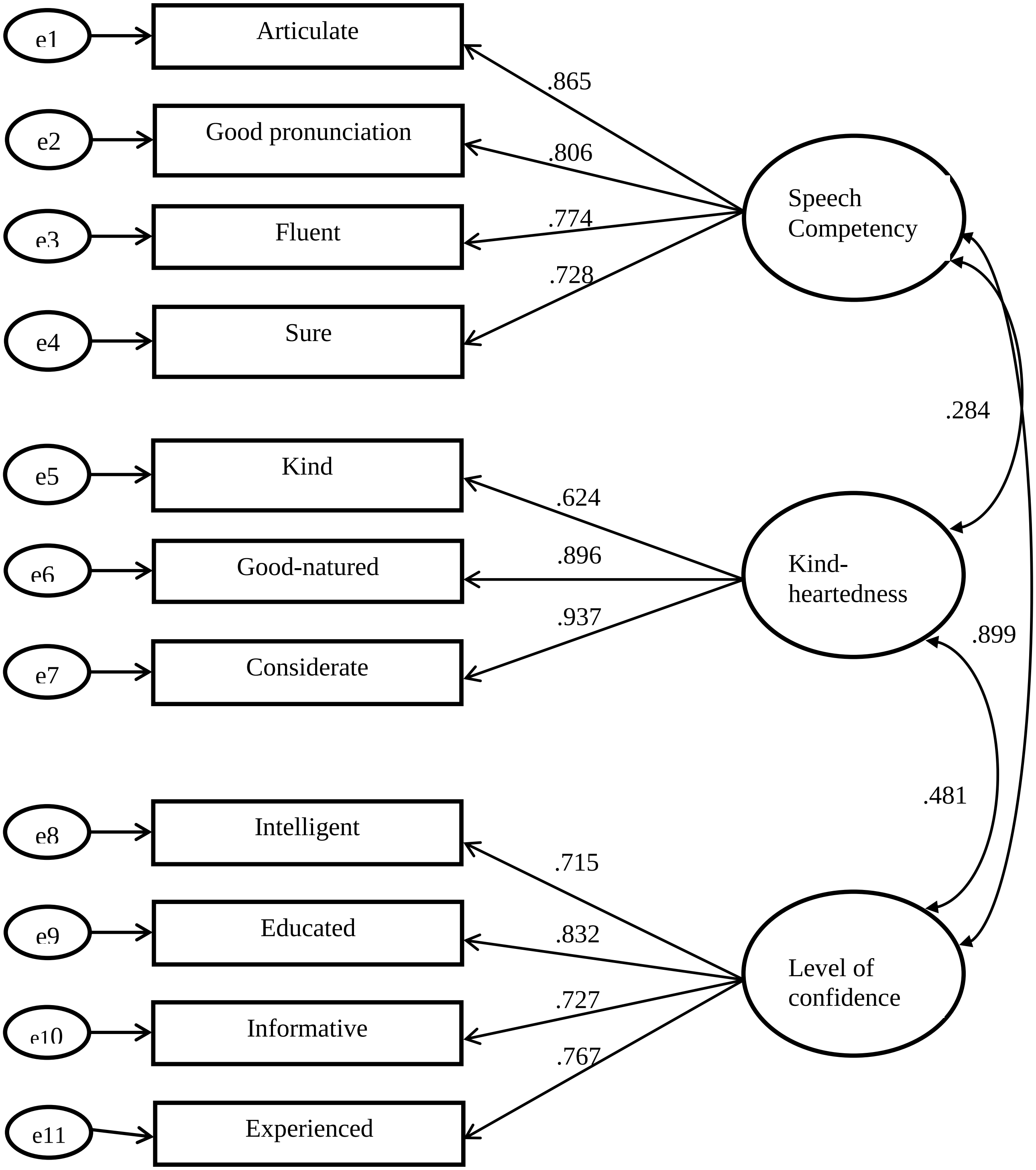


Figure 18. Model 2, the three-factor correlated model for rater feeling

Reliability estimate.

Finally, Cronbach's Alpha was re-calculated to estimate reliability based on the instrument and reconstructed subscales with two items deleted. Cronbach's Alpha for the reconstructed 11 item instrument is 0.897 versus 0.904 for the original 13 item.

Factor labeling.

Upon completion of factor determination and item selection, the factors were reviewed by the current researcher and two PhD students in humanities in an attempt to name the factors. Factor 1 contained four items that reflected the speech performance, including clear, good pronunciation, fluent and sure and was labeled "speech competency". Factor 2 was composed of three items about being kind, good-natured and considerate, reflecting speaker's characteristics or attractiveness to the listeners. Thus, this factor was labeled "kind-heartedness". Factor 3 included four items that reflected the degree of a speaker's confidence. These items were intelligent, educated, experienced and informative. Factor 3 was thus labeled "level of confidence".

Measure 2: Rater Belief and Rating Tendency

Descriptive Statistics and Test of Normality

Table 20 presents the mean and standard deviations for the four subscales. Eight missing data points were detected and mean substitution was used to replace the missing data. Note that the following three items are dichotomous: C4, C2 and C3. The normality assumption using skewness and kurtosis indices were inspected. As before, the acceptable range for normality is absolute value of skewness index lower than 3 and kurtosis index absolute value lower than 10(Kline, 2005). The skewness index for item C211 fell slightly out of the acceptable range (-3.063) so normality was assumed for this data set.

Table 20

Distribution of Items Measuring Rater Belief and Rating Tendency

	Item	Mean	SD	Skewness	Kurtosis
Section 1:Expectation (Belief)	C11	3.19	1.059	-.060	-1.207
	C12	4.12	.861	-1.154	1.620
	C13	2.75	1.654	.253	-1.590
	C14	3.08	1.149	-.251	-.852
	C15	2.15	1.046	.876	.507
	C16	2.80	1.120	.138	-.787
Section 4: Cultural factor (Belief)	C41	3.47	.994	-.601	-.004
	C42	3.85	1.105	-.892	.155
	C43	3.92	.991	-.757	.193
	C44	4.27	1.035	-1.443	1.332
	C451	.43	.497	.300	-1.951
	C452	.21	.408	1.459	.132
	C453	.75	.435	-1.173	-.638
	C454	.14	.344	2.165	2.744
	C46	4.45	.692	-1.090	.707
	C47	2.64	1.025	.356	-.166
	C48	4.14	.969	-1.128	.726
	C49	3.99	.946	-.665	-.425
	C410	3.78	1.028	-.311	-.570
	C412	4.45	.915	-2.214	5.268
	C413	4.62	.617	-1.696	3.019
Section 2: Rating tendency (Rating tendency)	C211	.92	.278	-3.063	7.540
	C212	.60	.492	-.433	-1.852
	C213	.70	.462	-.876	-1.260
	C214	.60	.492	-.433	-1.852
	C215	.67	.474	-.718	-1.516
	C22	3.16	1.173	-.119	-1.006
	C23	3.77	1.137	-1.013	.320
	C24	3.25	1.076	-.111	-.966
	C25	3.99	.747	-.292	-.341
	C26	4.23	.756	-.718	.073
	C27	2.55	1.139	.477	-.699
	C28	2.58	1.075	.278	-.553
Section 3: Familiarity (Rating tendency)	C311	.78	.773	-1.633	3.893
	C312	.42	.868	-1.864	4.226
	C313	.97	1.004	.846	.311
	C32	4.39	.800	-1.089	1.120
	C33	4.35	1.212	-.863	-.057
	C34	2.29	.416	-1.382	-.091
	C35	4.23	.496	.343	-1.923
	C36	3.66	.775	-2.474	8.560

Internal Consistency

Table 21 shows the reliability estimate for the entire Likert scale and each sub-scale. Cronbach's Alpha of .602 for the total scale shows somewhat acceptable internal consistency of the items (de Vaus, 2002; George & Mallery, 2003). The reliability estimates for each subscale in the current phase is generally lower than those in the exploratory phase, except for the last subscale, Interpersonal History, which was improved from .457 to the current .518. Cronbach's Alpha for other sub-scales are as follows: Expectation of Indian English (.474), followed by Perceived Cultural Factor (.383) and Rating Tendency (.361). Reasons that caused low reliability were most likely the small number of the items in the current sections (Symonds, 1928). Other potential reasons, as Symonds pointed out, could be the wider range of difficulty of items. In the current study, it could be explained by the fact that raters' beliefs in WE and rating tendency greatly differed from each other which led to low reliability. To improve the internal consistency, the 'alpha if item deleted' was checked which suggested removing item 24, "When examinees use unfamiliar expressions, it decreases their intelligibility", would improve alpha to .628 for the total scale. Thus, this item was discarded for further analysis.

Table 21

Reliability Coefficients

	Variables	Cronbach's Alpha in 2nd phase	Cronbach's Alpha in 3r^{d} phase
Belief	Expectation of Indian English	.726	.474
	Perceived cultural factor	.590	.383
Rating tendency	Rating tendency	.597	.361
	Interpersonal history	.457	.518
	Overall	.738	.602

Item Frequency

The participants were asked to respond to 32 items across the four sections. Items scored 1 indicate ‘strongly disagree’, 2, ‘generally disagree’, 3, ‘neutral’, 4, ‘generally agree’ and 5, ‘strongly agree’. An option of “un-ratable” was also included to allow any uncertainty in responding to items. Item responses were scored so that the higher the total score, the more positive the participants about their belief in WE and rating tendency. A comment box was placed at the end of each section to elicit rater qualitative feedback on items or issues concerned, except for that of “Interpersonal History” where the comment box was accidentally removed during editing.

Expectation of Indian English. As reported in Table 22, the results on one of the two measures in rater belief, Expectation of Indian English, were generally positive. Item 3 identified rater experience in rating Indian examinees and almost half of the raters (49.5%) disagreed with the statement, indicating that the other half of the raters had varying amount of experience in rating Indian examinees. Raters’ familiarity of Indian English was investigated in two items. Item 1 showed that near half of the raters (47.4%) had no difficulty comprehending Indian speakers in non-test situations; nevertheless, in the context of language assessment, 42.3% of the raters indicated a need to make more listening efforts to figure out Indian examinees’ intended messages. Raters’ positive attitude toward Indian English also demonstrated in two items concerning the status of Indian English and the extent to which Indian speakers can be categorized into native speakers of English. The majority of the raters (83.6%) agreed that Indian English is not an irregular dialect but a steady variety that present its own distinctive linguistic features. Related to this, more than a third of the raters (43%)

considered Indian speakers as native speakers of English while another third of raters (31.9%) disagreed to this statement.

Table 22

Frequency of Rater Expectation of Indian English

		Percentage (%)					
	Items	SD	GD	N	GA	SA	Un-ratable
1	I have no problem understanding Indian speakers in non-test situations.	2.1	33.0	16.5	39.2	8.2	1.0
2	Indian English is a steady variety that has its own linguistic features.	1.0	5.2	9.3	48.5	35.1	1.0
3	I have experience in rating Indian examinees.	37.1	12.4	13.4	10.3	25.8	1.0
4	Indian speakers may be treated as native speakers of English nowadays.	10.3	21.6	24.7	34.0	8.2	1.0
5	Indian speakers should not be exempted from English proficiency tests.	4.1	5.2	21.6	36.1	29.9	3.1
6	I need to make more effort to understand Indian examinees.	6.2	22.7	25.8	29.9	12.4	3.1

Qualitative feedback. Of the 96 respondents, 32 provided written feedback on items concerned. Of the four subscales/sections on the questionnaire, Expectation of Indian English elicited most inputs from the raters. Comments in this section were mainly concerned about three issues: pronunciation, status of native speaker, and the need for Indian students to English proficiency test. Some of the excerpts are presented below.

Pronunciation issue:

> 5-1[1]. Some I cannot understand at all, 95% because of pronunciation and 5% because of word order and word choice.
>
> 90-1. Any difficulty understanding Indian speakers stems from their pronunciation issues. (e.g., retroflexed “r”s, suprasegmentals such as intonations)

[1] The first number is the rater code and the second refers to the segment of the rater comments.

Status of native speaker:

> 82-1. There is too much variation among Indian speakers to be able to treat them all as native English speakers. [generally neutral]
>
> 42-1. While I would like to say that Indian speakers (and, for that matter, speakers of other varieties of English) may be treated as native speakers, I don't think that this is the pervasive opinion among the general public (in the US). [generally positive]
>
> 93-1. Besides the TESOL world, few people would be open to accepting Indian English as its own dialect and would ask students and employees to take ESL classes to improve their accents. [touching on ideological issues]

Need to take English proficiency test:

> 12-1. As a group, they should not be exempted from English tests. However, there are some Indian speakers who are native English speakers. It is ridiculous to keep testing them. There should be some kind of 'uber-certificate' that would exempt international speakers (not only those from India) who do have 'native like' skill levels from having to take any more language tests! [generally neutral]
>
> 11-1. Any native speaker of Indian English should not have to take an English test but they should be aware of the fact that their variety may be discriminated against by American and British English speakers. They might have to adjust their variety to meet other sociocultural expectations. [generally positive]

Perceived Cultural Factors. Another component for the belief dimension on the questionnaire, perceived cultural factors, was reported in Table 23. This section concerned raters' belief in the effects of WE in daily and cross-cultural communication, status of WE in ESL or EFL teaching and learning, and the necessity of adopting standard English in the oral proficiency assessment. When asked if a standard English, such as British or American English, should be used to judge examinee's performance in communicative-based testing, 63.2% of the rater agreed to this statement. Nevertheless, close to a third of the raters (27.1%) expressed neutral views on this statement. Raters' views on the promotion of the status of the variety to a legal and standard status seem less positive. 44.8% of the rater agreed to this statement whereas 35.4% expressed a neutral choice, which is the highest percentage for the "neutral" choice across all items on the questionnaire. In terms of ESL/EFL learning and rater training, the majority of the raters agreed that learners or raters should be exposed to different varieties in the

Table 23

Frequency of Raters' Perceived Cultural Factors

		Percentage (%)					
	Items	SD	GD	N	GA	SA	Un-ratable
1	Standard English (e.g. American English) should be used to judge examinees' performance in the test setting.	4.2	12.5	27.1	41.7	11.5	3.1
2	Varieties of English are not appropriate to use in cross-cultural communication.	32.3	37.5	15.6	9.4	4.2	1.0
3	Native speakers of English do not best serve as raters of oral English test (e.g. TOEFL, IELTS).	31.3	37.5	21.9	6.3	3.1	0.0
4	Varieties of English are not appropriate in everyday communication.	56.3	27.0	7.3	7.3	2.1	0.0
6	Language learners should develop an awareness of the global spread of English.	0.0	1.0	8.3	35.4	55.2	0.0
7	Unless varieties of English are promoted via educational efforts, such as by being codified in the dictionary, they can't obtain legal status and become standard.	5.2	13.5	35.4	32.3	12.5	0.0
8	Language learners should be exposed to different varieties of English.	1.0	8.3	9.4	37.5	42.7	0.0
9	Native speakers of English do not best serve as English language teachers.	34.4	39.6	16.7	9.4	0.0	0.0
10	Speakers of non-standard varieties (i.e., not British or American English) currently outnumber native speakers of standard English.	1.0	10.4	27.1	33.3	27.1	0.0
12	Raters of speaking tests (e.g. TOEFL, IELTS) should have opportunities to be exposed to varieties of English during training.	3.1	2.1	4.2	27.1	62.5	0.0
13	Raters of speaking tests (e.g. TOEFL, IELTS) should develop an awareness of the global spread of English.	0.0	1.0	4.2	26.0	68.7	0.0
5	In the region where I live, I think the following variety should be taught in English as a second or foreign language classes (select all that apply): a. Local English (42.7%) b. British English (29.8%) c. American English (75.0%) d. Other (please specify)						

context of learning or training and to develop an awareness of the global spread of English (see item 6,8,12, and13). Raters' beliefs on the role of the native speakers that serve as a rater and language teachers were not all the same. While the majority of the raters were positive about the role of the native speakers in the rating and teaching contexts (see item 3 and 9), closer to one third of the raters (i.e. 31.3% and 26.1% respectively) expressed neutral or less positive stance, which seems to imply raters' endorsement, to some degree, to the non-native speakers serving as raters in the oral proficiency assessment and ESL/EFL teachers respectively.

Qualitative feedback. A total of 17 feedbacks were elicited regarding different aspects of language use that raters focused on in decision-making processes. The following excerpts revealed raters' rating tendency to seek comprehensibility or consistency of the speech for scoring judgment:

> 93-2. While I give high scores to those that use words/phrases that I am familiar with that doesn't mean that I don't give high scores to those who use words/phrases I don't understand. Instead I seek to understand their meaning. For example, when grading the TOEFL many Indian speakers used the word "freshers" which I didn't understand. I contacted my scoring leader for clarification. The use of such word did not affect my rating [seeking the comprehensibility of the speech]

> 82-2. I am not familiar enough with many varieties of English to judge a speaker of them on correctness. I can, however, often judge on consistency within the sample [seeking consistency of the speech]

Rating Tendency. This section investigated raters' behavior tendency when making scoring judgment. Three items focused on intelligibility of the speech. As reported in Table 24, more than half of the raters (57.3%) did not agree that unfamiliar expressions presented by examinees was indicative of incomplete English learning process (item 7), implying raters' acknowledgment of their unfamiliar expressions as part of repertoire of examinees' variety. Items 3, 5, and 9 asked whether the high scores would be awarded to native like speech if produced by the examinees. Raters' views were mostly liberal and positive indicating the near nativeness was not prerequisite for high scores as long as examinees could get their message crossed.

Table 24

Frequency of Raters' Rating Tendency

		Percentage (%)					
	Items	SD	GD	N	GA	SA	Un-ratable
2.	The differences between standard English and varieties of English are creative and as correct as standard English.	7.23	27.1	18.8	32.3	12.5	2.1
3	Examinees do not need to speak like a native speaker in order for me to assign high scores.	6.2	10.3	9.3	45.4	25.8	3.1
5	I do not grade down examinees that speak a variety, as long as they express themselves well.	0.0	2.1	21.9	50.0	25.0	1.0
6	I do not penalize examinees who use negotiation strategies (e.g. asking for clarification, rephrasing).	2.1	13.5	41.7	39.6	3.1	3.1
7	When examinees use less familiar expressions, it suggests that they have not fully mastered English yet.	16.7	40.6	15.6	19.8	5.2	2.1
8	The rater is not responsible for examinees' intelligibility.	4.2	16.7	31.3	30.2	16.7	1.0
9	I give high scores to examinees that use expressions as used by the native speakers of English.	2.1	5.2	25.0	45.8	20.8	1.0

Qualitative feedback. Eighteen comments were provided, which can be classified into two broad categories: the acknowledgment of non-native speakers of English in the rating and teaching contexts, and the importance of WE awareness:

> Rater/ESL teacher of native speaker of English
>
> 5-4. We have several non-native English speakers teaching ESL at our institution. I think this is a huge asset to our program [positive]

> 12-4. Some native English speakers are excellent ESL teachers, others aren't. The same goes for non-native speakers. In general I think non-native speakers can explain English better for ESL students but I'm not sure that I can say they're overall better speakers, especially because some teachers also speak another language natively [generally positive]
>
> 93-4. Native speakers are not inherently better, all raters and teachers need training [generally neutral]
>
> <u>Development of WE awareness</u>
>
> 67-4. I think anyone involved in field of language teaching/learning, whether they be students or teachers, needs to be aware and sensitized to the different varieties of English and how their existence plays into the general interplay of communication, especially cross-cultural. [crucial to teachers and learners]
>
> 83-4. I think the language learners already have an awareness… it's the native speakers that need to be aware that there are varieties OTHER than their own. [crucial to native speakers]

Interpersonal History. The last section measured the extent to which raters' rating tendency may be influenced by their familiarity with the varieties. Table 25 presents the findings on the amount of rater exposure to the variety of English. Overall, the majority of the raters have exposed to the varieties in their daily life including neighborhood (78.1%) and workplace (96.9%) respectively. More than one third of the raters (41.7%) had experience with the varieties at home environment. A very high percentage in item 2 and 3 (91.6% & 92.6%) shows raters' comfort in listening to varieties and confidence in communicating with speakers of different varieties. Nevertheless, more than half of the raters expressed that the use of the varieties could cause cross-cultural misunderstandings (68.8%). In terms of raters' familiarity with language variations due to global spread of English, a majority of rater (85.4%) agreed that English had evolved into different steady varieties. 63.5% of the raters agreed to the statement that features of varieties were developed in the same way as American English developed from British English.

Table 25

Frequency of Raters' Interpersonal History

		Percentage (%)					
Items		SD	GD	N	GA	SA	Un-ratable
2	Comfortable listening to varieties of English.	1.0	2.1	5.2	40.6	51.0	0.0
3	Can't communicate well with people who speak a variety different from mine.	51.0	41.6	2.1	3.1	2.1	0.0
4	Use of varieties can cause cross-cultural misunderstandings.	3.1	12.5	15.6	50.0	18.8	0.0
5	English has evolved into different steady varieties.	0.0	5.2	7.3	44.8	40.6	2.1
6	Features of varieties are developed in the same way as American English developed from British English.	9.4	7.3	15.6	37.5	26.0	4.2
1	I have chances to speak English with people of different ethnic backgrounds (select all apply) a. In my neighborhood (78.1%) b. At the workplace (96.9%) c. At home (41.7%)						

Establishing a Confirmatory Factor Model for Combined Indicators

The conceptualized three-factor structure of the RAI was further tested to evaluate if a confirmatory factor model could be established. Toward this end, the first issue needed to be solved before the analysis was the unequal number of observations used in the two measures of the RAI, that is, the semantic differential scale and the Likert scale. Although the sample size for measuring rater feeling on the semantic differential scale was adequate (*N*= 576) for factor analysis, it was not the case for measures of rater belief and rating tendency (*N*=20 and 96 in the second and third phase respectively). The recommendations on minimum sample size required for factor analysis vary (MacCallum, Widaman, Zhang, & Hong, 1999; Velicer & Fava, 1998):

the lowest minimum sample size reviewed is 100 (MacCallum *et al*., 1999) and a subjects-to-variables ratio of five (Bryant & Yarnold, 1995 in Garson, 2008). Either way makes the RAI evaluation by factor analysis in the second phase unsuitable, rendering its use in the third phase questionable. This apparently constituted a limitation in providing evidence of construct validity other than the internal consistency estimate. Nevertheless, an attempt was made to use an alternative method to establish the measurement model without compromising the minimum sample size requirement. The three conceptualized components of attitude construct (i.e. feeling, belief and behavior tendency) were treated as latent factors and their sub-components as indicators in place of each individual item as normally analyzed. For example, rater belief (latent factor) includes two indictors: Perceived Cultural Factor and Expectation of Indian English. Thus, the

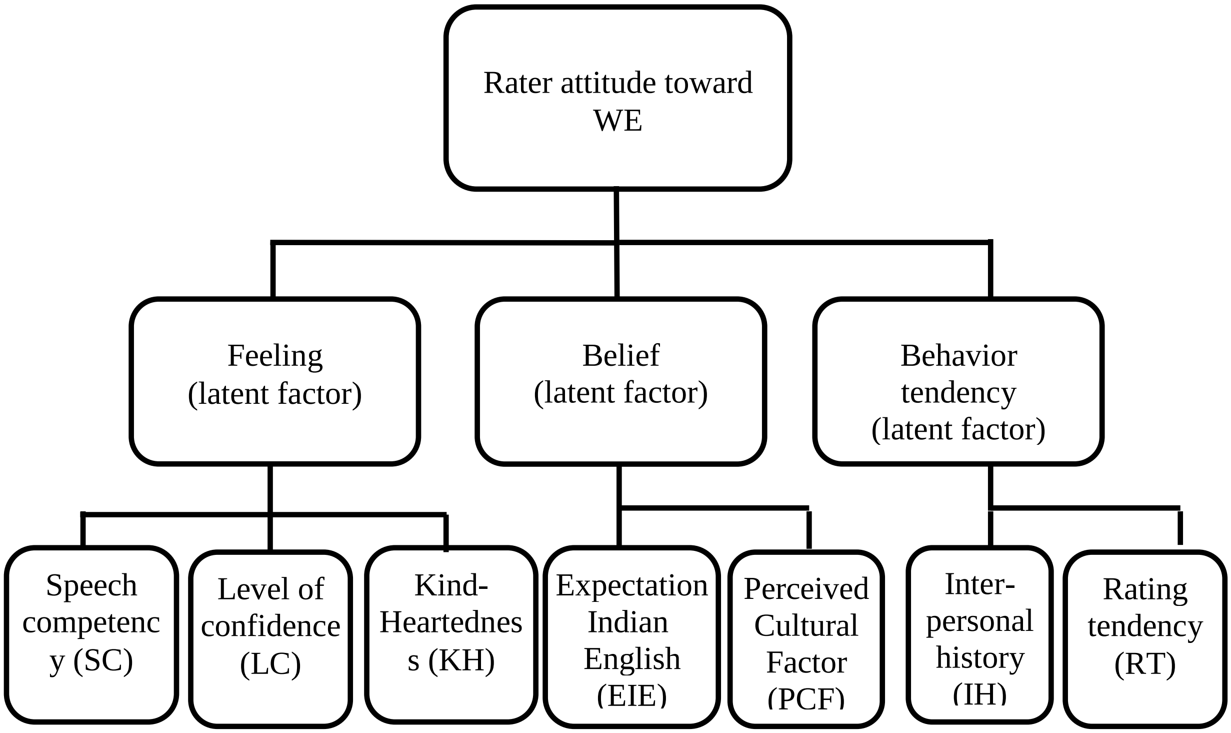


Figure 19. Structure of Measurement Model of three dimensions of attitude construct

conceptualized rater attitude model includes three latent factors and seven indicators, as illustrated in Figure 19.

Instead of examining each item in the indicator, item scores were aggregated to represent a single score for their respective indicator, with higher numbers implying positive attitude toward WE. Thus, the score on “Interpersonal History” for rater 1, summed up the scores this rater assigned on each of the six items making up the indicator. As for rater feeling, the indicator score was arrived at by summing up item scores across the six speech samples rated by each rater on the same criteria. For example, rater 1’s score for “speech competency” was a summation of the scores on the four items (*Articulation, Good Pronunciation, Fluency,* and *Sure*) this rater had assigned to the six examinees. Thus, the indicator score nested six speech samples within a rater. This resulted in seven indicators for 96 observations, which is close to the minimum sample size requirement for running factor analysis.

Prior to performing the CFA, the scores across two measures (i.e. measure 1: rater feeling and measure 2: rater belief and rating tendency) needed to be standardized as they derived from the different scaling methods (i.e. semantic differential scale and Likert scale) and were based on different point scales (e.g. 7 and 5 points respectively). The individual indicator scores were standardized by dividing them with their respective perfect scores and multiplied by 100 to yield the proportional scores. As such, the standardized scores were compared on a like basis. Then, the RAI composite score was calculated. Given the conceptualized tripartite attitude construct, each attitude component, that is, the latent factor, was allocated one third of the total attitude score. The following equation was applied when placing all the components in the same model:

RAI composite score= (SC+LC+KH)*1/3 + (EIE+PCF)*1/3 + (RT+IH)*1/3.

Testing a Three-Factor Measurement Model for Rater Attitude

The initial fit statistics for the 3-factor model did not meet the standards of a well-fitting model. Table 26 shows that χ^2 was 22.311, CFI was 0.934 which exceeded the recommended value. The other two fit indices did not meet the acceptable values: TLI of .894 and RMSEA of .087. Nevertheless, allowing one error covariance (i.e. expectation of Indian English and interpersonal history) to be correlated, the fit indices improved significantly ($\chi^2 = 16.559$ *, p =.167 ,* RMSEA=0.063 , CFI =0.968 , TLI=0.944) leading the data better fit into the model. The factor loading of each indicator on its respective factor was low to high, as illustrated in Figure 20. The factor loading was ranged from 0.257 (i.e. expectation of Indian English) to 0.928 (i.e. level of confidence). The first three indicators (i.e. speech competency, kind-heartedness and level of confidence) loaded strongly on the latent factor, feeling. The factor correlation between belief and rating tendency was the highest (0.925), indicating that raters who have positive belief in World Englishes tended to have positive rating tendency, that is, lenient rating. The factor correlation between feeling and belief was low (0.073) whereas it was negative (-0.019) for the correlation between feeling and rating tendency. The correlation between belief and rating tendency was strong (*r* =0.925), suggesting these two factors may actually be represented by a single factor.

Table 26

Goodness-of-Fit Indices for 3-Factor Measurement Model Before vs After Modification

Model	χ^2	*df*	*p*	RMSEA	CFI	TLI
Before	22.311	13	.000	.087	.934	.894
After	16.559	12	.167	.063	.968	.944

Note. RMSEA=Root Mean-Square Error of Approximation. CFI= Comparative Fit Index. TLI= Tucker-Lewis Index

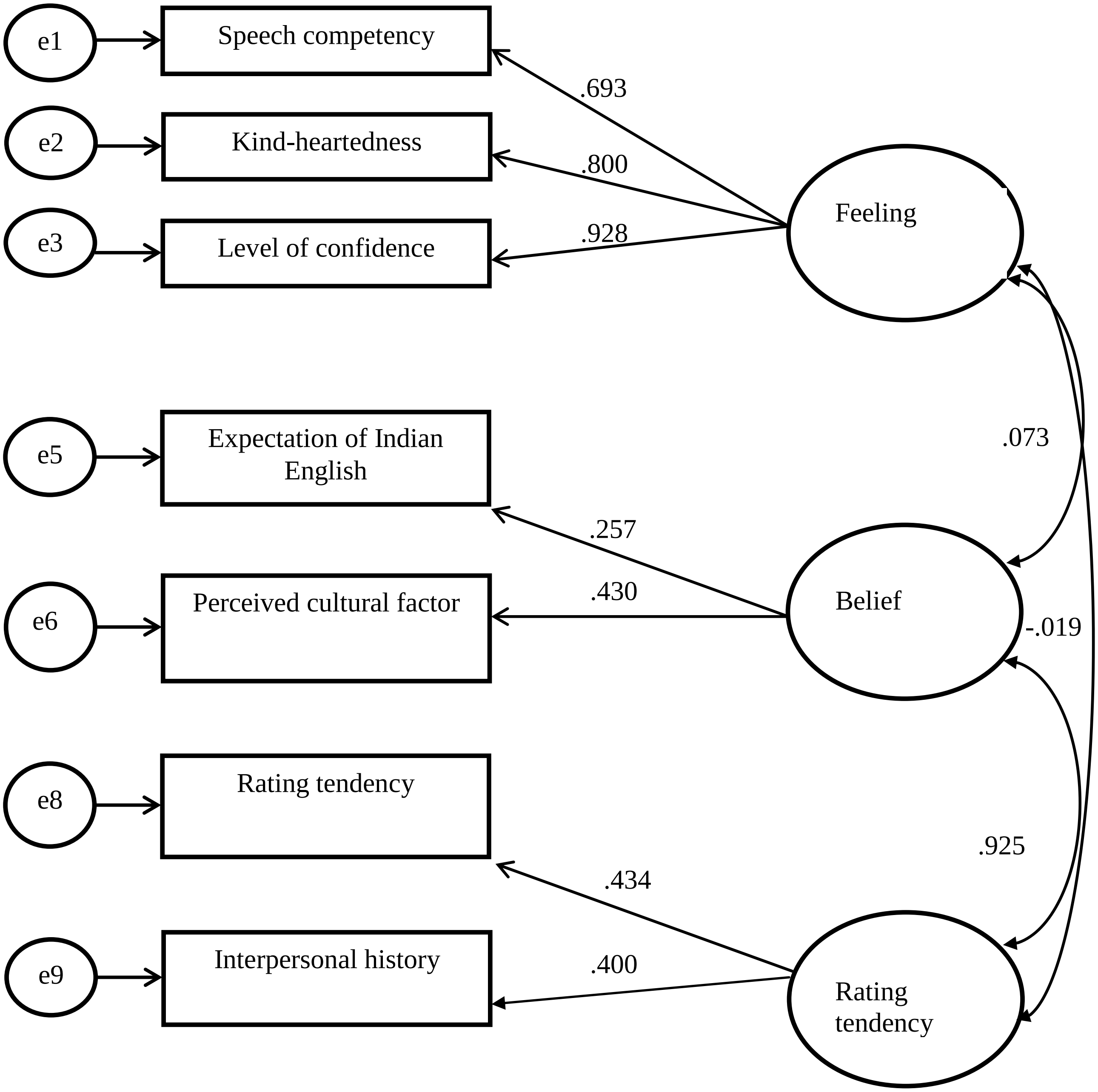


Figure 20. Model 2, the three-factor correlated model for rater attitude

Testing a Two-Factor Measurement Model

Based on the findings of the 3-factor model which implied the redundancy of the factors, the 2-factor model that combined belief and rating tendency was tested. The new latent factor was labeled “belief”. Unlike the 3-factor model, the 2-factor model treated each latent factor equally and applied the following equation when placing all the standardized indicator scores in the same model:

RAI composite scores= (SC +KH+LC)*1/2 + (EIE+PCF+RT+IH)*1/2

The initial fit statistics for the 2-factor model did not meet the standards of a well-fitting model. Table 27 shows that χ^2 was 26.965, CFI was 0.916 which fell below the recommended value. The other two fit indices did not meet the acceptable values either: TLI of .874 and RMSEA of .099. Then the model was modified according to modification index to correlate the error covariance, that is, expectation of Indian English and interpersonal history. The improvement on fit indices was modest, though not significantly ($\chi^2 = 20.052$, $p = .094$, RMSEA=0.076 , CFI =0.954 , TLI=0.926). As shown in Figure 20, the factor loadings of indicators on latent factor of feeling were strong, ranging from .750 (i.e. kind-heartedness) to .932 (level of confidence). Other factor loadings are either low or moderate, with the last indicator, interpersonal history, loading negatively on its respective latent factor. The factor correlation between two latent factors was .164. Figure 21 displays the modified two-factor confirmatory factor model.

Table 27

Goodness-of-Fit Indices for 2-Factor Measurement Model Before vs After Modification

Model	χ^2	*df*	*p*	RMSEA	CFI	TLI
Before	26.965	14	.019	.099	.916	.874
After	20.052	13	.094	.076	.954	.926

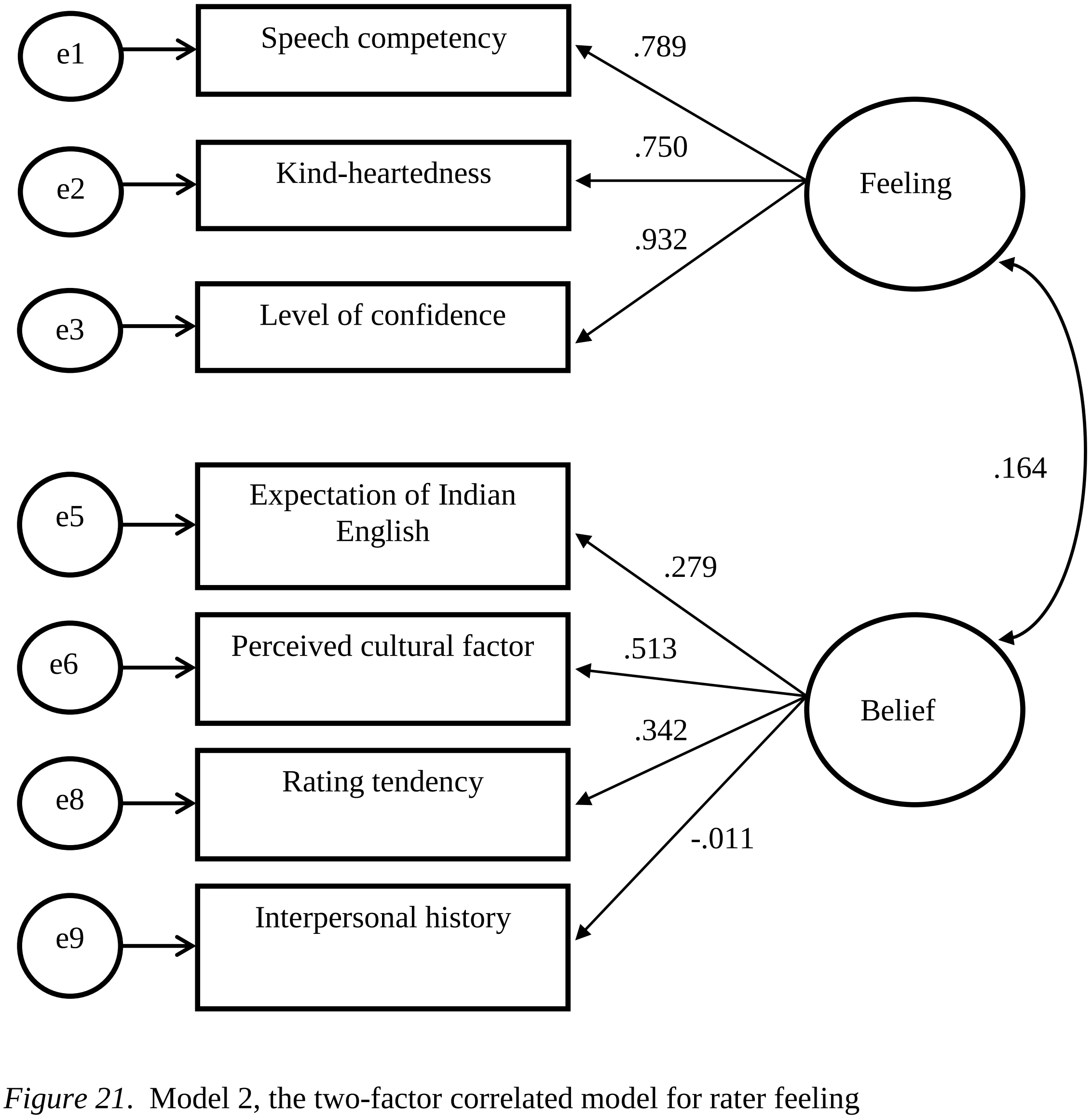


Figure 21. Model 2, the two-factor correlated model for rater feeling

Testing a One-Factor Measurement Model for Rater Attitude

As indicated in the literature (Fishbein & Aizen, 1975) that attitude may be formed by only one component, an attempt was to seek the feasibility of a one-factor measurement model. That is, whether the attitude construct represented by the three conceptualized components is in fact can be expressed by one single factor. Without any model modification, the initial fit statistics for the 1-factor model was not considered a good fit. Table 28 shows the fit indices for the 1-factor model: χ =27.695, p = .016; *CFI* =.904; *TLI* = .855; *RMSEA*=.101. The model was modified to add one pair of correlated error residuals (i.e. rating tendency and expectation of Indian English), which yielded acceptable fit values: χ^2 =17.848, p = .163; *CFI* =.966; *TLI* = .945; *RMSEA*=.063. The factor loading of each indicator after modification on its respective factor was low to high, as illustrated in Figure 22. The factor loading was ranged from -0.197 (i.e. expectation of Indian English) to 0.922 (i.e. level of confidence). The negative loading suggested the increase in the magnitude of rater perception of World Englishes is associated with the decrease in rater's expectation of Indian English and perceived cultural factor respectively. This will need to be further verified by examining the scores in Chapter 5. The two negative factor loadings apparently add the difficulty in interpreting the relationship between the latent factor and indicators.

Table 28

Goodness-of-Fit Indices for 1-Factor Measurement Model Before vs After Modification

Model	χ^2	*df*	*p*	RMSEA	CFI	TLI
Before	27.184	14	.018	.101	.904	.855
After	17.848	13	.163	.063	.966	.945

Note. RMSEA=Root Mean-Square Error of Approximation. CFI= Comparative Fit Index. TLI= Tucker-Lewis Index

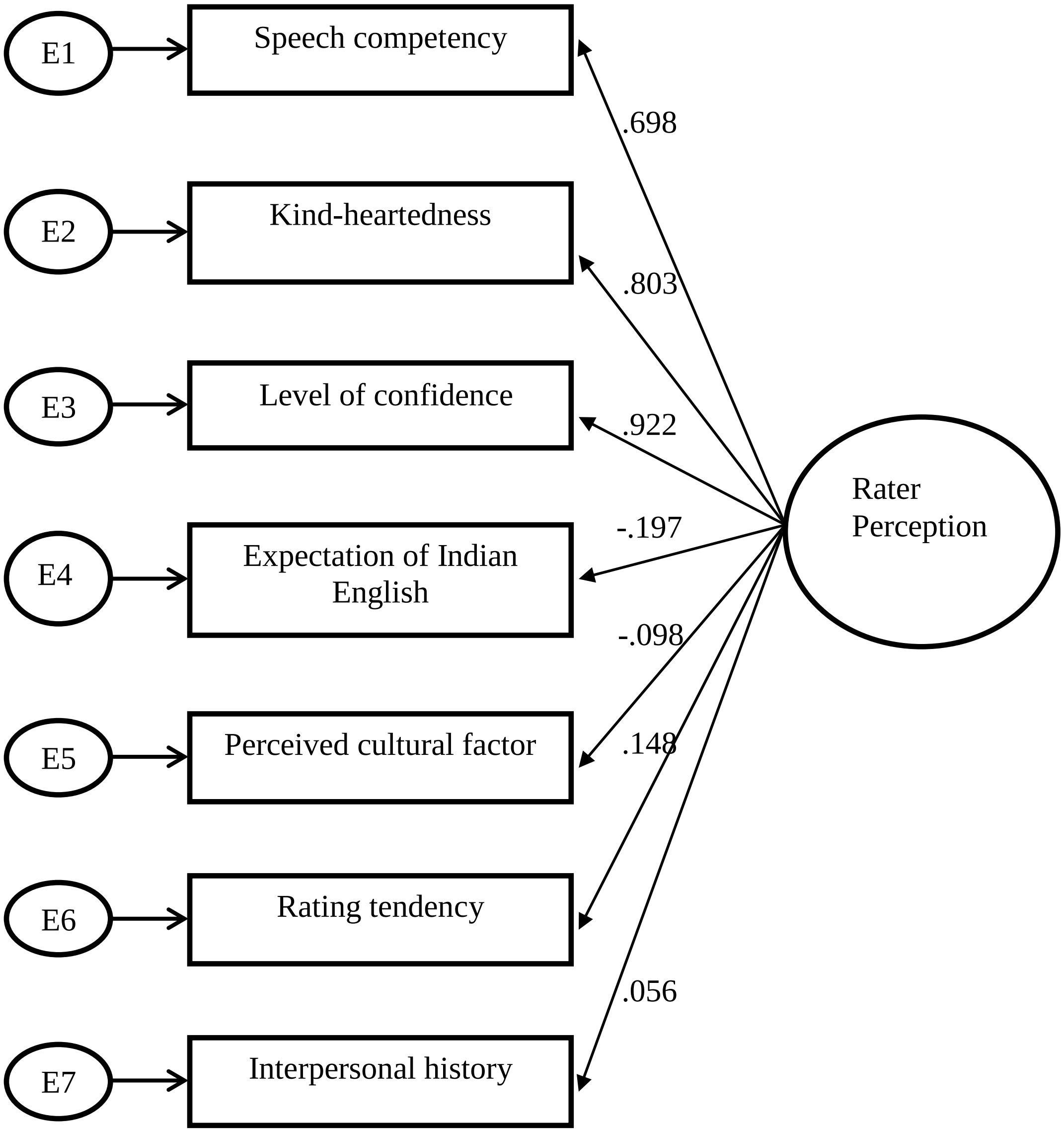


Figure 22. Model 2, the one-factor correlated model for rater attitude

Establishing a 2-Factor Measurement Model of Rater Attitude toward WE

Comparing the three measurement models, all showed good model fit after modification. To determine the best model that fit the data, the χ^2 difference test was conducted between each two models, which yielded critical values on and above 0.062, suggesting that none of the models provides a significantly best fit to the data. Thus, the selection of the best model had to be determined according to interpretability. Looking at the 3-factor measurement model, it established the conceptualized tripartite attitude construct into a confirmatory factor model. All indicators had moderate to strong factor loadings on their respective primary factor; however, the factor correlation between two of the factors, that is, belief and rating tendency, was very high suggesting the overlap of the factors. It indicates the items in these two factors may need to be revised to avoid duplication. In terms of the 1-factor model, the results of analysis support the literature of unified attitude construct. Nevertheless the CFA results suggest that two of the indicators (i.e. Expectation of Indian English and Perceived Cultural Factor) loaded negatively on the latent factor, which may cause interpretation difficulty. As for the 2-factor model, it supports the multi-dimensional attitude construct while avoids the factor redundancy as shown in the 3-factor model. Thus, comparing to three measurement models, the 2-factor measurement model appeared to best represent the constructs of rater attitude toward WE and will thus guide the analysis in the next chapter concerning the effects of rater perception on rating performance. The structure of the 2-factor RAI can be visualized in Figure 23.

Apparently, some of the overlapping questions in the measure of rater belief and rating tendency need to be further improved. Nevertheless, given the data available, the conceptualization of rater attitude toward WE may be best represented by the 2-factor measurement model. With future modifications on the content of the questions in Likert scale, the measurement model may be tested again against 3- and 2- factor model to compare the findings derived from the current study.

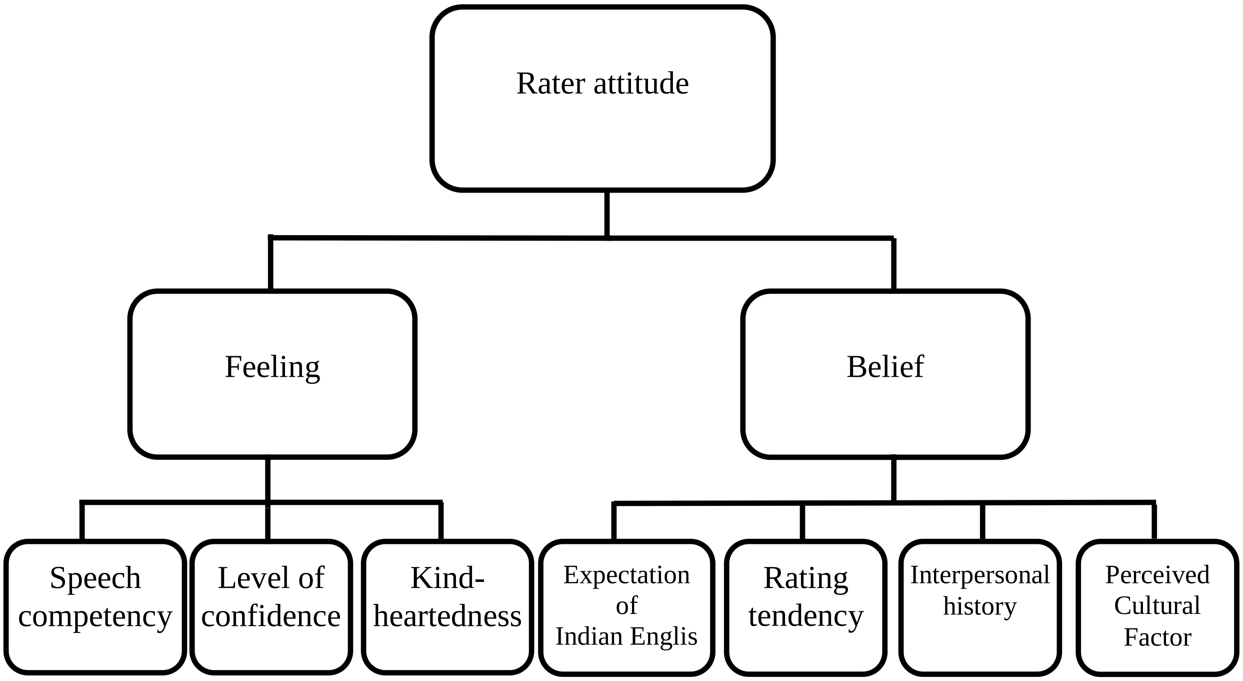


Figure 23. Measurement structure of rater attitude towards WE

CHAPTER 5

RATER ATTITUDE AND RATING TENDENCY

The major claim for study 2 is that: rater attitude towards varieties of English is a biasing factor that influences rater scoring performance on the IELTS descriptive tasks. Five hypotheses serving as warrants (Toulmin, 2003, see chapter 2) were tested using quantitative and qualitative approaches to evaluate the extent to which the claim can be supported. Evidence that supports or rejects each hypothesis will be presented in this chapter.

Hypothesis 1.

Rater Attitude towards World Englishes is Not Consistent and Can be Grouped Into Different Attitude Groups.

Mean Distribution

Initial screenings of the two components of the RAI, rater feeling and rater belief, during the scale construction revealed that the mean scores of each of the components were around or higher than the medium. The mean scores of three extracted factors that represented rater feeling on the 7-point semantic differential scale were 4.97 for Factor 1 (i.e. speech competency), 5.12 for Factor 2 (i.e. kind-heartedness) and 4.70 for Factor 3 (i.e. level of confidence). The mean distributions in the measure of rater belief on the 5-point Likert, excluding four dichotomous items, were 3.84 in Perceived Cultural Factor, followed by 3.78, Interpersonal History, 3.36, Rating Tendency and 3.02, Expectation of Indian English. All the mean scores were higher than the medium score, suggesting raters' attitude towards WE in the current language assessment context seemed to be positive. This was further verified by another statistical tool, FACETS, as discussed below.

FACETS Analysis

Multi-Faceted Rasch Measurement (MRFM) has been accepted over the past decade as a major statistical method of analysis in language performance tests. Specifically, the analysis counts the facets of interest simultaneously when generating the estimation of all facet values, such as rater severity, proficiency level of examinees, and difficulty of rating criteria (Weir, 2005, p.199). In other words, the different facets of interest are all taken into account when constructing the overall measurement picture. In the current study, a two-faceted design was employed, modeling raters and difficulty of RAI components. The latter refers to the seven subscales of the RAI, that is, the three factors representing the rater feeling (i.e. speaking competency, kind-heartedness, and level of confidence) and the four sections for the rater belief (i.e. Perceived Cultural Factor, Expectation of Indian English, Rating Tendency, and Interpersonal Hisotry). The examinee speaking proficiency was the controlled variable and did not factor in the measurement model. The analyses were carried out using the computer program, FACETS (Linacre 1989).

FACETS Summary

Figure 24 provides the relative severity of the raters and difficulty of the seven RAI subscales. The first column is the logit scale, which is the unit of measurement in Rasch analysis and the one in the far right column is the scale used in the scoring. The logit scale is treated as “a true interval scale” (Henning, 1987, p.129), as opposed to raw scores in which the discrepancy between intervals may not be equal (Brown, 1996, p.97). The second column shows the severity variation among raters. A measure of zero represents an average severity for rater performance. A rater who scores most severely, which may indicate negative attitude, is at the top and most lenient, suggesting positive attitude, is at the bottom. The third column shows the difficulty variations among rating categories. The more severely scored category was at the top and the least severely scored category was at the bottom. As noted in the output, the estimates for the raters cluster around the mean on the logit scale, ranging from between -1 and +1 on the logit scale. As for the estimates for the seven RAI subscales, they also cluster around zero with measure of rater feeling more severely scored and measure of rater belief more leniently scored. Note that in Figure 24, the codes appearing in the third column, rating criteria, represent each of the RAI subscales: speaking competency (A1), kind-heartedness (A2), level of confidence (A3), Expectation of Indian English (B1), Rating Tendency (B2), Interpersonal History (B3), and Perceived Cultural Factor (B4). A more detailed record of rater severity and difficulty estimates of RAI subscales are given below.

```
+---------------------------------------------------------------------------------------------+

|Measr|-rater                                                          |-Rating criteria|Scale|
|-----+----------------------------------------------------------------+----------------+-----|
|   2 +                                                                +                +(15) |
|     |                                                                |                |     |
|     |                                                                |                |  13 |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                | --- |
|     |                                                                | A1             |     |
|     |                                                                |                |     |
|     |                                                                |                |  12 |
|   1 +                                                                + A3             +     |
|     | 29                                                             |                |     |
|     |                                                                |                | --- |
|     | 30  55                                                         | A2             |     |
|     | 13                                                             |                |  11 |
|     | 31  70  79                                                     |                |     |
|     | 12  15  17  21  32  33  35  40  46  56  80  91                 |                | --- |
|     | 34  68  73  76                                                 |                |     |
|     | 16  18  2   41  51  63  81  88                                 |                |  10 |
|     | 28  5   53  65  86                                             |                |     |
*   0 * 61  8   90  94                                                 *                *     *
|     | 25  27  36  4   48  59  60  78                                 | B1             | --- |
|     | 11  37  47  49  54  77  89  95                                 |                |     |
|     | 10  14  23  24  26  43  44  45  57  6   66  67  69  74  82  96 |                |  9  |
|     | 20  22  9                                                      |                |     |
|     | 38  52  58  64  84  92                                         |                |     |
|     | 3   50  71  75  85                                             |                | --- |
|     | 39  83  93                                                     | B2             |     |
|     | 1   42                                                         |                |     |
|     | 72  87                                                         | B3             |     |
|  -1 + 62  7                                                          +                +  8  |
|     |                                                                |                |     |
|     | 19                                                             |                |     |
|     |                                                                |                |     |
|     |                                                                | B4             | --- |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |     |
|     |                                                                |                |  7  |
|  -2 +                                                                +                + (5) |
|-----+----------------------------------------------------------------+----------------+-----|
|Measr|-rater                                                          |-Rating criteria|Scale|
+---------------------------------------------------------------------------------------------+
```

Figure 24. FACETS summary (rater severity and category difficulty)

The estimate of rater severity is reported in Table 29. Raters' logit values extend from +.89 (Rater 29) to -1.22 (Rater19), a range of 2.11 logit. The extent to which the 2.11 logit is meaningful can be determined by the following three statistics provided by the FACETS analysis: the separation index, the reliability, and the fixed (all same) chi square, found at the bottom of the table. The separation index is the ratio of the adjusted standard deviation of rater severity estimate (i.e. .31 for this data set) to the root mean-square estimation error (RMSE) (i.e. .30). If the raters were equally or similarly severe, the standard deviation of the rater severity estimate should be equal to or smaller than the RMSE, leading to a separation index of 1.00 or less. The separation index for this data set is 1.06, indicating that rater severity did not vary considerably even though their level of severity was not equal. The reliability statistic produced by the FACETS analysis is different from the traditional sense of inter-rater reliability as the latter refers to the degree of the consistency between raters whereas the former reports the extent to which the analysis reliably distinguishes raters into different levels of severity. If the reliability is high, it means raters are reliably being separated into different levels of severity. The reliability for the current data set was .53, implying that raters may differ and do not share similar levels of rating severity. Lastly, the null hypothesis of the fixed chi-square test is that all the elements of the facet are equal. For the current data set, the chi-square of 197.4 with 95 df is significant at p = 0, indicating that the hypothesis was rejected. In other words, the raters were not equally severe. Based on the values of the three statistics, separation, reliability and fixed chi-square, it suggests the raters' attitude toward WE did not vary considerably but yet the individual differences did exist. As such, it is reasonable to group raters' relative attitude standing into three different groups according to their logit values for further analysis. Raters who had positive logits belong to "negative attitude", negative logits refers to "positive attitude" and zero logt is "neutral attitude".

Table 29

Rater Measurement Report

Rater	Measure logit	Model S.E.	Infit MnSq	Rater	Measure logit	Model S.E.	Infit MnSq
29	.89	.31	2.20	36	-.08	.29	1.17
30	.71	.30	.84	48	-.08	.29	3.15
55	.71	.30	1.08	59	-.08	.29	1.37
13	.62	.30	1.13	60	-.08	.29	.77
31	.53	.30	.38	78	-.08	.29	.39
70	.53	.30	.38	11	-.17	.29	.69
79	.44	.30	.55	37	-.17	.29	1.00
17	.44	.30	.51	47	-.17	.29	1.25
32	.44	.30	.82	49	-.17	.29	.84
40	.44	.30	.77	54	-.17	.29	2.07
56	.44	.30	.20	77	-.17	.29	2.44
80	.44	.30	.35	89	-.17	.29	.48
91	.44	.30	2.15	44	-.25	.29	.89
12	.35	.30	.32	45	-.25	.29	.37
15	.35	.30	.46	57	-.25	.29	.71
21	.35	.30	1.81	66	-.25	.29	.72
33	.35	.30	.29	69	-.25	.29	.43
35	.35	.30	2.55	96	-.25	.29	.68
46	.35	.30	1.13	10	-.34	.29	.65
34	.26	.29	.87	23	-.34	.29	1.34
68	.26	.29	1.24	26	-.34	.29	.97
73	.26	.29	.75	67	-.34	.29	.84
76	.26	.29	.99	74	-.34	.29	1.91
2	.18	.29	2.00	82	-.34	.29	1.75
16	.18	.29	.16	9	-.43	.30	.47
18	.18	.29	.40	20	-.43	.30	1.08
41	.18	.29	2.54	22	-.43	.30	.90
51	.18	.29	1.51	38	-.52	.30	.89
63	.18	.29	.49	52	-.52	.30	.36
81	.18	.29	1.49	58	-.52	.30	1.82
88	.18	.29	1.06	64	-.52	.30	.21
5	.09	.29	.59	84	-.52	.30	.84
28	.09	.29	.23	92	-.52	.30	.45
53	.09	.29	1.68	3	-.60	.30	.76
65	.09	.29	.60	50	-.60	.30	2.00
86	.09	.29	1.28	71	-.60	.30	.69
8	.00	.29	.21	75	-.60	.30	.53
61	.00	.29	.32	85	-.60	.30	.14
90	.00	.29	.58	39	-.69	.30	.76
94	.00	.29	2.73	83	-.69	.30	1.01
4	-.08	.29	.45	93	-.69	.30	1.77
25	-.08	.29	1.06	1	-.78	.30	.08
27	-.08	.29	1.71	42	-.78	.30	.60
95	-.17	.29	1.02	72	-.87	.30	1.18
6	-.25	.29	3.23	87	-.87	.30	.21
14	-.25	.29	.37	62	-.95	.30	.90
24	-.25	.29	1.07	7	-1.04	.30	.38

Table 29 (cont.)

43	-.25	.29	.79	19	-1.22	.30	.83
				Mean	-.11	.30	.99

RMSE = .30, Adj. SD= .31, Separation=1.06, Reliability=.53, Rixed (all same) chi-square = 197.4, d.f =95, Significance=.00

Table 30 shows the group measurement report based on their attitude classification. More than half of the rater (58.3%) is classified as Positive. Column 6 presents the infit mean square statistic. “Fit” refers to the difference between expected and observed scores. The infit mean-square index for the Negative group is 1.00 and that for the Neutral group is 0.96 and Positive group 0.99, smaller than 1. This finding indicates that a slightly more variation is found within the Negative group. However, the infit mean-square indices for the three rater attitude groups all fell within what Weigle (1998) claims to be the acceptable range of 0.5 to 1.5, which suggests the intra-group consistency of all different attitude groups of raters.

Table 30

Attitude Group Measurement Report

Rater attitude	N	Percent	Measurement logit	Model S.E.	Infit MnSq
Negative	36	37.5%	.34	.30	1.00
Neutral	4	4.2%	.00	.29	0.96
Positive	56	58.3%	-.04	.29	0.99
All groups			-.11	.30	0.99

Combining the finding of the statistics provided by FACETS as discussed above, the percentage of the positive attitude group (58.3%) together with mean of the measurement logit (-.11), it suggests that raters in general held positive attitudes toward WE and did not vary considerably, despite the finding that the individual differences did exist. This supports the first hypothesis that raters attitude did not differ dramatically from each other but displayed quite positive attitude toward examinees who speaking Indian English.

Table 31 reports the difficulty estimate for the seven subscales. The first column of the table lists the component and the second column shows difficulty logits, indicating the relative difficulty estimates among the seven subscales. As Table 31 reports, the most leniently scored component was B4 (logit = -1.42), the Perceived Cultural Factors, and the most harshly scored component was A1 (logit =1.27), the Speech Competency, resulting in a span between these two components of 2.69 logts. The logit difference can be interpreted as large because the reliability of separation index was very high (.99) and the chi-square of 892.6 with 6 df was significant at $p < .00$, suggesting that the null hypothesis that all components were equally difficult must be rejected. In other words, significant variation in difficulty did exist among the

Table 31

Difficulty Measurement Report for Seven Components

Criteria	Difficulty(logits)	SE	Infit MnSq
A1	1.27	.09	.74
A2	.74	.08	.91
A3	1.00	.09	.69
B1	-.05	.07	1.54
B4	-1.42	.09	.77
B2	-.66	.07	.85
B3	-.89	.08	1.16
M	.00	.08	.95
SD	1.03	.01	.30

A1=Speech Competency, A2=Kind-heartedness, A3=Level of confidence, B1=Expectation of Indian English, B4=Perceived cultural factor, B2=Rating tendency, B3=Interpersonal history. Reliability=0.99, separation index=12.69, fixed (all same) chi-square=892.6; significance = .00

seven scoring components. Overall, the three components of rater feeling (i.e. A1, A2, and A3) were most difficult, and the the B3, Interpersonal History, was the easiest scored component. The fourth column shows the fit values, which were within acceptable range of 0.5 to 1.5 (Weigle, 1998) for all components, with an exception of B1, the Expectation of Indian English, which slightly fell outside the range.

Hypothesis 2

The Rater Attitude Group Has a Significant Effect on IELTS Descriptive Tasks Scores.

Inter-rater reliability of data from IELTS descriptive task scorings was first calculated. As shown in Table 32, Cronbach's Alpha revealed an acceptable to high level of internal consistency for rater performance (i.e. above .526), except for one case. Alpha for Pronunciation in the neutral group (*N*=4) was negative and low. As noted earlier, response polarity was carefully analyzed and reversed where needed; in addition, the coding was the rating results, possibility about negatively wording issue, as may occur in the survey study, was eliminated. Several possible causes of the negative and low Cronbach's Alpha value include (1) small sample size (*N*=4) in the neutral group and (2) raters' judgment on pronunciation was considerably divergent, leading to the fact that variability of the individual rater exceeds their shared variance (Henson, 2001).

Table 32

Inter-rater reliability

	Positive	Neutral	Negative
Fluency	.825	.526	.930
Pronunciation	.674	-.017	.863
Sentence Structure	.810	.733	.885
Vocabulary	.829	.892	.886

A one-factor multivariate analysis of variance (MANOVA) was performed to determine how the variability in the four ratings can be explained by rater attitude groups.

The three groups of raters are the independent variables and the four rating criteria are dependent variables. Table 33 presents the means and standard deviation of the four dependent variables for the three levels (i.e. positive, neutral and negative) of the independent variables. An examination of these means revealed that the positive group rated the IELTS descriptive tasks higher than the other two groups on all the criteria, except for Pronunciation which was rated highest by the neutral group. The negative group consistently gave the lowest ratings across all the rating criteria, except for Fluency which was rated lowest by the neutral group.

Table 33

Means and Standard Deviations for Proficiency Variables by Three Groups of Raters

	Positive			Neutral			Negative		
	N (nx56)	Mean	SD	N (nx4)	Mean	SD	N (nx36)	Mean	SD
Fluency	336	7.30	1.65	24	6.65	2.12	216	6.03	2.10
Pronunciation	336	6.22	2.14	24	6.29	2.06	216	5.48	2.30
Sentence Structure	336	6.82	1.78	24	6.77	1.72	216	5.85	2.26
Vocabulary	336	7.11	1.84	24	6.81	1.62	216	5.74	2.27

The MANOVA, summarized in

Table **34**, revealed that the main effect for the group variable was significant (lambda=.866). That is, examinee's oral test scores in this study significantly depended upon which group of rater rated their speech. The tests of between-subjects effects showed that rater attitude had statistically significant effect on all the four dependent variables: Fluency (*F* (2, 573) =29.194; p<..0005; partial eta squared=.092), Pronunciation (*F* (2, 573) =8.268; p<.0005; partial eta squared=.028), Sentence Structure (*F* (2, 573) =16.327; p<.0005; partial eta squared=.054) and Vocabulary (*F* (2, 573) =30.918;

Table 34

MANOVA results of Dependent Variables of IELTS Descriptive Tasks

Groups	df (Hypothesis)	Wilk’s Lambda	F	*p*
Positive, Neutral, Negative	8	.866	10.642	.000*

p<.0005; partial eta squared=.097). Post hoc analysis of means using Tukey contrasts was performed to test for mean differences between the positive, neutral and negative group of raters. For all the four rating variables, there were significant differences between the positive and negative groups. Mean scores on Sentence Structure and Vocabulary were also found statistically different between neutral and negative groups. These differences can be visualized by the plots generated by MANOVA, as shown in Figure 25 to Figure 28.

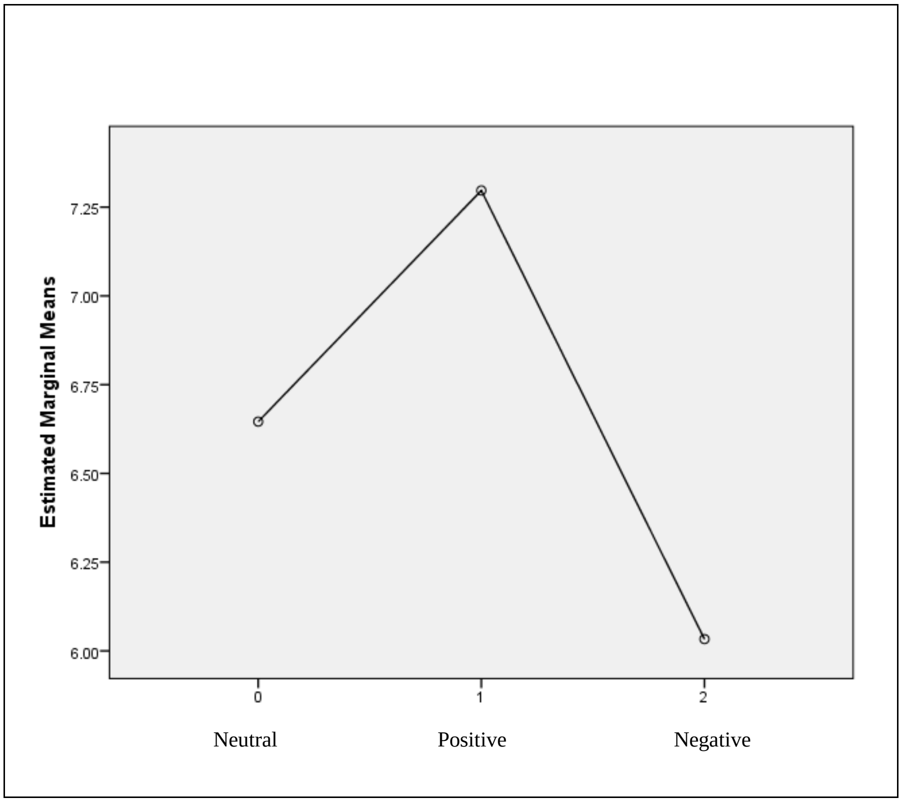


Figure 25. Estimated Marginal Means of Fluency

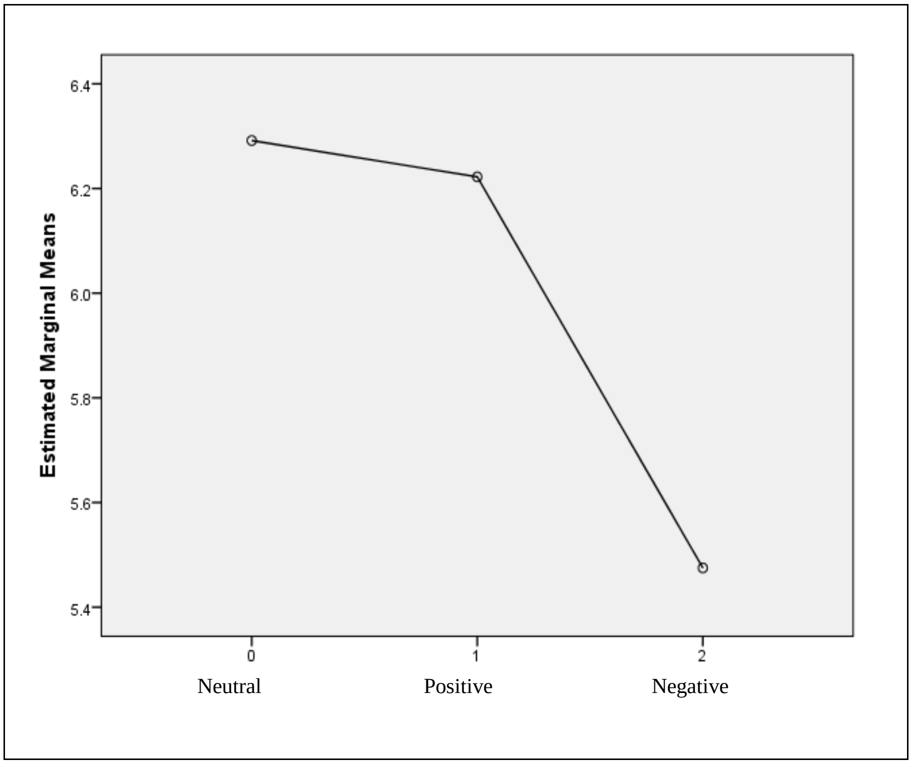


Figure 26. Estimated Marginal Means of Pronunciation

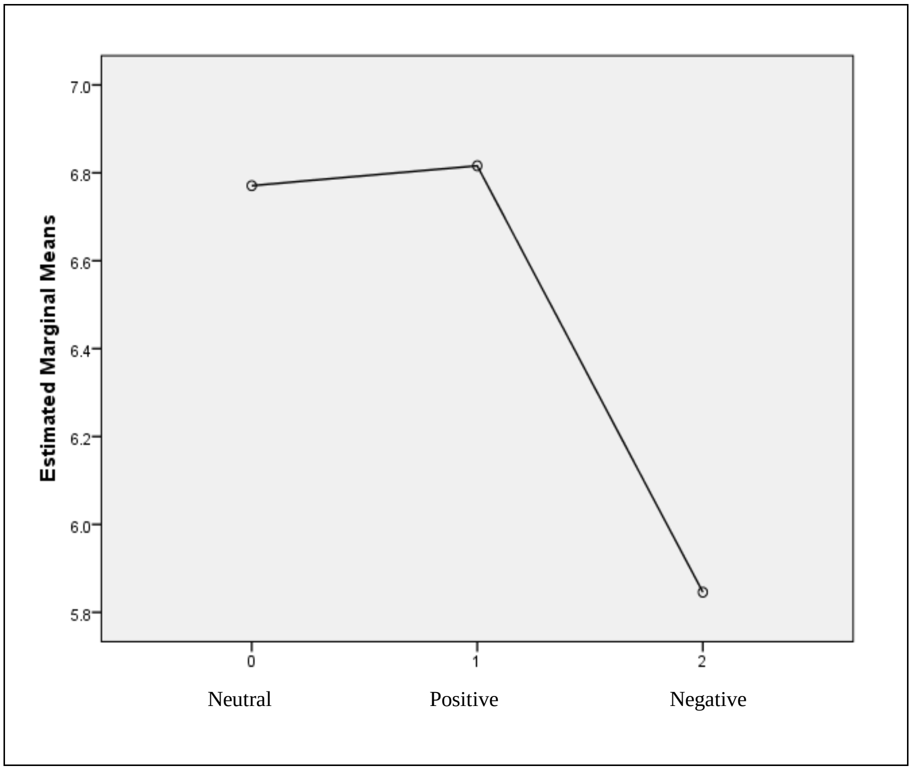


Figure 27. Estimated Marginal Means of Sentence Structure

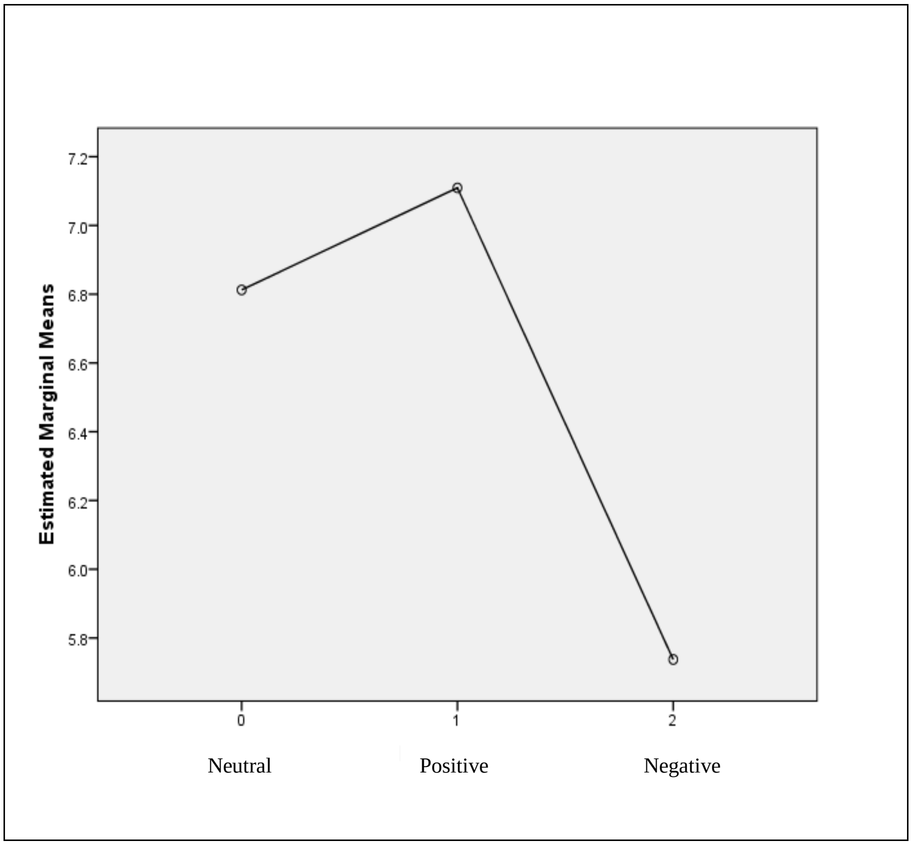


Figure 28. Estimated Marginal Means of Vocabulary

The results of Tukey tests are illustrated in Table 35. It shows that attitude that positive rater group held towards WE had significant mean differences on Fluency, Pronunciation, Sentence Structure and Vocabulary from the other two groups. Raters who had positive attitude toward WE provided higher mean scores than the neutral and negative groups. As for neutral and negative rater groups, the mean scores on Sentence Structure and Vocabulary were significantly different between these two groups. Raters in neutral group gave higher mean scores than the negative group.

Table 35

Tukey Multiple Comparisons of Four Analytic Scores Awarded by Different Attitude Group

Rating Criteria	Attitude group		Mean difference	Std. Error	Sig
Fluency	Positive	Negative	1.26*	.165	.000
Pronunciation	Positive	Negative	.75*	.193	.000
Sentence Structure	Positive	Negative	.97*	.316	.000
	Neutral	Negative	.93*	.316	.010
Vocabulary	Positive	Negative	1.37*	.176	.000
	Neutral	Negative	1.07*	.319	.002

*The mean difference is significant at the .05 level.

To look closer to the differences in raters' ratings on the six IELTS speech samples across the three attitude groups, Table 36 to Table 39 summarizes the means and standard deviation for each speech sample. Note that the speech sample number is indicative of examinee proficiency level, with 1 the lowest and 6 the highest in the current data set. All of the four ratings awarded to the five speech samples were generally consistent with the rank order of the scores obtained in the operational IELTS speaking test[2]. Except for speech sample 4, the higher the examinee's proficiency level (i.e. speech sample number), the higher the ratings across the three groups. In assessing Fluency, all the ratings provided by the positive group were higher than those by the neutral and negative group. Neutral group generally rated higher than negative group; however, the mean score of neutral group for speech sample 1 was lower (*M*= 3.25, *SD*=.854) than that of the negative group (*M*=4.81, *SD*=.298) and speech sample 2 received lower mean score from neutral group (*M*=4.00, *SD*=1.10) than negative group (*M*=5.08,

[2] The operational IELTS speaking test scores are the average scores of the three speaking tasks. This dissertation used only part 2 of the speaking test as stimulus.

SD=.327). The negative group is the only group that has the minimum rating of 0 and the maximum of 9, indicating raters in this group fully utilized the full range of the rating scale.

Table 36

Descriptive Statistics for Fluency

Speech	Positive				Neutral				Negative			
Sample	(n=6*56=336)				(n=6*4=24)				(n=6*36=212)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
1	6.00	.173	3	9	3.25	.854	1	5	4.81	.298	0	8
2	6.43	.219	3	9	4.00	1.10	1	6	5.08	.327	0	8
3	7.48	.189	4	9	6.75	.479	6	8	6.36	.326	0	9
4	6.56	.199	3	9	5.25	.946	4	8	5.03	.353	0	9
5	8.45	.102	6	9	8.25	.479	7	9	7.03	.289	0	9
6	8.65	.093	6	9	8.00	.408	7	9	7.61	.274	0	9

Table 37 reports the mean comparisons for Pronunciation. The highest mean scores for all the speech samples were rated by the positive group, except for sample 3 which oppositely received highest rating (M=7.75, SD=.479) by the neutral group. Similar to the ratings in Fluency, mean scores in neutral group were generally higher than negative group, except for speech sample 1 and 2 where the higher mean scores were assigned by the negative groups.

The mean difference between the three attitude groups on the rating of Sentence Structure is shown in Table 38. The positive group consistently gave higher ratings than the negative group. Nevertheless, the highest mean scores for speech 3 (M=7.75, SD=.479), speech 5 (M=8.25, SD=.479) and 6 (M=8.50, SD=.500) were awarded by the neutral group. For the mean comparison between the neutral and negative group, speech sample 2 received the same

Table 37

Descriptive Statistics for Pronunciation

Speech	Positive				Neutral				Negative			
Sample	(n=6*56=336)				(n=6*4=24)				(n=6*36=212)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
1	6.00	.173	3	9	3.75	1.03	1	6	4.08	.325	1	8
2	6.43	.219	3	9	3.25	1.03	1	5	4.11	.340	0	8
3	7.48	.189	4	9	7.75	.479	7	9	5.81	.313	0	9
4	6.56	.199	3	9	4.75	.629	3	6	4.36	.336	0	8
5	8.45	.102	6	9	7.75	.479	7	9	6.81	.281	2	9
6	8.65	.093	6	9	8.50	.500	7	9	7.47	.294	2	9

Table 38

Descriptive Statistics for Sentence Structure

Speech	Positive				Neutral				Negative			
Sample	(n=6*56=336)				(n=6*4=24)				(n=6*36=212)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
1	5.64	.175	3	8	5.25	.479	4	6	4.58	.348	0	8
2	5.57	.208	2	8	4.50	.500	4	6	4.50	.317	1	8
3	7.27	.177	4	9	7.75	.479	7	9	6.19	.335	0	9
4	5.85	.197	3	9	5.75	.750	4	7	4.44	.373	0	8
5	8.05	.128	5	9	8.25	.479	7	9	7.22	.290	2	9
6	8.45	.098	7	9	8.50	.500	7	9	7.78	.236	3	9

mean scores (M=4.50) from neutral and negative group. The rest of five speech samples received higher mean scores from the neutral group.

Table 39 summarizes the mean difference in rating Vocabulary. The ratings in Vocabulary generally reflected a more stable pattern, that is, the positive group gave the highest ratings of the three attitude groups, followed by the neutral and negative group across all the speech samples. The neutral group generally assigned higher scores than the negative group. Only one exception was speech sample 2 in which a slightly higher mean score was assigned by the negative group (M=5.06, SD=.333) than the neutral group (M=5.00, SD=. 000).

Table 39

Descriptive Statistics for Vocabulary

Speech	Positive				Neutral				Negative			
Sample	(n=6*56=336)				(n=6*4=24)				(n=6*36=212)			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
1	5.64	.175	3	8	5.00	.408	4	6	4.06	.331	0	8
2	5.57	.208	2	8	5.00	0.00	5	5	5.06	.333	1	9
3	7.27	.177	4	9	7.25	.750	6	9	6.22	.382	0	9
4	6.11	.226	3	9	5.75	.479	5	7	4.33	.361	0	8
5	8.08	.135	5	9	8.00	.707	6	9	7.00	.285	1	9
6	8.59	.091	6	9	8.00	.707	6	9	7.53	.216	3	9

Hypothesis 3

Rater Scoring Performance on IELTS Descriptive Tasks can be Predicted by Attitude Tendency

Are RAI scores able to predict raters' IELTS descriptive task scorings? Correlational analysis and multiple regression analysis were performed to address this question. Table 40 reports the correlation between scorings for IELTS descriptive tasks (i.e. total and four sub scores: Fluency, Pronunciation, Sentence Structure and Vocabulary) as dependent variables and RAI scores (total and two part scores) and five rater background characteristics as criterion variables. Note that the rater background variables are dichotomous, including Indian/non-Indian, native language, gender, teaching experience and highest level of education; thus, the point-biserial correlation is used. The IELTS descriptive task total scores and all of the four sub scores were significantly related to the RAI total score and part score 1, the rater feeling, ranging from .418 to .560 (*p*<.01) and .272 to .556 (*p*<.01) respectively. The strength of the association of these correlations can be considered moderate, except for the correlation of Pronunciation and RAI part score 1 (*r* =.272) which was weak. The RAI part score 2, rater belief, was significantly associated with the IELTS descriptive tasks total scores (*r* =.225, *p*<.05) and Pronunciation (*r* =.317, *p*<.01) only. The rest of the sub proficiency ratings (i.e. Fluency, Sentence Structure, and Vocabulary) was not significantly related to the RAI part score 2. Note that the RAI part score 2 was composed of the original measures of rater belief (i.e. Perceived Cultural Factor and Expectation of Indian English) and rating tendency (i.e. Rating Tendency and Interpersonal History). In order to examine the effects of rating tendency alone, the scores of rating tendency were compared with the IELTS descriptive total and sub scores. None of the IELTS scores was significantly associated with the rating tendency. In terms of the five rater background variables, only the Indian/non-Indian variable was significantly related to proficiency total score (*r* = -.252, p<.05), Sentence Structure (*r* = -.329) and Vocabulary (*r* = -

.303). The coding for Indian/non-Indian was 1 for Indian and 0 for non-Indian as it was hypothesized that Indian raters gave higher ratings to the Indian speech samples as used in this study. The negative correlation suggests that low proficiency rating is associated with high group membership; that is, as group membership increases, the proficiency rating decreases. In other words, Indian raters in the current data set gave lower ratings on the IELTS speaking samples than those of non-Indian raters. The rest of the background variables were non-significant: nationality, native language, gender, year of teaching experience and highest level of education.

Table 40

Correlations between IELTS Tasks Scores, Attitude Scores and Background Variables

	IELTS task	FLU	PRON	SS	VOC
Predictors	total scores				
RAI total score	.560**	.534**	.418**	.470**	.569**
RAI part score 1	.498**	.508**	.272**	.422**	.556**
RAI part score 2	.225*	.168	.317**	.177	.159
RAI rating tendency	.206	.125	.233	.236	.177
Indian/non-Indian	-.252*	-.192	-.063	-.329*	-.303*
Native language	.133	.128	.061	.164	.121
Gender	-.073	-.018	-.116	-.041	-.089
Teaching experience	-.128	-.137	.000	-.123	-.180
Education level	.002	-.056	-.021	.089	-.003

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

RAI part score 1=rater feeling, RAI part score 2= rater belief, FLU=fluency, PRON=pronunciation, SS=sentence structure, VOC=vocabulary

To examine how much of the variance of IELTS descriptive task ratings, either total or sub, is accounted for by the RAI total and part scores and rater background variables, regression analyses using stepwise methods were performed. Each regression analysis used one of the IELTS descriptive task scores, either total or one of the four sub scores, as dependent variable. A total of five regression analysis was performed. The results are summarized in Table 41. Variables that do not contribute significantly to variation in IELTS descriptive task scores are not listed.

With regard to the IELTS descriptive tasks total scores, RAI total score was the strongest predictor, accounting for 31.3% of the variance. The status of Indian or non-Indian was also a significant predictor, accounting for an additional 3.2% of the variance.

Table 41

Summary Results of Multiple Regressions for Rater Attitude towards World Englishes and Background Variables Predicting Ratings of IELTS Descriptive Tasks

	R	R2	R2 change	Standardized Beta	F change
IELTS descriptive tasks total score					
RAI total score	.560	.313	.313	.536	42.883
Indian/non-Indian	.587	.345	.032	-.180	4.511
IELTS descriptive task sub scores					
Fluency					
RAI total score	.534	.285	.285	.534	37.469
Pronunciation					
RAI total score	.418	.175	.175	.418	19.946
Sentence Structure					
RAI total score	.470	.221	.221	.433	29.596
Indian/non-Indian	.582	.293	.072	-.271	9.475
Vocabulary					
RAI total score	.569	.324	.324	.538	45.087
Indian/non-Indian	.613	.376	.052	-.230	7.773

By breaking the IELTS descriptive tasks total scores into four sub scores, Table 42 presents the strongest predictor for all the sub scores was the RAI total score. The variance it was accounted for ranged from 17.5% for Pronunciation, 22.1% for Sentence Structure, 28.5% for Fluency, to 32.4% for Vocabulary. The second predictor for the four IELTS descriptive tasks sub scores varied. For Fluency and Pronunciation, no second predictor was found significant at the .05 Alpha level. For Sentence Structure and Vocabulary, the second predictor was both Indian/non-Indian variable, contributed significantly to further 7.2% and 5.2% of the total variance respectively, though their contributions were relatively small.

Given that Indian/non-Indian variable served as significant predictor for three of the ratings, that is, the IELTS descriptive tasks total score, Sentence Structure and Vocabulary scores, independent T tests were conducted to compare which group of raters gave higher means of ratings. Table 42 to Table 44 reports the results of the T-tests. There were significant differences in the scores awarded by Indian and non-Indian raters for all of the three ratings. For the IELTS descriptive tasks total ratings, non-Indian raters (*M*=155.98, *SD*=28.305) gave higher scores than Indian raters (*M*=134.75, *SD*=17.571) did, t (94) =-2.522, *p*=.013. With regard to scores on Sentence Structure where Indian/non-Indian was the second strongest predictor, non-Indian raters (*M*=39.98, *SD*=7.384) gave higher scores than Indian raters (*M*=31.92, *SD*=5.143) did,

t (94) =-3.379, *p*=.001. For the Vocabulary, non-Indian raters (*M*=40.04, *SD*=8.440) also rated higher than Indian raters (*M*=32.33, *SD*=4.979) did, t (94) = -3.079, *p*=.003. In other words, whenever Indian/non-Indian variable was a predictor, non-Indian raters consistently gave higher scores than Indian raters did. It should be recalled that in the current data set, the non-Indian raters all lived in the US at the time they completed the study; the majority was American (N=67), followed by 14 raters with different nationalities, including 4 Chinese, 2 Korean and each of the following: Japanese, Brazilian, Russian, Greek, Malay, Filipino, Pakistan and Nigerian.

Table 42

Results of Independent Sample T-Test for IELTS Descriptive Task Total Score for the Indian/non-Indian Variable

T	Df	Significance
-2.252	94	.013

Table 43

Results of Independent Sample T-Test for Sentence Structure for the Indian/non-Indian Variable

T	Df	Significance
-3.379	94	.001

Table 44

Results of Independent Sample T-Test for Vocabulary for the Indian/non-Indian Variable

T	Df	Significance
-3.079	94	.003

Hypothesis 4

Rater Attitude is Associated with Rater Background Characteristics

As rater attitude towards WE was associated with raters' scoring tendency on IELTS descriptive tasks and served as a moderate predictor of IELTS descriptive ratings as found in Hypothesis 3, Hypothesis 4 is to further test if RAI scores can be predicted by rater background characteristics. Table 45 presents the results of correlational analysis. As seen in Table 45, only RAI part score 1 (i.e. rater feeling) was significantly related to the Indian/non-Indian variable (*r* = -.231, *p* <.05). The negative correlation revealed that the higher rating to the RAI part score 1 was associated with non-Indian raters when they were coded 0. RAI total score was not significantly related to any rater background variables, and neither was the RAI part score 2.

Table 45

Correlations between Rater Attitude Instrument Scores and Rater Background Variables

Predictors	RAI total score	RAI part 1 score	RAI part 2 score
NS/ NNS of India	-.134	-.231*	.123
Native language	.022	.020	.010
Gender	-.148	-.109	-.086
Teaching experience	-.057	-.129	.084
Education level	.014	-.089	.140

RAI part score 1=rater feeling, RAI part score 2= rater belief

To identify if the Indian/non-Indian variable is the possible determinants of the RAI part score1, a regression analysis with enter method was performed. As evident from Table 46, Indian/non-Indian variable significantly predicted the RAI part 1 scores. However, the R squared of the estimation was low (0.047), indicating that only 4.7% of the total variance in RAI part 1 score was accounted for by the Indian/non-Indian variable.

Table 46

Summary Results of Regressions Analysis for Indian/non-Indian Variable Predicting Rater Attitude Instrument Part 1 Score

Model	Unstandardized Coefficient (B)	Std. Error	T-statistic	Significance of T-statistic
Constant	24.988	.328	76.295	.000
India	-1.988	.926	-2.146	.034
R squared: .047				

To compare which group of raters gave higher means of rating, an independent sample T test was conducted. Table 47 reports the results of the T-test. There was a significant difference in the scores given by Indian raters (*M*=24.89, *SD*=3.00) and non-Indian raters (*M*=22.74, *SD*=2.44); *t* (94) =-2.146, *p* = 0.034. Specifically, raters of non-Indian gave higher scores on RAI part score 1 than native speaker of Indian did.

Table 47

Results of Independent Sample T-Test for Rater Attitude Instrument Part 1 Score for Indian/non-Indian Variables

T	df	Significance
-2.146	94	.034

Hypothesis 5

Rater with Like Attitudes May Score the IELTS Descriptive Tasks In a Similar Fashion by Weighing Particular Salient Features of Indian English More Heavily Than Others.

The richness of WE are defined in the various literatures. As noted in chapter 2, categories of language commonly discussed include phonology, syntax, vocabulary, pragmatic, communication and literature styles (Mesthire & Bhatt, 2008; Y. Kachru, 2005). Hypothesis 5 explores which of the above categories are applicable in the monologue descriptive tasks in the oral testing context and the extent to which raters with different attitude differ in the varietal features that they focus on when judging the tasks. A verbal protocol study was used and findings were compared to studies that use the same methodology for different oral task types, as discussed below.

Samples

Five of the six IELTS descriptive tasks used in study 1 and 2 were used again to elicit rater attitude and scoring performance. Speech sample 4 had somewhat poor sound quality and was not selected. The scores of the five IELTS descriptive tasks used in the rater cognition study are bands 4,5,7,8 and 9.

Raters

To look into the various dimensions of varietal features that influence rater judgment in relation to their attitudes towards WE, different combinations of rater attitude and rating tendencies were used. Eight raters were selected based on their relative severity of ratings on the two tasks: the RAI and IELTS descriptive tasks. Raters' relative severity in ratings to the RAI was analyzed by FACETS analysis as reported in Hypothesis 1. The same method of analysis was used to check raters' scoring judgment of IELTS samples. The outputs of the FACETS analysis modeling two facets (i.e., rater and rating criteria) for the IELTS speaking samples are displayed in Table 48. The selection targeted raters placed in the four different relative standings when aligning rater relative severity of two rating tasks. That is, raters of the following four combinations of raters were selected:

Table 48

Rater Severity on Rater Attitude Instrument Scores and IELTS Descriptive Task Ratings by FACETS Analysis

	Logit	RAI scoring	IELTS speech scoring
Score low	3		
	2		21
	1	29 30 55 13 31 70 79 12 15 17 21 32 33 35 40 46 56 80 91 34 68 73 76 16 18 2 41 51 63 81 88 28 5 53 65 86	29 15 13,33
	0	61 8 90 94 25 27 36 4 48 59 60 78 11 37 47 49 54 77 89 95 10 14 23 24 26 43 44 45 57 6 66 67 69 74 82 96 20 22 9 38 52 58 64 84 92 3 50 71 75 85 39 83 93 1 42 72 87	63 20 30 67 55 56 73 74 76 31 4 51 77 17 91 32 46 79 90 94
	-1	62 7 19	10 18 23 26 34 40 61 70 80 81 89 11 35 43 66 69 85 88 93 36 37 45 47 72 87 9 96 12 16 28 44 50 59 62 64 71 75 8 39 42 49 78 84 86 24 38 41 48 65 68 82 1 22 3 5 53 95
	-2		54 27 57 58 83 19 2 25 7 92 14 52
Score high	-3		6 60

Combination 1: positive WE attitude and high IELTS sample scoring

Combination 2: positive WE attitude and low IELTS sample scoring

Combination 3: negative WE attitude and low IELTS sample scoring

Combination 4: negative WE attitude and high IELTS sample scoring

The actual selection first checked raters' agreement for participation in the qualitative study as indicated during the RAI study. This therefore limited the selection of raters meeting the criteria above and lessened raters' relative severity of ratings. For example, two raters (4 and 77) in the positive yet near neutral attitude group were selected to represent combinations 3 and 2 respectively. Raters representing four varying levels of attitude tendency towards WE and severity of IELTS descriptive task ratings are shown in Figure 29.

		IELTS rating	
		Low	High
Attitude	Positive	23, 77	01, 54
	Negative	04	27, 48, 53

Figure 29. Raters selected in the verbal protocol study

Note that raters 77 and 23 are Indian and Brazilian respectively. This Brazilian rater has lived in the U.S. for 12 years at the time of the study. The rest of the raters are American. Among the eight selected raters, raters 77, 48 and 53 are accredited and experienced IELTS raters. Each interview lasted approximately an hour and was conducted online using Skype.

Collection and transcription of verbal reports

Prior to the verbal protocol study, the raters received a consent form (see Appendix F) and the five IELTS descriptive tasks to test the sound quality. Following Ducasse and Brown (2009), the raters were requested to perform two tasks during the study. First, they listened to an entire IELTS descriptive task without stopping and provided an overall impression of the speaker. The raters were told to focus specifically on features that they thought were Indian English, such as what was said and how it was said. Second, they had to pause and comment on the speech when they heard features they thought belonged to Indian English or something significant or noticeable that influenced their rating. The raters were given a practice run before the study to ensure they understood the instructions completely. It should be noted that the verbal report was produced individually with no prior discussion on what was meant by varieties of English. This is an important factor in obtaining unguided observations (xx).

The raters repeated the two steps for the five IELTS descriptive tasks. The 40 verbal reports (eight raters on five speech samples) were transcribed orthographically.

Analysis of the verbal report data

Each report was divided into units by the current researcher. Each unit focused on “a single event or task” (Green, 1998 p.19) or an “idea” (Ducasse & Brown, 2009), that excluded further elaborations, examples, or justifications. Next, each unit was read to search for rater orientation on the aspects of variety that influenced rater judgments. Then, another researcher coded the speech according to the seven categories that the current researcher that were observed to dominate the data. The seven categories were *degree to which the speech was native-like*, *pronunciation*, including intonation and proper pauses, *grammar,* including word choices, *comprehensibility, listener effort*, including clarity, *level of second language schooling*, and *fluency.* According to Hatch and Lazarton (1991), the inter-coder agreement was derived from the number of agreements as a proportion of the total number of codings. This resulted in an agreement of 70.53%. The disagreements were mostly on the high frequency of comments on the vocabulary use associated with Indian English and the extent to which that affected the listener's comprehension. The categories were re-organized after the discussion with the coder. After an iterative process of coding and discussion, the new set of categories were finalized comprising four linguistic performances, that is, *vocabulary*, *grammar*, *pronunciation* including intonation, stress and accent , *fluency*, and three non-linguistic performances of *comprehensibility* that include clarity of the speech, listener effort and organization, *level of language learning*, and *degree of near nativeness*. The first three categories form part of *language use* as claimed in the second language speaking construct (Fulcher, 2003). The inter-coder reliability was re-calculated which increased to 78.21%. As pointed out by Gass and Mackey (2001), inter-coder reliability is often close to 80 percent, suggesting that the new inter-coder reliability fell within the acceptable range. Each of the categories is discussed below with examples taken from the verbal-protocol transcriptions.

Validity of the protocols

Following Brown (2000), a validity check was made on the verbal protocol data to evaluate the representation of raters' actual scoring flow and judgment. The validity was based on the assumption that positive comments would be increasingly elicited as the scores got higher. The ranking of the mean score for each sample was compared with the proportion of the positive to negative comments. The distribution of positive and negative comments as reported in Table 49 reflected the rankings in comparison with the mean score. It can therefore be justifiably concluded that the comments represent rater's scoring processes.

Table 49

Protocol Rankings

Sample	Total no. of evaluative comments	Negative comments	%	Positive comments	%	Ranking	Mean[3] score	Score ranking
56	55	5	9	50	91	*1*	8.04	*1*
35	38	8	21	30	79	*2*	7.21	*2*
13	81	33	41	48	59	*3*	5.18	*4*
41	61	35	57	26	43	*4*	6.11	*3*
22	76	68	89	8	11	*5*	4.43	*5*

Comments by Category

As shown in Table 50, the first three large proportions of the comments all concern linguistic performance in vocabulary (22.3%), grammar, (19.7%) and pronunciation (18.8%), with overwhelming majority being negative. The major focus on the linguistic performance supports similar studies seeking rater orientations on the different tasks of the speaking tests through stimulated verbal recall (Brown, 2000; & Brown *et al*. 2005; May, 2010; Orr, 2002). The linguistic performance was the mostly frequent commented on, which may be explained by their salience and being first taught in ESL or EFL class, thus heavily drawing the raters' attention. The last four categories are non-linguistic performance: comprehension (16.9%), near-nativeness (10.5%), fluency (6.4%), and lastly proficiency level of language learning (5.4%). While comments in comprehension and fluency were mostly negative (62% and 80%), it is interesting to note that most comments in near-nativeness and proficiency level of language learning were mainly positive (73% and 95%). These two categories were less reported in the

[3] Average of the ratings awarded by the total of 96 raters in the study.

similar studies mentioned above and will draw examples of raters' comment as each category is discussed in turn.

Table 50

Comments by Category

	Voc-abulary	Grammar	Phonology	Compre-hension	Native like	Fluency	Language learner
N	70	62	59	53	33	20	19
%	22.3	19.7	18.8	16.9	10.5	6.4	5.4
Polarity	P N	P N	P N	P N	P N	P N	P N
N	21 49	18 44	21 38	25 33	24 9	4 16	18 1
%	30 70	29 71	36 64	47 62	73 27	20 80	95 5

The categories were compared with those selected in the RAI. Table 51 reveals that raters as a group (N=96) when responding to the RAI selected pronunciation as the most distinctive feature of a variety, followed by vocabulary use (13.96%), communication style (13.33%), sentence structure (12.08%), and pragmatic use (12.08%). The overlapping categories in the verbal protocol study and RAI included pronunciation, vocabulary use, and sentence structure. Two categories selected in the RAI, communication style and pragmatic use, were not mentioned by the raters in the verbal protocol study, possibly due to the descriptive tasks as the elicitation stimulus made these two categories less relevant. In general, pronunciation was the most salient category that constituted WE from the RAI results as compared to vocabulary use in the verbal protocol study. The latter most likely resulted from a speech sample that frequently used vocabulary unique to Indian English, which led to more comments on this category.

Table 51

Comparison of Categories Constituting World Englishes as Perceived by Raters in Two Studies

Verbal-protocol study	%	RAI	%
Vocabulary use	22.3	Pronunciation	18.33
Grammar	19.7	Vocabulary use	13.96
Phonology	18.8	Communication style	13.33
Comprehension	16.9	Sentence structure	12.08
Near nativeness	10.5	Pragmatic use	12.08
Fluency	6.4	n/a	
Level of language learning	5.4		

Category 1. Vocabulary

Comments about vocabulary were predominately negative and concerned its limited range and inappropriate use hindering comprehensibility of the speech. Comments were positive about lexical maturity and sophistication, which the raters considered to be evidence of a high level of English proficiency. While some comments made reference to examinees' overall lexical ability, others noted the use of specific lexical phrases. Given vocabulary's salience of signifying a variety as noted in the WE literature and a particular speech sample in this study using several vocabularies impeding speaker's intended messages, this is probably why vocabulary is the most commented category. When raters were less familiar with the vocabulary used by the examinees, it meant that speech comprehension was affected and greater listening effort was needed to grasp what they intended to say, thus leading to negative comments. The following are examples of comments in the vocabulary category:

negative comments on specific words or phrases:

54-12 Her choices of words were sometimes confusing. For example, you used "discover", in place of word, "explain", or perhaps "discuss", and the first time she said it, I couldn't make out the word.

1-8 So "locality". So he's trying to find a vocabulary that could fit. It may be translation from home, his language, vocabulary maybe, so it shows variety of English but also student he tries to find word that fit . He said "locality in the area", I think he wanted to say we play football locally, but not "in the locality" that doesn't fit what he was trying to say. He picked the word that doesn't fit and he going back to context [sic].

53-14 "means of football match" that seems unnatural, I don't know, it's kinda non-native phrasing.

positive comments on the strategy of use of words:

1-20 She knows "beard" but she said "under his nose". She's not sure the vocabulary and wants to clarify, which is a good language skill to clarify, but she doesn't need to.

27-15 "It was an inter-school competition", she could've said "contest". She was very clear of what some of the words that she used.

general comments on vocabulary range or usage:

77-21 His vocabulary is just okay for the topic at hands.

48-11 My overall impression [is that] the person is educated, speaks clearly most of the time, i can understand what he is say. He uses high vocabulary.

Negative comments on vocabulary (49) outweighed positive ones (21), and the three speech samples (speech 3, 2, and 1) in particular received a large proportion of negative comments, which reflect their lower level of proficiency among the six speech samples.

Category 2. Grammar

Positive comments on syntax were general, concerning examinees' overall good command of syntax whereas the negative comments were on specific aspects of syntactic use, including the typical structure of Indian English. The negative comments overwhelmingly concerned verbs and, in particular, tenses (17 comments), where the use of present continuous, present perfect, and past perfect each had three comments, with the latter two being commonly used in Indian English (Meierkord, 2004). Others included word order (6), incorrect use of indirect sentences (5), articles (5), pronouns (2), and plural, subject and verb agreement, preposition and substitution of noun with adjective, noun clause and objective pronoun (1 each). Examples of comments about syntax include:

> 1-5 The language structure is quite basic; she made lots of mistake with different verb tense. She tries present perfect a lot to get her points across a lot. She's always searched for structure, not naturally. [generally negative]
>
> 23-2 With the indirect speech, what her friend told her, She actually repeats the words of her friends, instead of using the grammar form of indirect speech. She said "I had to buy" , " I should've bought; I should've been buying" that shift in time from being indirect speech grammar, she doesn't state here. [specific negative]
>
> 48-10 There were some sentences where the word order was incorrect and I could not understand what she was saying. [specific negative]
>
> 23-5 Again, the way he constructed the sentences, it may not be very well with time, but the logics are there. It's grammatically correct. [generally positive]

Category 3. Phonology

Four aspects of phonology, that is, pronunciation, intonation, speed of the speech and accent, received comments from the raters. Raters commented on this category particularly when the pronunciation hampered comprehensibility or when they noted sounds distinctive to Indian speakers. Examples of each of the types are discussed below.

Phonemes

Raters commented on consonants that were un-aspirated and pointed out several pairs of phonemes unique to Indian English pronunciation, including the mix of *v* and *w*, *t* and *d, o* and *x*, *a* and *o,* and *t* and *th.* Twenty of the 59 comments regarding phonology were made specifically on this category.

> 54-13 Now her ‘w’ sounds more like ‘v’. In the word ‘way’, sound more like ‘vay’. I think the ‘v’ and ‘w’ are perhaps one sound in her native language.

> 53-21 Her pronunciation had a few little…what I call Indian English feature, such as instead of ‘th’, I think the word is “ three” but she says “tree”. But in the context, it was clear that she meant the number.

> 77-3 For the most Indian speaker of English, his “ oo” sound is articulated as the voiced labio-dental “ v “; this is the mistake made by many Indians.

Intonation

Given that English is a stress-timed language, the lack of the stress and intonation elicited negative reactions where 18 of the 59 comments were made on this category.

> 53-10 We [at the rater training] talk about intonation and stress. It sounds (the speech) kinda all run together. It doesn’t seem to stress on certain words that would help listeners to understand better. Nouns, and verbs, as opposed to everything seem to be equally stressed, and the rhythm. It’s just hard to listen to.

Speed of the speech

Predominantly, all raters made negative comments (10) on one speech that was spoken very fast and lacked stops between main ideas.

> 1-11 Non-stop without pauses. Native speaker would do message unit chunk because no natural pause so it’s very difficult to follow.

> 77-5 Most of the vowels and consonants sounds are articulated incorrectly, she speaks too fast at pace; there is no natural pauses, and she’s absolutely not concept of English being a timed vowel language. She seems to be in a hurry to complete her speech, as the cases with lots of Indian candidates who sit for the IELTS exams. It’s like she’s pouting.[?] English without borrowing, knowing any features of spoken language that should be kept in mind. Unfortunately this is the feature with many Indian speakers who may have not studied in good English medium school.

Accent

Accent is obviously a salient linguistic feature to identify a variety; nevertheless, it was rarely a stand-alone deciding factor of comprehensibility unless the speech was accompanied by other linguistic uses that cause intelligibility issues. The following demonstrates the comprehension problem due to accent along with other linguistic uses:

> 27-3 The content in some of the words she's choosing, she didn't need to use the word, "that", in that sentence, kinda redundant; double negative, wrong tenses, on top of that, the accent, and on top of it, quickly, she speaks quickly and she got thick accent and she's not using the language properly.

This is the only instance where a rater elaborated on the effect of accent on less efficient language use. Other comments (5) on accent were general, simply relating the accent to Indian speakers. Only one positive comment was made regarding a speaker's intention to duplicate the Northeastern U.S. accent.

Category 4. Comprehension

Comments can be categorized into three types: listener effort, clarity of the speech, and organization. Available data suggest that the causes of the comprehensibility issues vary as they could arise from less frequent use of single linguistic features, a combination of linguistic uses, or in the organization of the speech, all of which seem to lead to different degrees of listener effort in comprehending the speech. Nevertheless, despite these causes, speech may still be comprehensible mainly through listener effort, and so the exact causes of incomprehensibility is difficult to define. The following are explicit or implicit illustrations of the latter viewpoints.

Listener effort (23 comments)

> 53-5 So "her wife and children passed away" and it sounds like she "design from his job and return to his native". So I think I can figure out what she says but *it takes lots of effort* and I have to fill in there to figure out. [Italics added] [explicit indication of listener effort]

77-2 The speaker can be understood throughout and the description was quite clear. However, his accent and pronunciation has distinctive colloquial attached to it. Most of the time she addressed in the incorrect syllable, but he can be understood throughout. He's quite fluent, but he can't sustain his fluency as he drives up too soon, and has tendency to repeat himself. His vocabulary is just okay for the topic at hands. On the whole, one feels that he speaks as he has to say something but he doesn't have much to say. He does use complex structure, but his tenses are not sustained, and use of article is incorrect at times, on the whole, he is an average user of spoken English. [implicit indication of listener effort]

Clarity of the speech (18 comments)

1-13 Her story was clear but I couldn't follow the details; but her accuracy and pronunciation less accurate, it's more like a story a child would say. [generally negative]

54-2 Her "seven stand" or "second stance", not exactly sure what story she was trying to tell. It wouldn't make you think what she's talking about. It just I wasn't clear of that one part, but it doesn't take away the message. [generally positive]

27-16 She explained "you'll be giving 15 minutes", "you had to think about the topics", "you had to speak about pros and cons". Very clear, very precise. [specific positive]

Organization (12 comments)

48-15 It's not very articulate. If you ask me summarize what he said, I would have trouble just because it's not very organized. I don't think. But overall, I mostly understood what he said. [specific negative]

27-10 She was very clear about what she was saying; her thought was clear; her thought was very linear; she knew what the question was asked; she explained what she would gonna do. If I didn't know any better, I would've thought that she was reading an essay. It was very clear, very precise; that's how I teach my students: I do an outline, and you phrase supporting, supporting details, main ideas, separate by paragraphs. She was easy to follow, good grammar, vocabulary, spoke slowly. [general positive]

Category 5. Degree of near nativeness

Comments by raters seemed to indicate that their assessment of the speech was slanted against the native speaker of English. They included two broad levels, on the local level of the linguistic use, such as pronunciation and word choice, and at the global level of the speech evaluation, such as organization and overall evaluation of the speech. Of the 33 total comments, 24 were positive where raters appraised examinees' language use that reflected that of native speakers. Raters' reflection of examinees' proximity of native speaker of English on the local level:

> 27-5 " local guy" would be more like native speaker, instead of playing with my friends, playing with my family; I'd " play with local guy" that would be more native like.

> 04-9 A couple of instances she said "I very well remember". It's not wrong, but English people would say "I remember very well". But the word order is off sometimes. But it's not extent to which the listeners wouldn't understand. But it's clear to tell you that she's not native speaker of English.

the global level:

> 1-7 "You start my neighbor my home" so it doesn't start clearly, but the native speaker starts clearly to sets the scene of the story.

> 27-12 She sounds like she's an advanced speaker, but I wouldn't say she's professional and she's not native like.

It is interesting that the evaluation against the native speaker is based on two different perspectives. One seems to hold a lenient attitude and views the discrepancy, as compared to the language usage by native speakers and examinees, as simply different without negative indication of inferiority of the variety. Conversely, the second imposes rater interpretation of the difference and treats the inconsistency between native speakers of English and Indian English as mistakes. The two contrasting views were the concern of scoring validity as raters' different interpretations may reflect their rating decisions (Davies *et al*, 2003). The following examples demonstrate the two viewpoints on the discrepancy as being simply different:

> 48-4 Some the words he pronounces are different from the way a native speaker would say them.

53-8 “My brother, the colleagues, and the locality”, non-native phrasing and pronunciation.

and the discrepancy as mistakes:

54-9 This is not the question of doing one way or the other. *When it comes with stress and words, that would pretty much established things*, like everybody else, she makes stop, out of the… the American makes interdental, th fricative, so “ tink” instead of “ think”. (Italics added)

27-5 “ Then she’s explaining me”, that’s definitely coming from her L1, and she applied to her second language while she speaks English, and you know, you can’t use the same rule for the second language that you have in the first language. . . If she hangs out with her friends, she’s talking, we know what she’s saying. But at the high level group, it wouldn’t be right; when it comes to business setting, and academic settings, or something with higher level, you still can tell that she’s still very much influenced by her first language; it’s probably eventually, like my in-law, *it’s a fossilized mistake,* breaking that habit is gonna be impossible, unless they work on it very diligently.. . She sounds like she’s an advanced speaker, but I wouldn’t say she’s professional and *she’s not native like*. (Italics added)

Category 6. Fluency

Even though raters used the word, “fluency”, throughout the interviews, they seemed to interpret it differently and this can be categorized as: temporal fluency, accurate fluency and comprehension fluency.

<u>Temporal fluency (8 comments)</u>

This refers to the speech pace or naturalness regardless of the intelligibility or comprehensibility of the speech. Therefore, speech that was referred to as fluent may not make sense to the raters. The following two examples show generally negative and positive comments:

04-7 It’s quite hard to listen to. She speaks very fluently in her variety of English. She may use her dialect of English so maybe she brought up an environment different version of English is used, maybe simplified, or more slang variety. She sounded very fluent but what she said is very confusing to follow. [generally negative]

54-15 That’s quite good example of fluid; that’s interesting , the pronunciation is good, the speed is good, the linking is good, so he said quite naturally, not like a student would be thinking about structure, just saying naturally. [generally positive]

Flow fluency (7 comments)

The second sub-category is accurate fluency that views fluency as the ability to maintain the flow of the speech. Features which impede the continuity of the speech, such as filler words, hesitation, and repetition, would affect raters’ judgment of the speaker’s fluency level. The definition of this sub-category is specified in the rating descriptors of the English oral proficiency assessments, including IELTS speaking section. Examples of each of the features are presented below:

Filler word:

53-12 Okay, so I’ll say she repeats lots of “ like that”. It’s annoying. It would affect the fluency rating, maybe also for lexical, ‘cause it’s replacing what could be more exact and articulate expression of what she means. In the description it talks about hesitation while searching for correct words, she’s not hesitating ‘cause she just keep going, but she fills in with “ like that”, instead a more exact term. As for a rater, I’m not sure that’s because she doesn’t have vocabulary or she’s just kinda nervous speaker and filling in with time.

27-3 “I didn’t know that how to” was what she said, I didn’t know how to, so she has too many words there, “like that”, she’s trying to fill in; she was putting in a vocabulary that doesn’t need to be there.

Hesitation and repetition:

53-9 Here we’re getting more like IELTS because there are too many things that I don’t’ understand; really hesitation, searching for words, and repeating the same phrase, so she’s not as comfortable with English as the previous speaker.

Comprehension fluency (5 comments)

The third sub-category, comprehension fluency, is related to accuracy of language use and the extent to which raters are able to comprehend the speech. Speakers who were described as fluent in the other two sub-categories were not necessarily viewed as fluent here. This can be demonstrated in the following two examples made to the same speaker where raters indicated the speed of the speech was maintained but it was not considered as fluent:

> 77-21 She contains flow of speech but *her fluency is limited*, and sometimes it is difficulty to understand what she's trying to say. Her use of vocabulary is very limited. On the whole, her spoken English is very average. She just makes herself understood.

> 53-17 If you listen to this woman, it's just like constant language with very little pausing to help. So can you call this fluency? I guess it's overly fluent sort of because it sounds too fast for me to give up near native mark. Even though her vocabulary clearly she is comfortable with, she probably uses English all the time, but it's the speed and intonation that I think is off-putting. It'll be hard to listen for a long time and to interact, and to sit down to have a nice conversation.

Category 7. Level of the second language schooling

The final category with regard to the non-linguistic performance relates to the raters' reflection of the level of English education of the examinees. Of a total of 19 comments only one was a negative comment. The data shows that the lower level of linguistic maturity may not be negatively commented on but was indicative of a language learner. In other instances, a well-presented speech may be attributed to a good English education, instead of near nativeness as categorized previously. Two of the three raters (1 and 27) who used the level of language learning as an implicit scoring criterion seemed to score more leniently as shown in Table 49, indicating they were not judging against the precise forms as used by the native speaker but against the acceptable and expected language use by learners. The generally positive comments regarding good education received in English instruction are below:

> 27-20 I would say native like, she was very clear, higher end voc, she thought about the question, I assumed she did 'cause she was very organized, she spoke complete sentences, she spoke clearly and slowly, she enunciated, easy to understand, sounds very educated to me, formal schooling, *she sounds like she went to school to learn English, not just watch TV and learn the language, formal trained*, she was one of the best one. [Italics added]

1-41 This person can clearly communicate the message. I'd think probably someone study English, someone probably raise in the ESL environment, 'cause he's quite natural speaker; still with mistake structure, there is pronunciation issue, it shows varieties of English. But probably *someone raises not in the English environment but with access to English and then learn English well.* [Italics added]

27-18 So "he lived there" so she couldn't decide whether, are living there or I think she meant, he's currently living there 'cause she said "living there" so grammar and verb and article, things like that. I would say [she's a] high intermediate advanced student, some grammar class work on some grammar.

Rating tendency in relation to attitudes toward World Englishes

Ratings on the IELTS descriptive tasks were further examined to extract raters' attitude towards WE. Scores on each of the rating criteria were checked based on relative attitude groups of raters to establish if the attitude-behavior pattern as tested in Hypothesis 2 can be corroborated; that is, raters with relatively positive attitude rate leniently and those with relatively negative attitude rate harshly. The criterion measure is the group mean scores (*N*=96) on each of the rating category. This is to check if raters with positive attitude towards WE mostly rate higher than the mean score of each rating category and those with negative attitudes rate lower. Furthermore, rater' comments on each rating category along with scores awarded were evaluated to compare if the positive comments generally lead to higher scores and vice versa.

Table 52 shows the rating tendency on Fluency. For the five IELTS descriptive tasks used in this study, raters with relatively negative attitude displayed a clear pattern of rating lower than the mean scores on all speech samples, with the exception of two of the raters (Raters 48 and 53) in this group. For the relatively positive attitude group, ratings are rather evenly spread, indicating that raters in the positive groups did not necessarily rate higher than the mean scores. Raters' comments provide a more dynamic insight on the relationship between

Table 52

Rating Tendency on Fluency

Speech (Mean)	Relatively positive		Relatively negative	
	Higher than mean	Lower than mean	Higher than mean	Lower than mean
1(M=5.44)	R1, R23, R54, R77			R4, R53, R27, R48
2(M=5.82)	R1, R23, R54	R77		R4, R53, R27 R48
3(M=7.03)	R1, R54	R23, R77		R4, R53, R27, R48
5(M=7.91)	R1, R54	R23, R77		R4, R53, R27, R48
6 (M=8.23)	R1, R23, R54	R77	R53, R48	R4, R27

the types of comment and scoring tendencies. As presented in Table 53, positive comments are generally associated with higher scores and negative comments with lower scores. This is particularly true as the examinees' proficiency increases[4] (see samples 3, 5, and 6). Nevertheless, on the lower proficiency levels, negative comments may yield higher or the same scores as awarded by raters giving positive comments (i.e., sample 1 and 2).

Despite a relatively clear relationship between the types of comments and scoring tendency for Fluency rating, this was not clearly apparent for the rest of the rating categories. Table 54 to Table 59 show the rating tendencies and rater comments on Pronunciation, Vocabulary and Sentence Structure. Within each attitude group, ratings that are higher and lower than mean scores are generally evenly distributed suggesting that raters in the relatively positive group may score lower than the group mean scores and those with relatively negative attitude may score higher. It is also noted from rater comments that positive comments may lead to lower ratings whereas negative comments result in higher ratings than the group mean scores.

[4] It should be recalled that the larger the number of speech samples, the higher the examinees' IELTS official speaking scores.

Table 53

Rater Comments: Fluency

Rater (score)	Positive Comment	Rater (score)	Negative Comment
Speech 1			
Rater 1 (6)	That's native like language. She doesn't have to think about it. It comes naturally.	Rater 77(6)	She is willing to talk, but she's not very fluent.
		Rater 54 (7)	She adds unnecessary words which interfere her fluency.
Speech 2			
	Not found	Rater 1 (8)	What she said is very confusing to follow.
		Rater 48 (7)	Her fluency is not smooth at all.
Speech 3			
Rater 27 (9)	He was fluent.	Rater 77 (5)	He can't sustain his fluency as he drives up too soon, and has tendency to repeat himself.
Rater 1 (8)	She is quite fluent; speak clearly and naturally.		
Speech 5			
Rater 1 (9)	I wouldn't say he's native speaker but quite a fluent speaker.		Not found
Rater 23 (7)	Her line of speaking is definitely fluent, very progressive.		
Speech 6			
Rater 23 (9)	Her fluency in speaking caught my attention.	Rater 4 (8)	She said too quickly, I didn't get it. Is she in the competition?
Rater 1 (9)	She's very fluent. I would consider her a native speaker of the language.		

As shown in the Rater Comments on Vocabulary in

Table 57, Speech 5 elicited both positive and negative comments from raters. Rater 23 who made the positive comments that "Vocabulary is excellent" scored 5 on this speech sample whereas the negative comments provided by rater 4 ("Some words don't make sense, so it must be some other words".) and rater 53 ("It seems a little strange she used the word 'misplaced' instead of lost") scored this speech sample at 6 and 7 respectively. Furthermore, the same score may elicit different types of comments. Also as shown in Table 56, the same score may indicate different evaluations by raters. Both raters 48 and 1 assigned a score of 7 to Speech 3 whereas their comments differed. Rater 48 said, "He uses many high level words and his sentence structures are clear" while rater 1 noted that "In terms of vocabulary, it's quite limited and quite basic". The findings support Orr's (2002) study on the Cambridge First Certificate in English (FCE) Speaking test, leading the author to conclude that

> The varied nature of the raters' perceptions, with regard to what was heeded, and how it was judged, suggests that in normal circumstances it would be impossible to say how any one Speaking score had been reached. The validity of the interpretations that test users might with to make of the results is thus brought into question. (p.143)

Table 54

Rating Tendency on Pronunciation

Speech (Mean)	Relatively positive		Relatively negative	
	Higher than mean	Lower than mean	Higher than mean	Lower than mean
1(M=4.58)	R1, R23, R77	R54,	R4, R53, R27, R48	
2(M=4.21)	R1, R23, R77	R54	R53, R27, R48	R4
3(M=6.57)	R1, R23, R54,	R77	R53, R27, R48	R4
5(M=7.42)	R1, R77	R23, R54	R53, R27	R4, R48
6 (M=7.95)	R1, R23, R77	R54	R53, R27	R4, R48

Table 55

Rater Comments: Pronunciation

Rater (score)	Positive comment	Rater (score)	Negative comment
Speech 1			
Rater 1 (5)	Her pronunciation, intonation is very strong, very good.	Rater 54 (3)	Her ‘v’ in ‘haven’t’ was a bilabial, sounded more like a ‘b’.
Rater 48 (7)	Her pronunciation is clear but in some places it is not.	Rater 77 (5)	The speaker has incorrect intonation and stress.
Speech 2			
Rater 27 (8)	She sounds like she’s an advanced speaker.	Rater 27 (8)	She has pronunciation issue since she spoke quickly.
		Rater 48 (7)	She is also speaking so fast that her words are all jumbled up affecting her pronunciation.
Speech 3			
Rater 23 (7)	The way he pronounced the words interesting. He let out a little bit of his native accents, on a way I can feature him he was with a bunch of college kids watching soccer game.	Rater 77 (5)	His “oo” sound is articulated as the voiced labiodental “v “.
		Rater 48 (8)	His pronunciation of the word ‘Brazil’ does not sound like the way the word should be pronounced.
Rater 53 (8)	His pronunciation is pretty clear.		
Speech 5			
Rater 23 (7)	She got better intonation because she got very excited about the story.	Rater 1 (8)	Pronunciation issue, ‘cause it’s camel? Not too sure what he said.
Rater 53 (8)	I thought her rhythm and intonation is clear for listeners.		
Speech 6			
Rater 23 (9)	Perfect intonation that shows a question she’s asking herself.	Rater 53 (8)	There is a few things I didn’t’ understand, like “eee” is it a boom or? I didn’t understand the word.

Table 56

Rating Tendency on Vocabulary

Speech (Mean)	Relatively positive		Relatively negative	
	Higher than mean	Lower than mean	Higher than mean	Lower than mean
1(M=5.16)	R1, R23, R54	R77	R53, R48	R27
2(M=5.74)	R1, R23, R54	R77	R53, R27, R48	
3(M=6.98)	R1, R54	R23, R77	R53, R27, R48	
5(M=7.67)	R54	R1, R23, R77	R27	R53, R48
6 (M=8.17)	R1, R23, R54	R77	R27	R53, R48

Table 57

Rater Comments: Vocabulary

Rater (score)	Positive comment	Rater (score)	Negative comment
Speech 1			
Rater 1 (7)	She definitely got high level of language of vocabulary and phrase	Rater 54 (9)	There were two or three words that I can’t make out at all.
		Rater 77 (4)	She has limited vocabulary; just enough to get byway.
Speech 2			
Rater 53 (6)	Even though her vocabulary clearly she is comfortable with	Rater 4 (3)	She’s using simple vocabulary.
Speech 3			
Rater 48 (7)	He uses many high level words and his sentence structures are clear.	Rater 1 (7)	In terms of vocabulary, it’s quite limited and quite basic.
Rater 77 (6)	His vocabulary is just okay for the topic at hands.	Rater 4 (6)	There are lots of individual words and phrases that you can’t understand.
Speech 5			
Rater 23 (5)	Vocabulary was excellent.	Rater 4 (6)	Some words don’t make sense, so it must be some other words.
		Rater 53 (7)	It seems a little strange she used the word” misplaced’ instead of lost

Table 57 (cont.)

Speech 6			Not found
Rater 27 (9)	She was very clear of what some of the words that she used.		
	Her use of vocabulary is flexible.		
Rater 77 (7)			

Table 58

Rating Tendency on Sentence Structure

Speech (Mean)	Relatively positive		Relatively negative	
	Higher than mean	Lower than mean	Higher than mean	Lower than mean
1(M=5.23)	R54	R1, R23, R77	R53, R27, R48	R4
2(M=5.12)	R1, R23, R54	R77	R53, R27, R48	R4
3(M=6.89)	R1, R54	R23, R77	R53, R27, R48	R4
5(M=7.75)	R1, R54	R23,R77	R53, R27	R4, R48
6 (M=8.20)	R1, R23, R54	R77	R27	R4,R53, R48

Table 59

Rater Comments: Sentence Structure

Rater (score)	Positive comment	Rater (score)	Negative comment
Speech 1			
Rater 23 (5)	She used the expressions, like I used to, she used the past perfect, he has been . . . those are the signs of very high advanced student.	Rater 48 (6)	Her sentences are not correct sometimes because of word order and words she uses incorrectly.
Rater 77 (5)	She used complex structure.	Rater 53 (7)	She has fair number of grammar difficulties and mistakes.
Speech 2	Not found	Rater 1 (7)	She’s struggling with other structures.
		Rater 54 (6)	It was confusing.

Table 59 (cont.)

Speech 3			
Rater23 (6)	His English, grammatically speaking, he is very good.	Rater 1 (8)	She’s getting confused what tense to use.
Rater 77 (6)	He does use complex structure.	Rater 53 (6)	The time sequence is not clear.
Speech 5			
Rater 53 (8)	All of her grammars are in order.	Rater 27 (8)	She has lots of article and verb issues
Speech 6			
Rater 54 (9)	She’ll give examples of things in a good grammatical way.	Rater 23 (9)	He makes few grammar mistakes again with verbs.
Rater 27 (9)	She spoke complete sentences.		

Overall rater orientations

The overall orientation of raters emerged through the analysis of the verbal protocols. Raters in the relatively negative attitude group seemed to have a tendency to make negative comments with the exception of Rater 27 who had an almost equal number of positive and negative comments. With regard to the relatively positive attitude group, the generally expected pattern of positive comments was, nevertheless, not observed: two of the four raters made more positive comments and the other two mostly made negative comments. Figure 30 displays raters’ overall orientation of the five IELTS descriptive tasks in relation to rater attitude tendency. As shown, it is clear that raters’ views of an examinee’s speech performance vary to some extent but some generalizations based on the attitude group that the rater belongs to may be roughly observed. In the negative attitude group, Rater 27 seems to rely on his underlying criterion, native speaker performance, for scoring judgment and this was also observed in some of the verbal protocol reports by Rater 53. This is consistent with the observations in the RAI construction phase 1, where a rater displaying a negative attitude towards language variations pointed out the judgment on examinees' speech performance should take into consideration if the speech is to make sense to someone from the U.S. Midwest. Nevertheless, in addition to the implicit reliance on the native speaker model for judgment, the other three raters in this group focused more on the comprehensibility of the speech to guide their rating. This can be demonstrated in the opening remarks by Rater 53 on most of the speech samples, such as "Overall, I mostly understood what he said" (for speech 3) and "I follow everything she said" (for speech 5). In the relatively positive group, three of the four raters seemed to keep examinees' status as language learners in mind and frequently made such comments as "She's an intermediate language learner" (Rater 1), "She's a good user of English language" (Rater 77).

Positive	IELTS descriptive task scores	
	Low	High
Attitude	Rater 23 (Grammar & language learner) Rater 77 (Efficient language user)	Rater 1 (Language learner) Rater 54 (Pronunciation)
Negative	Rater 4 (Big picture of the story)	Rater 53 (Comprehensibility & Native speaker) Rater 27 (Native speaker) Rater 48 (Listener effort)

Figure 30. Raters' overall orientation of the five IELTS descriptive tasks

Of the three raters, Rater 23 in particular attended to use of grammar and generally rated harshly on this category. This can be observed in his ratings for some of the speech samples that were lower than the group mean scores. For Rater 54, she is the only rater in the positive attitude group that made most of the comments concerning pronunciation features and generally rated lower than the group's mean scores on this rating category. It is not surprising that the raters' focus on the aspects of language use vary differently particularly since they were not given any training prior to the study. This also supports earlier studies seeking a rating process using verbal protocol studies on different task types (Brown, 2000; Brown *et al*, 2005; May, 2010; Orr, 2002). Several particular categories appear to be more salient to some raters, while others may relate to most of the categories when judging the speech. In addition, as the level of examinees' proficiency increases, raters made more non-linguistic comments, such as a speaker's level of confidence and mood elaborating on the topic at hand.

CHAPTER 6

DISCUSSION AND CONCLUSION

This study aims to broaden the understanding of the impact of WE on rater scoring performance in oral proficiency assessment. In view of the rich research findings in the field of language attitude suggesting that non-standard English is less preferred by listeners, of greater pertinence in language assessment is whether listeners transfer such attitudes to behavior. Test fairness will be in question and the inferences of score use and interpretation will be challenged if rater attitude towards examinees of multiple varieties is biased and, further, if it affects rater scoring judgment. Two separate yet inter-dependent studies were conducted to address this issue: (1) development and validation of the Rater Attitude Instrument (RAI) to evaluate the measurable portion of rater attitude towards WE and (2) an examination of the relationship between rater attitude towards World Englishes and scoring tendency through the use of IELTS descriptive tasks produced by Indian examinees as an elicitation stimulus. To strengthen the inferences from study findings, two inference arguments (Toulmin, 2003) were outlined and guided the study procedures.

Study 1, covering the development of RAI, included three phases. First, an extensive literature review of attitude constructs and an elicitation of rater views and attitudes towards WE informed the RAI item construction. Second, an exploration of the RAI internal structures with findings facilitated the RAI item revisions. Third, a verification process determined multi-dimensional constructs of rater attitude towards WE. Study 2 involved 96 ESL/EFL teachers, including 23 IELTS raters, to respond to the RAI and six IELTS descriptive tasks. The results were cross-analyzed to test the five hypothesis put forward in study 2: (1) raters' attitude is not consistent and can be grouped into different attitude groups; (2) the rater attitude group is a main effect on

on IELTS descriptive task scores; (3) ratings of IELTS speaking descriptive tasks may be predicted to some extent by rater WE attitudes; (4) rater attitude may be associated with rater background characteristics; and (5) rater with similar attitudes may score the IELTS descriptive tasks in a similar fashion by weighing particular salient features of Indian English more heavily than others. Multiple sources of evidence were collected and cross-analyzed in part in mixed-methods fashion to strength the inference arguments outlined in the two studies, justify the divergence of findings that emerged in the study and broaden our understanding of the complexity of rater attitude towards WE and scoring tendency.

Summary of Findings and Discussions

The summary of findings and discussions is based on the two inference arguments proposed in each study. Where necessary, evidence collected in support of one warrant will be used to support or refute the findings in other warrants.

Development and validation of the RAI

The claim that the RAI provides supportable evidence of inferences about multidimensional aspects of rater attitudes towards WE is supported by three warrants along with backings and evidence and is discussed respectively in this section. Table 60 presents the validity evidence and counter evidence in support of the validity of RAI.

Table 60

Validity Evidence in Support of the Rater Attitude Instrument

Claim	The Rater Attitude Instrument (RAI) provides supportable evidence of inferences about multidimensional aspects of rater attitude towards World Englishes.	
Warrant	Backing	
	Supported evidence	Counter evidence
1.1. A measurement model of multi-dimensional rater attitude was established.	• Establishment of a 2-factor measurement model of rater attitude towards WE • A 3-factor internal structure of rater feeling	• Items in sections of rater belief and rating tendency evaluated only by Cronbach's Alpha • Low item internal consistency in sections of rater belief and rating tendency revealed by Cronbach's Alpha
1.2. RAI specific tailored to language assessment needs	• Items tightly connected with findings of in-depth interviews with raters of Berlitz Proficiency Interviews • Items informed by the literature reviews in language assessment research in relation to World Englishes	• Items in the measure of rater feeling informed by the undergraduates' views as opposed to raters of oral proficiency assessment
1.3. Evidence of content validity supported the RAI item development.	• Items reviewed by content experts in every stage of the RAI construction • Item revised as a result of consensus by researcher and content experts • Item revision as informed by the qualitative feedback from raters	• Lack of interaction among content experts • Qualitative feedback not elicited from all the raters

Warrant 1.1. A Measurement Model of Multi-Dimensional Rater Attitude was Established.

The conceptualization of the 3-factor attitude model was tested by factor analysis and classical testing theory which yielded to a 2-factor measurement model that best represented the current data set. The measure of each rater attitude component was evaluated by appropriate statistical methods: rater feeling as measured by the semantic differential scale was inspected for its internal structure using exploratory and confirmatory factor analysis (*Backing 1.21*); and rater belief and rating tendency, assessed on the Likert scale, mainly used Cronbach's Alpha to determine item suitability(*Backing 1.22*). Finally, all three conceptual attitude components were tested together in a series of confirmatory factor analyses (CFA) to establish a measurement model (*Backing1.23*).

The 25 pairs of adjectives measuring rater feeling were evaluated by exploratory factor analysis (EFA) via oblique rotation and yielded three meaningful factors that accounted for 68.28 % of the total variance. According to the results of the EFA, two a priori 3-factor models conceptualizing rater feeling were established. The first model was informed by the results of factor extraction and the second was proposed with several items swapped into different latent factors according to interpretability. Series of CFAs were performed and indicated that the removal of two items, *Quick* and *Talkative,* in both models with low square multiple correlations would considerably increase the fit indices. This resulted in the second model yielding better fit indices all exceeding the recommended cutoff values ($\chi^2 =198.208, p =.000, RMSEA=0.082, CFI =0.959, TLI=0.945$). The three factors were labeled speech competency (i.e. Fluent, Articulate, Good Pronunciation and Sure), kind-heartedness (i.e. Kind, Considerate and Good natured) and level of confidence (i.e. Intelligent, Educated, Experienced and Informative).

The item consistency on measures of the other two attitude components, rater belief and rating tendency, on a 5-point Likert scale revealed less desirable results by Cronbach's Alpha. Even though the alpha for the entire Likert scale in both the exploratory and verification phases of RAI construction was acceptable (.738 in the exploratory phase and .628 in the verification phase), the alpha for each subscale was less satisfactory, ranging from .361 (rating tendency) to .726 (expectations of Indian English). This apparently indicates a need for further item modification in future research. Following the common practice of scale construction, more items than currently remained would be deleted due to low alpha. Nevertheless, as the RAI is among the first instruments to measure rater attitude in language assessment research, its scope in this study aims for comprehensiveness of concerns addressed by raters and testing professionals. As Kattan (2009) argues, "measures of internal consistency give information on reliability, not validity" (p.580) and, as such, items that covered the different dimensions of rater belief and rating tendency were therefore not sacrificed as a result of internal consistency checks. On the contrary, the low alpha in the sections provides valuable implications of rater uncertainty to the questions and indicates a need for further study to investigate the cause.

To establish the conceptualized 3-factor rater attitude model to confirmatory factor measurement model, it was first necessary to standardize the scores across three factors and use summated subscale scores for analysis to compensate for the insufficient sample size if individual items are used. The three components of the RAI, rater feeling, belief and rating tendency, were treated as latent factors and seven subscales (i.e. speech competency) were indicators. A 3-factor model was tested by CFA which showed the correlation between latent factors 2 (rating tendency) and 3 (rating belief) was over 0.90 suggesting an overlapping between the two factors and a need to reduce the number of factors. A 2-factor model was run which yielded good fit indices (χ^2 =20.052, *p* =.094, *RMSEA*=0.076, *CFI* =0.954, *TLI*=0.926) all well exceeding the recommended cutoff values. A 1-factor model was also run to ascertain if it provided better fit indices and interpretation. Though the model yielded good fit indices, the chi-square test of difference showed no statistical difference between the 1- and 2- factor models. Besides, two negative factor loadings in the 1-factor model between the latent factor (i.e. rater attitude) and the indicators (i.e. expectation of Indian English and interpersonal history respectively) made for increased difficulty in interpreting the results. Thus, although the CFA failed to confirm the hypothesized 3-factor attitude model, the resulting 2-factor measurement model appeared to best demonstrate evidence of adequate construct validity, indicating that two rater attitude dimensions of rater feeling and belief subsumed rating tendency. Thus, the goal of this study was achieved; namely, to identify and establish the internal structure of rater attitude model within the language assessment context. Figure 20 on page 120 presents the 2-factor measurement model.

The 2- factor attitude measurement model has valuable implications. First, the internal structure of 3-factor rater feeling presents a mix of consistent and contradictory findings to previous language attitude research. Factor 1, Speech Competency, explained the greatest percentage of variance. Unlike other language attitude scales (Bradac, Bowers & Courright, 1979; Bradac, Desmond & Murdock, 1977; Zahn & Hopper, 1985), it combines speaking quality (clear-unclear, fluent-not fluent, good pronunciation-bad pronunciation) and confident certainty of the speech (sure-unsure). This indicated that factor 1 covered a broader range of evaluations and implied the traditional criteria for good speaking based solely on speaking quality was not sufficient enough to constitute raters' good feeling of one's speech competency. The item, *sure*, may indicate or be associated with accuracy of speech, such as linguistic use.

Raters' evaluation of this factor may reflect the long professional discussions of approaches to ESL/EFL teaching and learning on whether fluency or accuracy should be accorded greater emphasis in the classroom context (Hammerly, 1991). Factor 1 suggests raters conceptualization of good speech is a balance of both. Factor 2, Kind-Heartedness, is in many ways similar to other language attitude studies (Carranza & Ryan, 1975; Ryan, Carranza & Moffie, 1977) displaying overlapping items. This factor suggests raters' concern with the qualities or attractiveness of the speakers along with their speech, leading to a broader dimension of evaluating the speech that take speaker's character and likeability into consideration. The third factor, Level of Confidence, includes items from Zahn and Hopper's (1985) superiority factor. It consists of elements such as educated, experienced, informative and intellectual. They displayed rater perception of a speaker's social status and intellectual achievement. This factor was well correlated (*r* =.899) with the first factor, speech competency, suggesting that speech competency as perceived by raters are aligned with their feeling towards or implications about the examinees' level of education. This conforms to the findings in the verbal protocol study, described later, where raters attributed an examinee's high level of oral proficiency to good ESL/EFL education.

Looking further at the traits associated with the standard or "non-standard" varieties, the current findings in measuring rater feeling do not fully support earlier language attitude research (Cargile & Giles, 1998; Paltridge & Giles, 1984; Wilson & Bayard, 1992) which suggest that a standard variety is most often associated with power and was rated highly on traits, such as competence, intelligence and social status whereas the "non-standard" variety is linked to lower socioeconomic success (Fishman, 1971) and the traits rated lower even by listeners who themselves share the same variety as the speaker or with a "non-standard" accent. When speakers are evaluated along traits related to kindness, attractiveness and solidarity, those with a “non-standard” accent are rated favorably. While authors in these language attitude studies do not engage in WE discussions, the term, “non-standard variety” may be interpreted as “new varieties of English” (D’Souza, 2001) and expanding circle variety, or simply non-inner circle varieties.

In this study, rater responses to the traits measuring speaker’s Kind-Heartedness had the highest mean score (*M*= 5.12) of the three factors that conformed to the expected response pattern in “non-standard” variety. The other two factors, Speech Competency and Level of Confidence, both labeled as superiority in Zahn and Hopper (1985) and associated with standard variety traits, had mean scores greater than medium, suggesting raters generally evaluated positively on their feelings of examinees of Indian English. This apparently contrasts to findings of previous attitude studies and may be attributed to two reasons. First, as opposed to the predominant use of undergraduate students in previous language attitude studies, this study involved raters of ESL/EFL teachers that are most likely to have awareness of WE, though perhaps in varying degrees. Second, most of the raters were located in New York city where exposure to diverse language learners is very common. Both may lead to more accepting views and feelings toward examinees of Indian English. To consolidate the current findings, contrasting varieties between outer- and expanding-circle, for example, Indian English and Chinese English, may be used simultaneously to compare responses to traits related to standard and “non-standard” variety as suggested in previous attitude studies. Furthermore, listener response may be context-specific (Zahn & Hopper, 1985). As such, it is worth the research effort to use the same data set with groups of listeners besides raters of oral proficiency assessments and verify whether the current findings are rater-sensitive to increase the validity argument of the RAI score use and interpretation.

In the section of rater belief and tendency, rater responses revealed several contrasting points that warrant attention by language testing professional. To a large extent, raters acknowledged the role of WE in daily and cross-cultural communication and the need to increase awareness of rater and ESL/EFL leaners in the classroom of the global spread of English through, among others, exposure to WE during rater trainings. Knowledge of English language spread was also demonstrated in raters' concurrence that non-inner circle speakers outnumber their inner-circle counterparts' (Crystal, 1997; Graddol, 1997). Nevertheless, when it comes to evaluating examinee oral proficiency, more than half of the raters preferred standard English as the criteria on which to base the scores. This seems to imply a secure blanket provided by standard English to guide a fair scoring process as opposed to the uncertainty of handling unfamiliar expressions by examinees on the multiple-variety tests. Although raters in this study were generally accepting in treating unfamiliar expressions as part of examinees' variety rather than as indications of not having fully mastered English, raters' preferences for using standard English to judge performance were not in conformance with the arguments put forth by WE-view language testing researchers (Chalhoub-Deville & Wigglesworth, 2005; Davidson, 1993; Lowenberg, 2002; Spolsky, 1993). Opening up to WE varieties suggests a more active involvement and listener effort in negotiating intended meanings (Canagarajah, 2006; Elder & Davies, 2006; Jenkins, 2006), or the "test accommodation" approach as enunciated by Elder and Davies (2006). Nevertheless, whether it is feasible in generating fair scoring results in test situations challenges raters' willingness to engage in meaningful interaction and interviewing styles and techniques, which may ultimately affect examinees' performances and the scores awarded (Brown, 1995; 2000). Furthermore, the demands placed on raters in judging examinees of WE varieties may introduce even more uncertainty particularly in assessing monologue tasks, such as used as stimulus in this study, where raters would not have the opportunity to clarify any questionable responses. Raters' voice should be factored into any significant change in the speaking test practice as their readiness or reluctance for change in assessing examinees would presumably affect the scores they award and any inference made based on test scores about examinee speaking competency.

Despite the preference for a standard English, rater responses to rater scoring tendency revealed contrasting results where the native speaker model was not used to judge an examinee's oral proficiency level. Instead, clarity and speech comprehensibility helped determine the final scores. This reflects the call by testing professionals for an *efficient language user* rather than *native speaker* as the benchmark for assessing language ability (Bachman & Savingnon, 1986; Barnwell, 1989; Davies, 2003; Hamilton *et al*., 1993). Even though responses to rating tendency generally revealed a positive and accepting stance of considering examinees' varieties in the oral proficiency assessment, the section on rating tendency, as tested in study 2, was not significantly related to IELTS descriptive task scores and was not a significant predictor, suggesting the surface interpretation of liberal views of embracing multiple varieties in the test may not be the case in the real testing situation. It is possible that raters gave "socially acceptable responses" (Bernreuter, 1933; Lenski& Leggett, 1960; Vernon, 1934 as cited in Ajzen & Fishbein, 2005) and did not express or remember their true decision-making process in the testing situation. This also questions the feasibility of measuring behavior tendency in the attitude measurement (Allport, 1935; Woodmansee & Cook, 1967 as cited in Ajzen & Fishbein, 2005). The underlying rating tendency seems to be better reflected in the verbal protocol study where raters, with varying degrees, had the tendency to compare examinees' speaking performance with the expected patterns in native speaker model(i.e. American English), which better supports the findings for a preference for standard English in the test context.

Besides the general attitude toward WE, specific expectations of Indian English were measured, suggesting raters were generally positive about the latter as a steady variety that had its own linguistic features and treated its speakers as native speakers of English, though if may not be a pervasive opinion among the general public in the U.S. (cf. Llurda, 2009). Despite that, a majority of raters believed Indian speakers should not be exempted from English proficiency tests. This may be explained by the fact that the amount of English-medium instruction each Indian examinee received differed leading to varying levels of English proficiency (Hohenthal, 2003). In reviewing current language requirement for university admission in the U.S., the requirement of proof of English language proficiency for Indian applicants broadly falls into three categories: exemption, conditional exemption and mandatory. In the exempt situation, a list of inner- and outer- circle countries are included and students from listed countries, including India, might be exempted from having to evidence English language proficiency scores, such as TOEFL and IELTS. Conditional exemption usually applies to students who have received instruction in English outside of the U.S. and requires a letter from their institution stating that the language of instruction was English. The last category requires TOEFL or IELTS from applicants where English is not the "ubiquitous language" (retrieved from UCLA admission website, *http://www.anderson.ucla.edu/x21453.xml*). In the UCLA admission websites, Indian applicants, for example, are specifically highlighted as requiring TOEFL or IELTS. The different language proficiency requirement for Indian applicants signifies each institution’s acknowledgment of the varied status of Indian English; nevertheless, it may have practical connotations. Requiring TOEFL or IELTS scores for all international students, including those from India, could serve to standardize and facilitate the administration process as no additional labor is needed to review the letter proving applicant’s had received education in English medium universities.

Warrant 1.2. RAI Specific Tailored to Language Assessment Needs

The findings of interviews with the raters of Berlitz Proficiency Interviews not only paved the way for the development the RAI but greatly shed light on rater opinions, awareness and thoughts on WE which, despite rater calibration, were transmitted to their rating beliefs. Iterative reviews of interview data revealed that the formation of perceptions in WE was greatly influenced by rater education, hometown environment, personal interests and job achievements, all of which influenced in varying degrees their rating tendency and commitment to a fair scoring process. The initial exploration of rating tendency was partly aligned with Kim’s (2005) dissertation study. Raters with more liberal views on WE placed less emphasis on linguistic use than on task fulfillment and communicative ability in assessing English language oral performance. Alternatively, raters with less open-minded views attributed hesitation in rating due to language variations to the problem of naïve raters. Despite acknowledging the global spread of English, one rater claimed that speaking performance should be conscious of acceptance by listeners from the U.S. Midwest who may be less exposed to varieties. From this perspective, the rating on each linguistic use should be rigorously judged against standard American English. Nevertheless, given the variations within standard American English (Wolfram, 2006), the significant enrolment of international students in Midwest universities and the associated increase in interaction with locals, it is debatable whether Midwesterners are necessarily less familiar with or less tolerant of English varieties.

In terms of the attempt to devise a rater-sensitive instrument to increase the generalizability of findings in the context of language assessment research, the involvement of raters in the entire cycle of RAI development was not fully accomplished due to financial constraints and lack of accessibility to raters at the time certain studies were conducted. To accommodate the gap between rater background and recruited participants, the alternatives in participant selection were to either match as closely as possible to the background of raters of oral proficiency assessments (i.e. ESL/EFL teachers) or implement a rigorous study procedure for participants to be thoroughly familiarized with investigation aspects to counteract validation criticism. This constraint on subject selection has challenged the capability of a single PhD candidate researcher. As such, undergraduate students, despite being criticized in attitude studies reviewed by Reddington (2008), could be the most efficient and feasible resources on condition that study procedures are carefully designed and administered. The issue of rater accessibility in particular posed concerns on scale construction where a large number of subjects would be preferable.

Warrant 1.3. Evidence of Content Validity Supported the RAI Development.

Content validity of the RAI construction played a crucial role in determination of item quality, clarity, removal and retention particularly when construct validity indicated otherwise. This is particularly crucial for the current study as the RAI attempts to capture comprehensiveness of issues addressed by testing and WE professionals. In each of the three phases of RAI construction, content review was concurrently conducted with a measurement inspection of the RAI before proceeding to the next phase of the study. Items suggested to be removed by construct validity evidence may be retained as a result of expert judgment. The attempt to demonstrate and increase content validity is to respond to the current practice of modern language assessment inquiry to value multiple evidences to support and strengthen the inference from study findings (Kane *et al*, 1999; 2001; 2002; 2004; 2006, Messick, 1998; Mislevy, 2004). It also responds to current thinking of introducing expert judgment by those most familiar with the study context to enhance the credibility of study findings (Moss, 1992, 2004; Watt, 2007). It was argued that the judgment and consensus reached between content reviewers and current researchers are valuable evidence in claiming that the “concept” of reliability is achieved through consistent interpretations and justification by people most knowledgeable about the context of assessment, as opposed to the “value” of reliability from a positivist stance.

The RAI, as an initial study in language testing inquiry, contributes to the literature by revealing a measure of rater attitude of concern in second language oral assessments influenced by WE. Rater attitude towards WE speakers are not completely the same as those governing general language attitude studies. It also contributes to our understanding of the discrepancy between rater views in dealing with oral proficiency assessments with WE examinees and researchers’ call for a WE-oriented oral language assessment. Equally important, the multiple evidence collected in the construction and validation of RAI supports the arguments that the RAI has compelling content validity, adequate psychometric properties with further modifications needed to render a more powerful tool, and a clearly interpretable factor structure, that is, rater feeling and rater belief.

Relationship between Rater Attitude towards World Englishes and Rating Tendency

Concerns with regard to rater attitudes toward WE being a biasing factor affecting rating tendency was investigated (*Claim*) in study 2. For brevity and cohesiveness, two warrants pertaining particularly to rater attitude-behavior relationship are discussed together. The validity evidence for the second claim is presented in Table 61.

Table 61

Validity Evidence In Support of the Rater Attitude-Rating Tendency Relationship

Claim	Rater attitude towards World Englishes is a biasing rater factor that influences rater scoring performance of IELTS descriptive tasks.	
Warrant	Backing	
	Supported evidence	Counter evidence
2.1 Raters grouped into three relative attitude groups: positive, neutral and negative	• Unequal severity level of rater attitudes towards WE as suggested by FACETS analysis • Raters grouped into three attitude groups according to measurement logit generated by FACETS analysis	• Groupings made relatively rather than absolutely as mean scores and FACETS suggesting rather generally liberal views in WE
2.2 Rater attitude group as a main effect on IELTS descriptive task scores	• Rater attitude effect revealed by MANOVA • Significant mean difference between positive and negative attitude group on all criteria • Significant mean difference between neutral and negative attitude group on several criteria	• Not detected
2.3 RAI and rater background characteristics as predictors of IELTS descriptive task scores	• IELTS descriptive task scores significantly related to attitude scores • Attitude scores and Indian/Non-Indian predicting IELTS descriptive tasks scores	• Indian/non-Indian variable contributing only to small variance in IELTS descriptive tasks scores
2.4 Associations between rater attitude and background characteristics established	• Rater feeling significantly related to Indian/Non-Indian variable • Indian/Non-Indian variable a significant predictor of rater feeling scores	• Weak relationship between two variables probably resulting from occasional occurrences
2.5 Different salient variety features attended to by raters with similar attitudes	• Rater with positive attitude considering expected performance of language learners; some raters with negative attitude focusing on native speaker model	• Exceptions found in each attitude group

The five warrants providing multiple evidence to support the claim in study 2 include evaluation of rater attitude inclination (*Warrant2.1*), rater attitude group as a main effect on IELTS descriptive task scores (*Warrant 2.2*), rating performance on IELTS descriptive tasks in relation to rater attitudes toward WE (*Warrant 2.3*), an association of rater characteristic background and attitudes towards WE (*Warrant 2.4*), and salience of Indian English variety as attended to by raters with similar attitudes (*Warrant 2.5*).

Warrant 2.1. Raters Grouped into Three Relative Attitude Groups: Positive, Neutral and Negative.

It was hypothesized that within the measurable portion of rater attitudes towards WE, raters' attitude is not consistent and can be classified into different attitude groups (*Warrant 2.1*). The severity of rater attitudes was evaluated by descriptive statistics (*Backing 2.11*) and FACETS analysis (*Backing 2.12*). The mean scores for each component of the RAI indicate that raters generally hold a positive attitude toward WE, which is further supported by the negative mean of measurement logits (M= -.011) in FACETS analysis, implying a generally lenient rating in response. The summary of FACETS analysis showed that Raters' logit values extend from -1.22 to +.89, a meaningful range of 2.11 logit given the three statistical indices: the separation index (1.06), reliability[5] (.53) and fixed chi-square test (χ^2 =197.4, p = 0). These indices indicate raters' severity did not vary considerably in responding to the RAI but that the individual differences did exist. Therefore, it is valid to group raters in their relative standing according to the respective measurement logit into three attitude groups: positive attitude group (N=56) as result of negative measurement logit, neutral attitude group (N=4) that had

[5] Note that the reliability statistic produced by the FACETS analysis is different from the traditional sense of inter-rater reliability as the latter refers to the degree of consistency between raters whereas the former reports the extent to which the analysis reliably distinguishes raters into different levels of severity. High reliability means that raters are being reliably separated into different levels of severity. The reliability for the current data set was .53, implying that raters may differ and do not share similar levels of rating severity.

measurement logit of zero and negative attitude group (*N*=36) with positive measurement logit. This classification in rater attitude toward WE in relation to oral proficiency assessments served as the basis for further analysis. Unlike other attitude studies that used raw scores of the criterion measure for grouping (Coniam, 2010; Kim, 2005) by rank-ordering scores placing raters with higher half scores into the positive group and the rest into the negative group, this study used FACETS analysis to examine raters' relative severity of perception rating and to justify the grouping by assessing the three statistical indices. This approach informs that the grouping is relative rather than absolute according to rater relative standing that guides interpretations of findings. This grouping approach by FACETS analysis that is popularly used in language assessment research to monitor rater consistency of rating (Kondo-Brown, 2002; Lynch & McNamara, 1998; Weigle, 1999; Wigglesworth, 1993; Zhang & Elder, 2011) is useful to inspect relative severity of rater attitude towards WE and is recommended for other attitude study.

The difficulty estimate for the seven rating components as measured by FACETS analysis shows that the three components constituting rater feeling (Speech Competency, Kind-Heartedness and Level of Confidence) were found most difficult to rate. The other four Likert-scale components measuring rater belief and tendency had negative measurement logit, suggesting more lenient ratings. It was probable that the response format (i.e. the 7-point semantic differential scale) of rater feeling was less familiar to raters as it required them to think outside the box and placed themselves out of the typical assessment scenario in presenting their immediate reactions and feelings about an examinee's voice and examinee him/er self. It apparently differed from the common rating practice to assess examinee's level of English oral proficiency, which caused concerns and hesitation in responding as indicated by a few raters in the comment section. Raters expressed concerns that the response format failed to provide ratings on certain items, such as intelligent and unintelligent, due to difficulty in judging one's intelligence level based on a 90-second speech. As Dillman, Phelps, Tortrora, Swift, Kohrell, Berck & Messer (2009) cautioned, using a format less familiar to respondents may increase response error. Nevertheless, it may be argued that raters were probably more cautious and careful when responding to this section, as the measure of rater feeling was found significantly related to IELTS descriptive tasks ratings (total and all of the four sub scores), which was revealed in the evidence supported by the following two warrants.

Warrant 2.2 &2. 3. Rater Attitude Group as a Main Effect on IELTS Descriptive Task Scores, and RAI and Rater Characteristic Background as Predictors of IELTS Descriptive Task Scores

The data further probed the relationship between rater scoring performance and their respective rater attitude group (*Warrant 2.2*) and supports the previous language attitude studies (Lindemann, 2002; Rubin, 1992) that a positive attitude contributes to positive behavior. The one-factor multivariate analysis of variance (MANOVA) (*Backing 2.2*) that evaluated the variability as explained by rater attitude groups against the four rating criteria (Fluency, Pronunciation, Sentence Structure and Vocabulary) of the IELTS descriptive tasks showed unfavorable yet expected results: the main group effect was significant, implying that examinees' scores on IELTS descriptive tasks in this study significantly depended upon the group rating their speech. The tests of between-subjects effects further showed that rater attitude had statistically significant effect on all the four dependent/rating variables. Tukey contrasts analysis showed significant differences between positive and negative attitude groups, with the former consistently giving higher mean score ratings on each of the four rating criteria. Furthermore, neutral and negative attitude groups were found to have significant differences in mean ratings on sentence structure and vocabulary, with the neutral group giving higher mean score ratings on these two criteria.

The results partly support the similar language assessment study by Kim (2005) using Korean speech sample as attitude stimulus. Kim found that raters with a positive attitude gave higher mean score ratings than neutral and negative groups on three criteria: grammatical accuracy, rate of speech (“fluency” in the current study) and task fulfillment. Significant differences between neutral and negative attitude groups were not found as in the current study. Despite difficulty in drawing comparable conclusions in findings due to different measures, designs, rating criteria and elicitation stimulus, a strong yet unfavorable indication of test unfairness clearly arises, that is, biaseness in English oral proficiency assessment scores due to attitude raters hold towards WE. This indicates a clear need to monitor and evaluate rater attitudes towards WE to prepare raters to be more confident and objectively handle the multiple varieties encountered in oral tests and, ultimately, enhance scoring validity. Unlike other researched rater biasing factors, such as residency, nationality and English language background, which cannot be ignored or changed, rater attitude is a psychometric trait shaped by multi-dimension external factors (Cargile *et al*, 1994) and is susceptible to change over time, though gradually (Miller, 2008). Pertinent to the magnitude of the attitude-behavior relationship, it was argued that direct experience as compared to secondhand information would strengthen the stability of attitudes (cf. Eagly & Chaiken, 1993; Fazio & Towles-Schwen, 1999). As such, actions should be taken by testing professionals and agencies, for example, to use oral data from WE varieties that raters would be most likely to assess, provide opportunities for raters to interact with the WE speakers and increase raters’ exposure with the speakers via designed activities to increase their direct experience with WE varieties as opposed to only listening to or watching training speech samples typically offered in training programs (Fulcher, 2003; Luoma, 2004; Taylor, 2002). Previous research also shows strong evidence that direct intercultural contact facilitates listener's comprehension of speech (Derwing & Munro, 1997; Field, 2003; Gass & Varonis, 1984; Kang 2008; Powers, Scheldi, Leung, & Butler, 1999).

A further investigation of the predictive power of the RAI and rater background variables on IELTS descriptive task scores (*Warrant 2.3*) suggest a moderate relationship between the two. RAI total scores and RAI part score 1 (rater feeling) were significantly related to both IELTS descriptive task composite score and each of the four analytic scores. Nevertheless, RAI part score 2 (i.e. rater belief) revealed lower magnitude of association as it significantly related only to pronunciation scores. In terms of rater background variables, the Indian/non-Indian rater background variable was the only predictor significantly related to the IELTS descriptive total score, analytic score on Sentence Structure and Vocabulary. The RAI total score was the strongest predictor of IELTS descriptive task total and any of the four sub-scale ratings. The total variance it accounted for ranged from 17.5% in the Pronunciation score to 32.4% for Vocabulary, including a quite surprisingly high proportion of variance (31.3%) in the IELTS descriptive task total scores. That is to say, setting aside extraneous variables other than rater attitude that may affect scores of oral proficiency assessments (Barnwell, 1989; Brown, 1995; Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Cumming, 1990; Shi, 2001), very proficient examinees in English oral competency can ensure that only 70% of the variance in total score is contributed to by their own level of English oral proficiency and the balance on the chance of being assessed by a rater having positive attitude towards WE. Thus, the RAI provides testing professionals and agencies a powerful tool for demonstrating efforts at ensuring test fairness through incorporating the element of rater attitude towards WE as an essential component of their research and training agendas.

In the examination of other rater background variables, the Indian/non-Indian component was the second strongest predictor for the IELTS descriptive task total scores, Sentence Structure and Vocabulary scoring categories, despite only contributing to approximately 10% of the total variance in each of the category above. Indian raters were found to be significantly harsher with ratings on the three scoring categories than were non-Indian raters. These results paralleled findings in previous studies (Brown; 1995; Kang, 2008). That is, raters of NNS of English were substantially harsher than raters of NS of English on linguistic items. Perhaps NNSs of English had experienced the learning process and difficulty and were more easily able to detect other learner's language learning issues. One Indian rater interviewee in the verbal protocol study pointed out one common issue with Indian examinees was the fast speaking rate, noting that many of them did not keep in mind the features of spoken English and tended to articulate continuously and attributed this to the quality of English medium schools they attended. Santos (1988) reported that NNS raters who had achieved high levels of English language proficiency often judged the errors of other NNSs more severely than NSs.

Warrant 2.4. Associations between Rater Attitude and Background Characteristics Established.

With regard to whether rater background characteristics can predict the tendency of rater attitude toward WE (*Warrant 2.2*), the results of both correlation analysis and multiple regression analysis show that of the five background characteristics (i.e. Indian/non-Indian, native language, gender, teaching experience and highest level of education), only the Indian/non-Indian variable was significantly related to the score on rater feeling (r = -.231, p <.05). The measure of rater belief was not associated with any of the rater background characteristics. The negative correlation showed that the lower ratings on rater feeling were associated with Indian raters. Furthermore, regression analysis showed that only 4.7% of the total variance in scores of rater feeling was accounted for by the Indian/non-Indian variable. Although it conforms to findings in some language attitude studies (Barona, 2008; Giles & Billings, 2004; McKenzie, 2008) that listeners may share the same variety with speakers yet harbor negative attitudes of the speakers, the current finding should be interpreted with caution. Even though the Indian/non-Indian variable was a statistically strong predictor at the .05 level, it did not necessarily imply practical significance (Krueger, 2001). The small shared variance suggests that either the Indian/non-Indian variable matters very little with the measure of rater feeling or the weak relationship between the two may be a spurious occurrence. A more compelling interpretation about the impact brought by the Indian/non-Indian variable on ratings of rater feeling of speakers of multiple varieties should be further investigated based on a larger sample size of Indian English speakers.

Warrant 2.5. Different Salient Variety Features Attended to by Raters with Similar Attitudes

In terms of raters' focus of scoring category within similar attitude groups that was explored by the verbal protocol study, the results support previous literature on the rating process (Brown, 2000; Brown *et al*, 2005; May 2006; Meiron, 1998; Orr 2002; Pollitt & Murray, 1996). The criteria used by raters to judge examinees' oral proficiency varied even within the same attitude group, though they may comment on the same scoring criteria. This study shows that the largest group of comments relate to vocabulary (22.3%), followed by grammar (19.7%), phonology (18.8%), comprehension (16.9%), degree of near nativeness (10.5%), fluency (6.4%) and level of English language learning (5.4%). Except for the category of degree of near nativeness and level of language learning, other categories overlapped some of those generalized in Brown's (2000) study which used entire IELTS oral interviews as stimulus. Categories that resulted from interaction such as comprehension of interviewer questions were not found in the current study. While an attempt was made to generalize rating tendency according to rater attitude group, it was difficult to find a consistent pattern within similar rater attitude groups; what is more, features that raters emphasized in their rating may overlap across different rater attitude groups. Some criteria may be more salient than others and the performance judged against one or two of the particular language behaviors; in other cases, different aspects of the criteria were used to make scoring judgments. All made the general observation in rating while keeping WE in mind challenging. Nevertheless, some noticeable patterns were observed although it should be borne in mind that there were exceptions which were not applicable to all raters from the same attitude groups. Raters with a positive attitude towards WE typically kept in mind the examinee status as language learners and evaluated their English speaking proficiency accordingly, though some focused more on particular linguistic categories. On the other hand, raters with negative attitude towards WE seemed to be more concerned with the success of the intended messages delivery with some tending to compare examinees with native speaker performance. This yields important implications for rater trainings. As discussed in chapter 2, language assessment and WE research do not advocate using the native speaker model for scoring judgment as native speakers may not be absolutely defined (Davies, 2003; Mesthrie & Bhatt, 2008) and do not necessarily outperform non-native speakers in testing situations(Hamilton *et al*, 1993). Despite that, the native speaker model seems to be favored by some raters in the negative attitude group and have apparently become a latent standard even if not used as a rating criterion in this study. These raters, including the accredited IELTS raters, either revealed a tendency to implicitly use the native speaker model to aid in scoring decisions, or openly point out that utterances did not conform to expected native-speaker performance via comments such as " non-native phrasing" and even more strongly " fossilized mistakes" as shown in chapter 5. This should draw attention of testing agencies, particularly IELTS, or rater trainers given their effort to highlight examinees' communicative competency rather than the proximity to native speaker performance. The rating scale (see Appendix B) used in this study adopted the IELTS rating category and followed the principles of IELTS's rating descriptors of public version (see https://www.teachers.cambridgeesol.org/ts/digitalAssets/114292_IELTS_Speaking_Band_Descriptors.pdf) to avoid using the expected native speaker model to guide scoring decision. This was further verified by the IELTS rater trainer in a workshop conducted at the UIUC in Spring 2012 that raters were trained not to evaluate examinees' speech against the native speaker model. Nevertheless, the verbal protocol study showed that the native speaker model was clearly an underlying rating criterion for some raters with negative attitude towards WE, suggesting that raters' disregarded the training and rating descriptors, and most importantly, undermined the scoring validity and inferences from test scores about examinees' speaking proficiency. This raises the question on how test scores are interpreted and what L2 speaking performance actually means. Despite IELTS rating scales being revised driven by empirical data and feedback from raters worldwide to increase the usability of the scales (Taylor, 2001), the tendency of raters towards the native speaker model may indicate the inadequacy of training. One approach to improve rater performance in discarding native speaker model can be considered. As verbal protocol reports revealed that raters with positive attitude generally commented on performance according to the expected levels of the language learners' ability, testing agencies may consider the following approaches to improve the fairness of scoring processes. First, raters' attitude tendency towards WE may be evaluated by means of the RAI developed in this study, and second, the verbal protocol approach may be used to elicit the rating processes of the expert and positive attitude raters and justify their scores. This can serve as a good example for other raters in reaching the final scoring decision. As Orr (2002) urges, "the raters should focus on what they *do* differently from the expert judges, not just on score differences" (p.153) (Italics in the original).

Other findings from the verbal protocol reports point to the diverse nature of rater judgment on the same performance. Nevertheless, the diversity does not necessarily result in variations in scores. A rich source of information is generated below.

1. *Variations in judgments not necessarily leading to variations in scores*. Raters varied in their judgments, but, consistent with previous research (Brown, 2000; Brown *et al*, 2005; May 2006; Meiron, 1998; Orr 2002; Pollitt & Murray, 1996), the variations are not necessarily reflected in the scores. Thus, one rater offered a score of 6 in sentence structure because an examinee ". . . tries to use correct grammar in speaking", while another gave the same score but noted "his tenses are not sustained and use of article is incorrect at times". Raters that provided similar interpretations on a rating category may give different scores. For example, two raters commented that an examinee's pronouns are "generally mixed up" but awarded scores of 5 and 7 respectively on sentence structure.

2. *Variations in conceptualizing rating categories.* Raters did not treat the scoring criteria in the same way and had their own ways of interpreting on specific criterion. Thus, while discussing vocabulary, raters' comments on specific words or phrases, strategy in use of words, and vocabulary itself ranged widely. In terms of some raters' underlying criterion to judge against native speaker model that was not explicitly fully addressed in similar verbal

protocol studies, it was found that the difference between an examinee's linguistic use and the native speaker model was attributable to "simply different" or errors. While language testing researchers (Davies *et al*, 2003) are concerned that raters' different interpretation may reflect their rating decisions, it was not entirely the case in the results of this study. One rater classified in the low perception group expressed a strong stance similar to interlanguage theory (Selinker, 1974) that adult L2 learners are most likely to make fossilized mistakes and cannot attain complete target language grammar after a certain critical period. She commented on an examinee's sentence structure revealing influences from L1, stressed the difficulty in breaking the fossilized mistakes, and expressed its inappropriate use in formal settings. Nevertheless, her ratings on this examinee on four of the rating criteria were higher than the group mean score (*N* =96), except for the pronunciation category which she did not attribute to L1 influence and for which she assigned the lowest score among the eight interviewee raters. On the other hand, raters that associated the discrepancy between examinee variety and native speaker model as being different did not necessarily give lower ratings on the commented category. Another rater also with low perception in WE as measured by the RAI gave higher scores than the group mean score on the category of vocabulary when revealing comparisons against the native speaker model and described the examinee's use of vocabulary simply as "non-native phrasing".

This finding has an important implication. Despite the tension in assessing spoken L2 proficiency against the native speaker model (Bachman & Savingnon, 1986; Barnwell, 1989; Hamilton *et al*., 1993), which was evidenced earlier, rater variations in interpreting the difference, either treated as fossilized mistakes or simply different from the native speaker model, were found not to be associated with the scores they assigned. Both raters in the example rated leniently which may be best explained by raters' own severity of rating. Another possibility is the compensation of overall comprehensibility of the speech leading raters to be lenient on the specific scoring category, as inductively generated in the current verbal protocol study. This supports comments about intelligibility in WE research that a higher level of understanding (with regard to comprehensibly and interpretability) is most crucial in cross-cultural communication even if some utterances are not entirely intelligible (Smith & Christopher, 2001; Y. Kachru, 2008). Nevertheless, future studies are encouraged to further investigate whether the native speaker model if an underlying rating criterion, no matter how raters interpret it, is associated with the scores rater award.

3. *Variations in scoring judgment being minimized probably due to rater training.* In addition to evidence provided above, cross examination of attitude-rating relationship was not marked. Besides the possibility of inherent leniency in the rater, the inconsistency between attitude and rating behavior may be justified by raters' awareness of being raters and did not allow their own underlying judging criterion, for example, the native speaker model, to be activated and affect their ratings (cf, Ajzen & Fishbein, 2005, p.182). This is most likely the result of rater training to minimize subjectivity to scoring decisions and reduce rater severity or leniency (McNamara & Adams, 1991; Lumley and McNamara, 1995; McNamara, 1996; Weigle, 1998) and "making raters more self-consistent" (McNamara, 1995). Another possibility is that his attitude towards World Englishes is not activated (Fazio, 1986, 1990, 1995; Fazio & Towles-Schwen,1999) perhaps due to fatigue or low motivation (e.g., being a rater simply to gain extra money), which led to inconsistency between attitude and behavior. All suggests the prediction power of rater attitude to scoring behavior may vary across individual raters.

4. *Little association between the Indian/non-Indian variable and rating focus*. Very

little association can be established based on rater nationality and salience of variety features they attend to. One Indian rater provided a balanced account of evaluations on different aspect of examinee's language use, which is also reflected in the comments by a few non-Indian raters. The most distinctive difference between this Indian rater and other non-Indian raters is the evaluative comments on the effect of examinee's language use as a conclusive remark on each of the speech performance as in "he is a good user of English language", suggesting her stress on effective use of language as an overarching factor in score judgment.

In sum, based on the quantitative data, the findings seem to suggest that rater attitude towards WE is a relative steady and group-based construct. More unexpected results emerged in the qualitative inquiry revealing diverse and dynamic views to expand our understanding that effects of attitude may not be manifested in rater rating behavior. This also shows the difficulty to isolate elements of attitude and other affecting factors, such as rater own severity, examinee's own speaking proficiency and task difficulty from the final scoring judgment.(Chalhoub-Deville, 1995; Gass, Mackey, Alvarez-Torres, & Fernandez-Garcia, 1999; Iwashita, McNamara, & Elder, 2001; Wigglesworth 2001). McNamara (1996) made a reflective statement regarding maximum testing professional's endeavor to score use and interpretation:

> "We must remain skeptical about the meaning of our test scores, and do everything we can to improve our understanding of what they mean, in the interests primarily of fairness to the test candidates, but also of the informativeness of our reports on candidates to test users" (p.246).

Guided by the post-Messick validation approach that is value- and social-laden, this study demonstrates potential undesirable evidence deriving from rater bias in WE. It is timely for language testing professionals and agencies to perceive second language speaking performance differently in the contemporary world by presenting defensible evidence to examinees in claiming measures have been taken to ensure that raters' bias in WE are reduced to the minimum. In the same vein, by seriously incorporating rater attitudes towards WE into all aspects of testing practice, testing professionals and agencies will demonstrate their social responsibility in bringing into greater focus the larger context and its impact on an integral part of the second language oral assessment agenda.

Suggestions and Implications

Discussions on the impact of WE in second language assessment over the past few decades have focused more on the theoretical rather than practical (Xi, 2010). This section suggests three practical guidelines for testing professionals and practitioners to engage the research and practice of English language oral assessment within the WE context in a more systematic fashion and to help drawing comparable interpretations and meaningful discussions for the future research. The guidelines relate to (1) constructs of WE within second language oral assessment research; (2) modifications of speaking test performance; and (3) test fairness design. These attempts seek to enhance interpretations and comparisons of research findings, allow for meaningful discussions between researchers and encourage the emergence of insightful empirical testing projects in the near future.

1. *Initial construct definition of World Englishes within second language oral assessment research*

What does WE mean when discussed within the context of second language oral proficiency tests? Prior to answering this question, the way we look at testing second language speaking (Fulcher, 2003; Skehen, 1998) may be re-considered when the 'second' may imply second first, foreign language or World Englishes. This re-conceptualization involves the inappropriate use of idealized native speaker model as the underlying judgment criteria and points to a need for a state-of-art definition of WE within second language speaking assessment research. As noted in chapter 2, Literature Review, WE's distinctive references to linguistic features, such as phonology, morphology, sentence structure and non-linguistic features including cultural norms, communication styles and literature styles, have been well researched. In fact, the findings of this dissertation demonstrate that each category exerts either no or varying amounts of influence on raters' judgment of examinee's oral proficiency performance, thus suggesting a need for a clearer construct definition of WE-oriented second language assessment to facilitate meaningful discussions among researchers. According to this study's findings, raters' judgments on English speaking proficiency focused on the examinees' salient language use particularly in v*ocabulary*, *pronunciation, sentence structure*, and *fluency*, reflecting the language competence described in Bachman's (1990) communicative language competency model and Fulcher's (2003) framework of speaking constructs. These are in fact legitimate categories because they are what constitutes language and make varieties different from each other. Language testing research concerning WE has shown research agenda investigating these linguistic categories in relation to score impact, including variations in morphosyntactic structure in the reading test (Lowenberg, 2002), sentence level in the writing test (Kenkel & Tucker, 1989) and pronunciation in the listening tests (Hardings, 2008). The other three non-linguistic categories as found in this study probably contribute more to the understanding of rater orientation toward WE and should be incorporated into the construct constitution: *comprehensibility of the speech, native speaker model* and *level of English education.* Comprehensibility as argued a crucial element in cross-cultural communication (Smith & Christopher, 2001; Y. Kachru, 2008) was noted to be a deciding factor in raters' scoring decisions. The findings of the verbal protocol study further suggest that raters' comprehension of examinees' speech is likely to compensate for their less proficient performance in individual linguistic categories, indicating its overriding effect on rater judgment. Furthermore, the fact that raters explicitly or implicitly fall back on the native speaker model to judge speaking performance reflects the extent of its permeation in testing practice and this will continue to generate theoretical debates (Davies, 2003). Thus, any testing research related to WE may not avoid addressing the impact and role of nativeness into inquiry and should justify the inclusion or exclusion of the native speaker model to assess examinee performance, for example, in the rating scale construction or any decision-making, such as the debate over norm selection between WE and standard English perspectives. Either view defends its perspectives by using the native speaker model as a starting point for argument. Finally, when speech comprehension was impeded as a result of expressions or speech style highly associated with variety, it was frequently attributed to the quality of English language education received by examinees. It indicates that second language oral assessments should not ignore the quality of English language education the examinees may have received either in or outside their home country and incorporate their English language education into any decision-making. Rather than treating all examinees equally as having acquired or learnt English language in the same manner and at the same pace, the element of language education should be incorporated into any decision-making process, for example, the language proficiency test scores required for admission into U.S. universities.

Having said that, the current proposal of the construct definition of WE within second language oral assessment research should be interpreted with caution. Given that the existing data that did not take into account the examinee's interaction ability, the proposed construct definition is better only applied to the descriptive task at this stage. Other functional strategies, such as negotiation, as an important communication strategy in cross-communication interaction, may be further investigated to evaluate their effects on rater scoring judgment and if necessary, to further expand the current construct definitions.

2. *Modifications of speaking test performance*

The model of speaking test performance for examining variables that interact in the test process and impact the final test scores as put forward by Skehen (1998) and expanded by Fulcher (2003) is now further refined. Rater attitude as this dissertation illustrates is a potential biasing factor that affects scoring decisions. Nevertheless, compared to other background variables, the positive aspect of attitude is that it can be monitored and changed over time, and bias can be converted into a neutral or, even more, an accepting and positive stance towards examinee varieties. Thus, almost a decade after Fulcher's (2003) model of speaking test performance was proposed and cited in speaking test performance research (Brooks, 2009; Bygate, 2009; Davies, 2009; Lazaraton, 2008; Lee, 2005; May; 2009; Segalowitz, 2010;Taylor, 2009; Tavakoli, 2009; Weir, 2005), there is a continuing need to factor in the psychological traits of raters into the L2 speaking performance model to highlight the potential impact they exert on the scoring process and to be systematically and carefully monitored as an essential part of rater training. As displayed in Figure 31, rater attitude merits a key position on the speaking test performance model, urging testing researchers to keep in mind the social dimensions of language assessment. They should be aware of the unintended test consequences arising from raters' preference for the examinee's variety and highlighting that testing practice, and research has to remain abreast of the state-of- art issues resulting from the global English language spread. Equally important, it assures examinees that a fair scoring process is being carefully monitored from the very beginning of the speaking test, their own variety is recognized and respected for communicative purposes, either within or cross-culturally, and that raters are trained to broaden their WE knowledge and to fairly handle their scoring decisions.

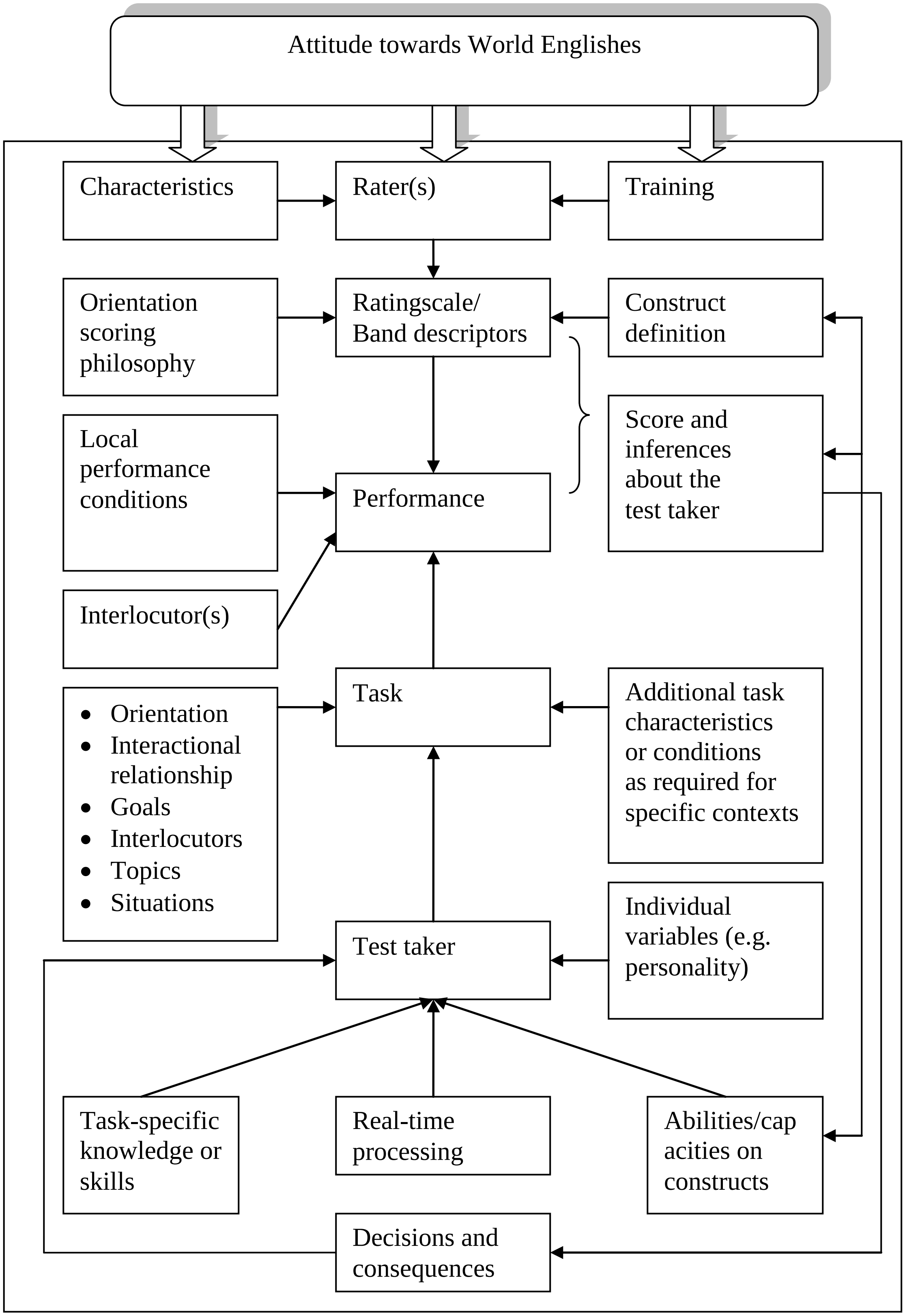

Attitude towards World Englishes
Characteristics
Rater(s)
Training
Orientation scoring philosophy
Ratingscale/ Band descriptors
Construct definition
Score and inferences about the test taker
Local performance conditions
Performance
Interlocutor(s)
Task
Additional task characteristics or conditions as required for specific contexts
Orientation
Interactional relationship
Goals
Interlocutors
Topics
Situations
Test taker
Individual variables (e.g. personality)
Task-specific knowledge or skills
Real-time processing
Abilities/capacities on constructs
Decisions and consequences

Figure 31. Expansion of Fulcher's (2003) model of speaking test performance.

3.Test fairness design

A practical approach to increase test fairness in response to potential bias by raters and subsequent scoring judgment is to not treat fairness as a form of checklist that is only checked till during the test validation study or after the test administration is completed. Inferences made about the examinee's speaking performance in the relevant real world context based on such a test score would be weakened if test fairness is not carefully monitored and ensured over the entire cycle of test development and administration. In ensuring fair scoring procedures, testing agencies can help raters increase certainty in handling and assessing L2 speaking performance in multiple variety situations by applying the test specification approach (Davidson & Lynch, 2002). Test specifications originally served as generative blueprints to document the constructs to be measured, the tasks selected to measure the constructs and the expected examinee response. In addition to these practical guidelines for test development, test specifications also record mandates shaping the test, constraints for test development and administration and feedback from stakeholders to improve the content of the test specification. The formulation of a test specification is an evolving process through a series of problem-solving activities and negotiations with stakeholders, which also serves as important validity evidence before a testing event, or termed "a priori validity evidence" (Weir, 2005, p.17). By applying the test specification approach, plans to increase rater awareness of WE, training of rater dealing with multiple varieties, any activity designed to reduce rater bias in examinee variety and issues raised by raters and other stakeholders can be documented in the test specification to guide rater training procedures. This will serve as powerful and defensible validity evidence to justify the testing agency's endeavors to ensure test fairness.

Limitations and Suggestions for Future Research

This study makes important theoretical contributions to the understanding of rater bias in World Englishes and its association with IELTS descriptive task scores. The limitations of this research should be taken into account when interpreting the study findings. First, the 2-factor rater attitude measurement model was established based on the conceptual 3-factor attitude structure constituting two different measurement scales and was directly tested by confirmatory factor analysis (CFA). A more rigorous and theoretical approach would perform exploratory factor analysis to examine the internal structure of the entire Rater Attitude Instrument (RAI) prior to CFA for further confirmation. However, this was not possible for the current data set due to insufficient sample size. Note that this limitation does not apply to the measure of rater feeling. The number of Likert scale items in the measure of rater belief and rating tendency was 35, which would need at least 175 raters to meet the minimum of a 5 subject-to-item ratio requirement (Costello & Osborne, 2005). Given the accessibility to raters and financial constraints at the time the study was conducted, it was unfeasible to reach this number. An alternative attempt at the measurement model was to use summated item scores across each of the seven sections that constituted the three attitude factors as indicators. This resulted in seven indicators that were underlined by three conceptual attitude factors. Though satisfying the sample size requirement, the summated scores had the disadvantage of an obscure rater response pattern: high scores on several items but low scores on others may reflect a moderate or neutral position. A similar score may occur to rater who expressed neither a positive nor negative position. There is thus a clear need to increase the number of raters in future research to further test the psychometric property of the RAI and to evaluate whether the current measurement model is defendable against a larger sample size. Alternatively, the number of the

RAI items may be further reduced for a more time-efficient and user-friendly scale and its internal structure needs to be compared against the current 2-factor measurement model.

To expand the scope of this study, future research may focus on different stakeholders and investigate their perspectives on the inclusion of WE varieties in the English oral proficiency assessment as language testing professionals have the responsibility to counsel stakeholders about test use and changes for the test. Stakeholders may include examinees and the score users, such as employers, university advisors and decision-makers. As Taylor (2006) has cautioned, “we must avoid acting as ‘liberators’ only to impose a new ‘bondage’ ” (p.52).

Relevant to the scale construction is the method of statistical analysis. This study relied principally on factor analysis to determine item appropriacy and was constrained by its demand on a large sample size. Future studies may consider using different statistical tools, such as FACETS analysis, multidimensional scaling analysis and structural equation modeling. An advantages of FACETS in evaluating scale structure is its relatively low demand on subjects needed to obtain reasonable estimates (Hambleton, *et al*, 1991) and has been applied in the instrument validation (Jackson, Draugalis, Slack, Zachry & D’Agostino, 2002).

In terms of rater groups, the findings suggest that Indian raters may score more harshly on certain rating criteria than non-Indian raters on Indian examinee’s speaking performance. It would be useful to recruit raters of other outer- and inner-circle varieties to verify whether raters sharing examinee’s variety tend to rate lower than raters who do not. Looking at the prediction power of the RAI on the speaking test score, it will be interesting to examine if similar results in the present study would apply to different rater groups. Improvements on other aspects of methodology design include a more balanced number between rater groups (e.g., Indian vs non-Indian) and varying lengths of ESL/EFL teaching experience (e.g., experienced vs naive teacher/MA TESOL students).

Another limitation of this study is the choice of stimulus. As only Indian English was used as stimulus the results need to be interpreted with caution as they may not apply to raters' attitude toward examinees of other WE varieties. Extending the current study using alternative stimuli, such as single outer- or expanding-circle varieties or a combination, would also provide insights into the generalizability of these findings. Also with respect to the stimulus aspect, growing discourse-based research on oral assessment (Lazaraton, 1992; Yoshida-Morise, 1998) suggests that future study may apply discourse analysis to investigate examinee responses and evaluate the extent to which distinctive features of examinees' varieties really matter in test scores.

This study used descriptive tasks as elicitation stimulus of rater attitude and rating performance. Given that task types may affect test scores (Chalhoub-Deville, 1995; Gass *et al*., 1999; Iwashita *et al*., 2001; Wigglesworth, 2001), further research may focus on interaction-oriented speech tasks or a combination of different task types, as in the entire IELTS speaking section, to seek comparable results and broaden understanding of how raters perceive examinee's communication strategies in relation to their attitude toward WE and, most importantly, to what extent it matters in the scores they award. This will also clarify whether the interaction strategies as highlighted in the cross-cultural communication by WE researchers are actually the case and feasible in the testing situation.

Finally, the findings of the verbal protocol study may have differed if different raters were selected. It would be worth the research effort to interview other raters to seek support or divergence for the present findings. Relevant to the qualitative study is the exploration of rater attitude formation and change to elicit insightful inner voices of raters' likeability of the WE examinees, to probe what factors shape attitudes other than those identified in this study and in the literature, and to examine what internal and external forces change their attitude to further enhance our understanding the power of attitude on rater scoring judgment.

Conclusions

The post-Messick test validity paradigm highlighted the social dimensions of the test to make it socially responsible, which will bring new perspectives on the value implications of the test as part of the test validation process. This study argues that test fairness encompasses the property of validation by looking at the extent to which raters' own bias or preference in examinee's variety of English affects test fairness. Bias tendencies are inherent, but when it is transmitted into real action towards the objects and causes harm, it needs to be investigated, if not to promote liking, to at least avoid negative consequences. This study established a Rater Attitude Instrument that captures and measures rater feeling, belief and rating tendency covering issues and concerns in language assessment relevant to WE. There is much evidence supporting the RAI as an adequate tool to serve its intended purpose to measure rater perception of WE examinees. Rater's scoring on IELTS descriptive tasks presented by Indian English examinees revealed the expected but unfortunate results: that rater attitude is significantly related to the scores awarded. Raters with positive attitudes consistently rated higher than their negative counterparts on all rating criteria. Neutral and negative groups also have significant rating differences on certain rating criteria. It was found that the RAI is a significant predictor of scores awarded by raters, accounting for at least 17.5% of variance in the total and each of the analytic scores. It was also noted that the ratings by Indian and non-Indian raters differed significantly on certain criteria. The rater verbal protocol study revealed that linguistic and non-linguistic features that are associated with Indian English affected rater scoring judgment. This study elicited insightful and diverse perspectives on a mix of attitude-behavior relationships implying that ratings involve a complex host of decision making processes and that behavior tendencies towards attitude may be countered or suppressed by rater training or the rater’s own sense of the need to minimize subjective elements in the rating.

This study contributes to testing literature about additional potential rater biasing factors, adding more rater variability in any decision making process and affecting the test score use and interpretation. For ideological and pragmatic reasons, language testing professionals have to consider what does English speaking assessment mean, what drives test fairness to the maximum, what the constraints are that weaken it, and what to do to strike a balance between respecting the socio-cultural identities of the examinees and maintaining test integrity. As found in this study, raters’ inclination towards standard English in judging speaking performance is an important consideration for delivering WE-oriented English speaking tests. Other stakeholders’ views in WE may be further investigated to justify any change and test use.

Language testing is broadening its scope as a result of collaboration with other disciplines. Future research linking language assessment, World Englishes and language attitude studies is needed to better define the constructs of L2 speaking proficiency, develop appropriate assessment criteria and implement assessment training programs.

.

REFERENCE

Aiken, L. R. (1996*). Rating scales and checklists: Evaluating behavior, personality, and attitudes.* New York: John Wiley.

Ajzen, I. & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.

Ajzen, I. & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracı´n, B. T. Johnson, & M. P. Zanna (Eds.), *Handbook of attitudes and attitude change* (pp. 173–221). Hillsdale, NJ: Erlbaum.

Ajzen, I. & Timko, C. (1986). Correspondence between health attitudes and behavior. *Basic and Applied Social Psychology, 7(4),* 259-276.

Albarracin, D., Johnson, B.T., Fishbein, M., & Muellerleile, P.A. (2001). Theories of Reasoned Action and Planned Behavior as Models of Condom Use: A Meta-Analysis. *Psychology Bulletin, 127,* 142-161.

Albarracin, D. Johnson, B.T, & Zanna, M.P (Eds) (2005). *Handbook of attitudes.* Mahwah, NJ: Erlbaum

Alderson, J.C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14(2),* 115-129.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). Standards for Educational and Psychological Testing. Washington, DC.

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1–34.

Bachman, L. F. & Palmer, A. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modem Language Journal, 70,* 380-390.

Barnwell, D. (1989).‘Naive’ native speakers and judgments of oral proficiency in Spanish. *Language Testing 6 (2),* 152–63.

Barnwell, D. (1989). Naïve’ native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6(2),* 152–163.

Barona, D.B. (2008). Native and non-native speakers' perceptions of non-native accents. *Language and Literature Journal, 3(2).* Retrieved March 20, 2011 from http://ojs.gc.cuny.edu/index.php/lljournal/article/viewArticle/430/428

Beadnell, B., Baker, S., Gillmore, M.R., Morrison, D., Huang, B., & Stielstra, S. (2008). The Theory of Reasoned Action and the Role of External Factors on Heterosexual Men's Monogamy and Condom Use. *Journal of Applied Social Psychology, 38(10),* 97-134.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin 107,* 238-46.

Bernreuter, R.G. (1933). Validity of the personality inventory. *Personality Journal, 11,* 383-386.

Berns, M. (2008). World Englishes, English as a lingua franca, and intelligibility. *World Englishes, 27(3/4),* 327-334.

Bhatt, R.M. (2001). World Englishes [Electronic version]. *Annual Review of Anthropology, 30,* 527-550.

Bolton, K. (2005). Where WE standards: approaches, issues, and debate in World Englishes. *World Englishes, 24(1),* 69-83.

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.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: co-constructing a better performance. *Language Testing 26(3),* 341-366.

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. (2000). An investigation of rater's orientation in awarding scores in the IELTS interview. In R. Tulloch (Ed.), *IELTS research reports, 3* (pp.1-19).

Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of Rater Orientations and Test Taker Performance on English for Academic Purposes Speaking Tasks*. (Monograph Series MS-29). Princeton, NJ: Educational Testing Service.

Brown, J. B. (2004). 'What do we mean by bias? Englishes, Englishes in testing, and English language proficiency?' *World Englishes, 23(2),* 317-319.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice-Hall.

Brutt-Griffler, J. (2002). *World English: A Study of its Development*. Clevedon: Multilingual Matters.

Bygate, M. (2009). Teaching and testing speaking. In M.H. Long & C.J. Doughty (Eds), *The Handbook of language teaching* (pp.412-440). Oxford: Blackwell Publishing Ltd.

Byrne, B. M. (2001). *Structural equation modeling with AMOS*. Mahwah, NJ: Erlbaum.

Canagarajah, S. (2006). Negotiating the local in English as a lingua franca. *Annual Review of Applied Linguistics, 26,* 197-218.

Caracelli, V.J. & Greene, J. (1997). Crafting mixed-methods evaluation design. In J.C. Greene, & V.J. Caracelli (Eds), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms* (pp.19-32). San Francisco, Jossey-Bass.

Cargile, A. C., & Bradac, J. J. (2001). Attitudes toward language: a review of speaker-evaluation research and a general process model. In W. B. Gudykunst (Ed.), *Communication yearbook 25* (pp. 347-382). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cargile, A. C. & Giles, H.(1998). Language attitudes towards varieties of English: an American-Japanese context. *Journal of Applied Communication Research 26,* 338-356.

Cargile, A. C., Giles, H., Ryan, E.B., & Y Bradac, J.J.(1994). Language attitudes as a social process: A conceptual model and new directions. *Language & Communication, 14*, 211-236.

Carranza, M., & Ryan, E. (1975). Evaluation reactions of bilingual Anglo and Mexican-American adolescents toward speakers of English and Spanish. *International Journal of the Sociology of Language 6,* 83-104.

Chapelle, C.(1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 254-272.

Chapelle, C. , Enright, M. & Jamieson, J (2008). *Building a Validity Argument for the Test of English as a Foreign Language.* Mahwah, NJ: Lawrence Erlbaum.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12(1),* 16–33.

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

Christ, T. J., & Boice, C. H. (2009). Rating scale items: A brief review of nomenclature, components and formatting. *Assessment for Effective Intervention, 34,* 242-250.

Cluver, A.D. (2000). Changing language attitudes: the stigmatization of Khoekhoegowap in Nmibia. *Language Problems and Language Planning, 24(10),* 77-100.

Cohen, A. & Upton, T. (2007). ‘ I want to go back to the test’: Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24(2),* 209-250.

Coniam, D. (2010). Validating onscreen marking in Hong Kong. *Asia Pacific Educational Review, 11,* 423-431.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly, 33(2),* 185-209.

Costello, A & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation, 10(7,)* 1-9.

Crane, G. (1994). The English language in Brunei Darussalam. *World Englishes, 13(3),* 351-360.

Crystal, D. (1997). *English as a Global Language*. Cambridge: Cambridge University Press.

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

Davidson, F. (1993). Testing English across cultures: summary and comments. *World Englishes, 13(1),* 113-115.

Davidson, F. (2006). World Englishes and test construction. In B, Kachru, Y. Kachru & C. Nelson (Eds.), *The Handbook of World Englishes* (pp.709-717). Malden, MA: Blackwell Publishes Ltd.

Davidson, F., & Lynch, B. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. Yale University Press.

Davies, A. (2003). *The native speaker: myth and reality*. Clevedon: Multilingual Matters Ltd.

Davies, A. (2006). ILTA code of practice background discussion material. Retrieved November 12, 2006, from http://www.iltaonline.com/ILTA-COP-Disc.pdf ILTA: Draft code of practice: Version 3. (2005, June 21). Retrieved November 12, 2006, from http://www.iltaonline.com/ILTA-COP-ver3-21Jun2006.pdf

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003). Whose norms? International proficiency tests in English. *World Englishes, 22(4),* 571-584.

Davies, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26(3),* 367-396.

DeCoster, J. (1998). Overview of Factor Analysis. Retrieved November 11, 2011 from http://www.stat-help.com/notes.html

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

Derwing, T. & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39 (3),* 379-397.

Deterding, D. & Kirkpatrick, A. (2006). Emerging South-East Asian Englishes and intelligibility. *World Englishes, 25(3),* 391-409.

De Vaus, D. (2002). *Analyzing social science data: 50 key problems in data analysis*. Los Anageles, CA: Sage.

DeVellis, R. F. (2003). *Scale Development: Theory and Applications*. Thousand Oaks, CA: Sage

Dillman, D.A., Phelps, G., Tortrora, R., Swift, K., Kohrell, J., Berck, J. & Messer, B (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research, 38,* 1-18.

D'Souza, J. (2001). Contextualizing range and depth in Indian English. *World Englsihes, 20(2),* 145-159.

Ducasse, A. & Brown, A. (2009). Assessing paired orals: Rater's orientation to interaction. *Language Testing, 26,* 423–443

Eagly, A.H., & Chaiken, S. (1993). *The psychology of attitude*. Fort Worth, TX: Harcourt Brace Jovanovich.

East, M. (2007). Bilingual dictionaries in tests of L2 writing proficiency: do they make a difference? *Language Testing, 24(3),* 331-353.

Educational Testing Service (2003). ETS fairness review guidelines. Princeton, NJ: Author.Elder, C. & Harding, L. (2008). Language testing and English as an International language: constraints and contributions. *Australian Review of Applied Linguistics, 31(3).* 34.1-34.11.

Elder, C. & Davids, A. (2006). Assessing English as a Lingua Franca. *Annual Review of Applied Linguistics, 26,* 282-301.

Fazio, R. H. (1986). " How do attitudes guide behaviors?" In R.M. Sorrention & E.T. Higgins (Eds), *The handbook of motivation and cognition*: foundations of social behavior (Vol.1, pp.204-243). New York: Guilford Press.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 75–107). NewYork: Academic.

Fazio, R. H. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247–282). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Fazio, R.H., Powell, M.C., & Williams, C.J. (1989). The role of attitude accessibility in the attitude-to-behavior process. *Journal of Consumer Research, 16(3),* 280-288.

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology* (pp. 97-116). New York: Guilford.

Fazio, R.H. & Zanna, M.P. (1977). On the predictive validity of attitudes: The roles ofdirect experience and confidence. *Journal of Personality, 46(2),* 228-241.

Firth, A., (1996). The discursive accomplishment of normality. On 'lingua franca'English and conversation analysis. *Journal of Pragmatics, 26,* 237–259.

Firth, A. & Wagner, J. (1997). On discourse, communication, and (some)fundamental concepts in SLA research. *Modern Language Journal, 81,* 285–300.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: an introduction to theory and research*. Reading, MA: Addison-Wesley Pub.

Fishman, J. A. (1971*). Sociolinguistics*: A brief introduction. Boston: Rowley.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal, 57 (4),* 325-334.

Field, J. (2005). Intelligibility and the listener: the role of lexical stress. *TESOL Quarterly, 39(3),* 399-423.

Friederici A.D, Kotz S.A, Scott S.K, Obleser J. (2010). Disentangling Syntax and Intelligibility in Auditory Language Comprehension. *Human Brain Mapping 31*, 448–457.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment*. London and New York: Routledge.

Graddol, D. 1997. *The Future of English?* London: British Council.

Garrett, P., Coupland, N., & Williams, A. (2003). *Investigating Language Attitudes*. Cardiff: University of Wales Press.

Garson, D. G. (2008). Factor Analysis: Statnotes. Retrieved January 22, 2011, from North Carolina State University Public Administration Program, *http://www2.chass.ncsu.edu/garson/pa765/factor.htm*.

Gass, S. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NY: Lawrence Erlbaum.

Gass, S., Mackey, A, Alvarez-Torres, M.J., & Fernandez-Garcia, M. (1999). The effects of task repletion on linguistic output. *Language Learning, 49(4),* 549-581.

Gass, S., & Varonis, E.M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning, 34(1),* 65-89.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Giles, H., & Billings, A. C. (2004). Assessing language attitudes: Speaker evaluation studies. In A. Davies & C. Elder (Eds.), *The handbook of Applied Linguistics* (pp. 187-209). Malden, MA: Blackwell.

Giles, H., Katz , V., & Myers, P. (2006). Language attitudes and the role of community infrastructure: A communication ecology model. *Moderna Sprak, 100,* 38-54.

Graddol, D. (1999). The decline of the native speaker. In D. Graddol & U. Meinhof (Eds), *English in a changing world* [AILA Review 13]. United Kingdom, 57-68.

Graddol, D. (2001). English in the future. In A. Burns and C. Coffin(eds*), Analyzing English in a global context: A reader* (pp.26-37). New York: Routledge.

Gramley, Stephan & P¨atzold, Kurt-Michael (2004). *A Survey of Modern English. 2nd edition.* London: Routledge.

Grant, J. & Davis, L. (1997). Selection and use of content experts for instrument development. *Research in Nursing and Health, 20,* 269-274.

Green, A. (1998). *Verbal Protocol analysis in language testing research: A handbook (Vol. 5).* Cambridge: Cambridge University Press

Greene, J. C. (2007) *Mixed methods in social inquiry*. San Francisco: Jossey-Bass.

Greene, J., Caracelli, V., & Graham, W. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11(3),* 255-274

Griego-Jones, T. (1994). Assessing students' perceptions of biliteracy in two-way bilingual classrooms. *Journal of Educational issues of Language Minority Students, 13,* 79-93.

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

Hamilton, J., Lopes, M., McNamara, T. F., & Sheridan, E. (1993), Rating scales and native speaker performance on a communicatively-oriented EAP test. *Language Testing, 10(3),* 337-353.

Hammerly, H. (1991). *Fluency and accuracy*. Clevedon, England: Multilingual Matters.

Harding, L. (2008). *The use of speakers with L2 accents in academic English listening assessment: A validation study.* An unpublished doctoral dissertation. The University of Melbourne, Australia.

Hatch, E. M. & Lazaraton, A. (1991). *The research manual: design and statistics for applied linguistics.* New York: Newbury House.

Henson, R.K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34,* 177-189.

Higgins, C. (2003). "Ownership" of English in the outer circle: an alternative to the NS-NNS dichotomy. *TESL Quarterly, 37(4),* 615-644.

Hohenthal, A. (2003). English in India: Loyalty and Attitudes. Language in India, 12,3, May2003. http://www.languageinindia.com Accessed on 22nd March 2011.

House, J. (1999). Misunderstanding in intercultural communication: Interactions in English as a lingua franca and the myth of mutual intelligibility. In C.Gnutzmann (Ed.), *Teaching and learning English as a global language* (pp.73–89). Tubingen: Stauffenburg.

House, J. (2002). Pragmatic competence in lingua franca English. In K. Knapp & C. Meierkord (Eds.), *Lingua franca communication* (pp. 245–267). Frankfurt:Lang.

House, J. (2003). English as a lingua franca: A threat to multilingualism. *Journal of Sociolinguistics, 7(4),* 624–630.

Hrubes, D., Ajzen, I., & Daigle, J. (2001). Predicting hunting intentions and behavior: An application of the theory of planned behavior. *Leisure Sciences, 23(3),* 165-178.

Hsu, HL (2007, June). *The impact of World Englishes on rater judgment.* Work-in-Progress Presented in the Language Testing Research Colloquium, Barcelona, Spain.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation modeling, 6,* 1-55.

Iwashita, N., McNamara, T. & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning 21,* 401–36.

Jackson, T., J. Draugalis, M. Slack, W. Zachry & J. D'Agostino. (2002). Validation of authentic performance assessment: a process suited for Rasch modeling. *American Journal of Pharmaceutical Education, 66,* 233-243.

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford: Oxford University Press.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23,* 83-103.

Jenkins, J. (2006). Current perspectives on teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly, 40(1),* 157-181.

Jenkins, J. (2007). *English as a lingua franca: attitude and identity*. Oxford: Oxford University Press.

Jianda, L. (2007). Developing a pragmatics test for Chinese EFL learners. *Language Testing, 24(3),* 391-415.

Johnson, M. (2001). *The art of nonconversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Joint Committee on Testing Practices (1988). Code of fair testing practices in education. Washington, DC: Author.

Joint Committee on Testing Practices (2004). Code of fair testing practices in education. Washington, DC: Author.

Kadir, K. (2008). *Framing a validity argument for test use and impact: the Malaysian public service experience*. Unpublished doctoral dissertation, University of Illinois at Urbana-Chamapign, Urbana, IL.

Kachru, B. (1985). Standards, codification and sociolinguistic realism: the English language in the outer circle. In R. Quirk, & H.G. Widdowson (Eds), *English in the world: teaching and learning the language and literatures* (pp. 11-30). Cambridge: Cambridge University Press.

Kachru, B. (1992). *The Other Tongue (2nd edition*), Urbana: University of Illinois press.

Kachru, B. (1997). World Englishes and English-using communities. *Annual Review of Applied Linguistics, (17),* 66-87.

Kachru, B. (1998). *English as an Asian Language*. Links & Letters. 89-108.

Kachru, B. (2008). Symposium on intelligibility and cross-cultural communication in world Englishes: introduction. The first step: the Smith paradigm for intelligibility in world Englishes. *World Englishes, 27(3/4),*293-296.

Kachru, B., & Nelson C. (2001). World Englishes. In A. Burns & C. Coffin (Eds), *Analyzing English in a global context: A reader* (pp.9-25). New York: Routledge.

Kachru, B., Y. Kachru & C. Nelson (2006). *The Handbook of World Englishes*. Oxford: Blackwell.

Kachru, Y. (2005). Teaching and learning of World Englishes. In E. Hinkel (Ed.), *Handbook of research in second language learning and teaching* (pp. 155–173).Mahwah, NJ: Lawrence Erlbaum.

Kachru, Y. (2008). Cultures, contexts, and interpretability. *World Englishes, 27(3/4),* 309-318.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112(3),* 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21(1),* 31-41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. Measurement: *Interdisciplinary Research and Perspectives, 2,* 135-170.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18(2),* 5-17.

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.

Kenkel, J. M., & Tucker, R. W. (1989). Evaluation of institutionalized varieties of English and its implications for placement and pedagogy. *World Englishes, 8(2),* 201-214.

Kioko, A.N., & Muthwii, M.J. (2003). English variety for the public domain in Kenya: speakers' attitudes and views. *Language, Culture and Curriculum, 16(2),* 130 - 145

Kim, H.J. (2005). *World Englishes and language testing: the influence of rater variability in the assessment process of English oral proficiency*. Unpublished doctoral dissertation. University of Iowa, Iowa city, IA.

Kim, H.J. (2006). World Englishes in language testing: a call for research. *English Today, 22(4),* 32-39.

Kim, J.Y.(2008). *Development and validation of an ESL diagnostic reading-to-write test: an effect-driven approach.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

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.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61,* 213–218.

Kirkpatrick, A. (2007). *World Englishes: Implications for international communication and English language teaching*. Cambridge: Cambridge University Press.

Kline R. B. (2005). *Principles and practice of structural equation modeling (2nd ed.)*. Guilford Press, New York.

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

Kortmann, B. & Schneider, E.W. (Eds.). (2005*). A Handbook of varieties of English: A multi-media reference tool*. Berlin & New York: Walter de Gruyter.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (ed), *Fairness and validation in language assessment* (pp. 1-15). University of Cambridge Local Examinations Syndicate.

Kunnan, A. J. (2004). Test fairness. In M. Milanvoic, C. Weir, & S. Bloton (Eds.), *European year of language conference papers,* Barcelona (pp.27-48). Cambridge, UK: CUP.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27(2),* 183-189.

Lambert, W. E., Hodgson, R., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology, 60,* 44-51.

Lazaraton, A. (1992).The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373–86.

Lazaraton, A. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly 5(4),* 313-335.

Linacre, J. M. (2004). *A user's guide to FACETS: Rasch measurement computer programs.* [Software manual]. Chicago: Winsteps.com.

Lindemann, S. (2002). Listening with an attitude: A model of native-speaker comprehension of non-native speakers in the United States of America. *Language in Society, 31*, 419-441.

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics, 7 (3),* 348-364.

Lee, J. (2005). The native speaker: an achievable model? Asian EFL Journal, 7(2), Retrieved from http://www.asian-efl-journal.com/site_map_2005.php.

Lenski, G.E.,& Leggett, J.C.(1960). "Caste, Class, and Deference in the Research Interview." *American Journal of Sociology, 65 (March),* 463-67.

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

Llurda, E.(2004). Non-native speaker teachers and English as an International language. *International Journal of Applied Linguistics,14(3),* 314-323.

Llurda, E. (2009). Attitudes towards English as an international language: The pervasiveness of native models among L2 users and teachers. In F. Sharifian (Ed), *English as an international language: perspectives and pedagogical issues*. Bristol: Multilingual Matters.

Lowenberg, P.H. (1993). Issues of validity in tests of English as a world language: whose standards*? World Englishes, 12(1),* 95-106.

Lowenberg, P.H. (2002). Assessing English proficiency in the Expanding Circle. *World Englishes, 21(3),* 431-435.

Lumley, T. & McNamara, T (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12(1),* 54-71.

Luoma, S. (2004). *Assessing speaking*. New York: Cambridge University Press.

Lynch, B.K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18(4),* 351-372.

Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing 15,* 158–80.

MacCallum, R. C.,Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99.

Major, R.C., Fitzmaurice, S.F., Bunta, F. & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: implications for ESL assessment. *TESOL Quarterly, 36(2),* 173-190.

Matsuda, A. (2003). Incorporating World Englishes in teaching English as an international language. *TESOL Quarterly, 37,* 719-729.

Matsuura, H., Chiba, R.& Fujieda, M. (1999). Intelligibility and comprehensibility of American and Irish Englishes in Japan. *World Englsihes, 18(1),* 49-62.

May, L. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. *Language Testing 26(3),* 397-421.

May, L.A. (2006) An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11(1),* 29-51.

McArthur, T. (2001). World English and world Englishes: trends, tensions, varieties, and standards. *English teaching, 34,* 1-20.

McArthur, T. (2003). *The Oxford guide to world Englishes*. Oxford: Oxford University Press.

McKay, S. L. (2003). Toward an appropriate EIL pedagogy: re-examining common ELT assumptions. *International Journal of Applied Linguistics, 13(1),* 1-22.

McKenzie, R. M. (2008). Social factors and non-native attitudes towards varieties of spoken English: a Japanese case study. *International journal of Applied Linguistics, 18 (1),* 63-88.

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

McNamara, T. (1997). 'Interaction" in second language performance assessment: Whose performance? *Applied Linguistics, 18,*446-466.

McNamara, T. (1998). Policy and social considerations in language assessment. *Annual Review of Applied Linguistics, 18,* 304-319.

McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford: Blackwell.

Meierkord (2004). Syntactic variation in interactions across international Englishes. *English World-Wide, 25(1),* 109-132.

Meierkord, C. (2000). Interpreting successful lingua franca interaction: An analysis of nonnative-/ non-native small talk conversations in English [Electronic version]. Linguistik online, 5 (1/00). Retrieved July 20, 2010, from http://www.linguistik-online.com/1_00/index.html

Messick, S. (1989). Validity. In Linn, R.L.(Ed.), *Educational measurement (3rd ed.)* (pp.13-103). New York: American Council on Education & Macmillan.

Mesthrie, R., & Bhatt, R. (2008). *World Englishes: The Study of New Linguistic Varieties*. Cambridge: Cambridge University Press.

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing, 19(4),* 477-496.

Mislevy, R. (2004). Can there be reliability without "reliability"? *Journal of Eudcational and Behavioral Statistics*, 29, 241-244.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Educational Research, 62,* 229–258.

Moss, P. A. (2004). The meaning and consequences of "reliability". *Journal of Educational and Behavioral Statistics, 29(2),* 245–249.

Mueller, D.J. (1986). *Measuring Social Attitudes: A Handbook for Researchers and Practitioners.* New York: Teachers College Press.

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

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

Nelson, C. (1982). Intelligibility and nonnative varieties of English. In Kachru, B. (Ed.), *The other tongue: English across cultures* (pp. 58–73). Urbana, IL:University of Illinois Press.

Nelson, C. (1995). Intelligibility and world Englishes in the classroom. *World Englishes, 14,* 273–279.

Nelson, C. (2008). Intelligibility since 1969. *World Englishes, 27(3/4),* 297-308.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19* (2), 169-192

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System, 30(2),* 143-154.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Paltridge, J., & Giles, H. (1984). Attitudes towards speakers of regional accents of French Effects of regionality, age and sex of listeners. *Linguistiche Berirhte, 90,*7l-85.

Pennycook, A. (1994). *The cultural politics of English as an international language*. New York: Longman.

Pennycook, A. (2003). Global Englishes, Rip Slyme, and performativity. *Journal of Sociolinguistics, 7(4),* 513 – 533.

Pett, M., Lackey, N.R. & Sullivan, J.J. (2003). *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*, Sage, London.

Phillipson, R. (1992). *Linguistic Imperialism*. Oxford: Oxford University Press.

Pollit, A., & Murray, N.L. (1996). What raters really pay attention to. In M. Milanovic & Saville, N. (Eds.), *Studies in Language Testing 3: Performance Testing, Cognition and Assessment* (pp.74-91). Cambridge University Press, Cambridge.

Powers, D. E., Scheldi, M. A., Leung, S. W, & Butler, F. A. (1999). Validating the revised test of spoken English against a criterion of communicative success. *TOEFL Research Report, 99 (5),* 63.

Rajadurai, J. (2008). Revisiting the concentric circles: conceptual and sociolinguistic considerations. *Asian EFL Journal, 7(4),* 111-130.

Rajagopalan, K. (1999). 'Of EFL teachers,conscience, and cowardice'. *ELT Journal, 53(3),* 200–206.

Rampton, M. B. H. (1990). 'Displacing the "native speaker": expertise, affiliation, and inheritance'. *ELT Journal, 44(2),* 97-101.

Reddington, E. (2008). Native speaker response to non-native accent: a review of recent research. Retrieved from http://journals.tc-library.org/templates/about/editable/pdf/Reddington_AL.pdf

Reeves, J. (2006). Secondary teacher attitudes toward including English-Language learners in mainstream classrooms, *Journal of Educational Research, 99,* 131-142.

Regan, D.T. & Fazio, R.H. (1977). One the consistency between attitudes andbehavior: Look to the method of attitude formation. *Journal of Experimental Social Psychology, 13,* 28-45

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

Ryan, E. B., Carranza, M.A., & Moffie, R.W. (1977). Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech, 20(3),* 24-26.

Saif, S. (2007). Aiming for positive washback: a case study of international teaching assistants. *Language Testing, 23(1),* 1-34.

Schumacker, R.E & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling*. New Jsersy: Lawrence Erlbaum Associates.

Segalowitz, N. (2010). *Cognitive base of second language fluency*. New York: Routledge,

Seidlhofer, B. (2004). Research perspectives on teaching English as a Lingua Franca. *Annual Review of Applied Linguistics, 24,* 209-239.

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

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24(4),* 331-345.

Shohamy, E. (2001). *The Power of tests: a critical perspective on the uses of language tests.* England: Pearson Education Limited.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 167-185). Essex, UK: Longman.

Smith, L. E. (1992). Spread of English and issues of intelligibility. In B. B. Kachru (Ed.), *The other tongue: English across cultures* (2nd ed., pp. 75-90). Urbana: University of Illinois Press.

Smith, L.E., & Christopher, E. (2001) "Why can't they understand me when I speak English so clearly?" In Edwin Thumboo (Ed.), *The three circles of English: language specialists talk about the English language* (pp. 91–100). Singapore: UniPress.

Smith, L.E, & Nelson, C. (1985). International intelligibility of English: Directions and resources. *World Englishes, 3,* 333-342.

Spann, M. (2000). Enhancing fairness through a social contract. In Kunnan (Ed), *Fairness and Validation in Language Assessment* (pp. 35-39). University of Cambridge Local Examinations Syndicate.

Spatz, C. (2001). *Basic statistics: Tales of distributions (7th ed.).* Belmont, CA: Wadsworth.

Spolsky, B. (1993). Testing across cultures: An historical perspective. *World Englishes, 12(1),* 87-93.

SPSS (Version 17.0 for Windows) [Computer Software]. (2009). Chicago: SPSS Inc.

Standards for Educational and Psychological Testing. (1985). Washington, DC: American Psychological Association.

Steiger, J.H. & Lind, J.C. (1980). Statistically-based tests for the number of common factors. Paper presented at the annual Spring meeting of the Psychometric Society, Iowa City, IA.

Symonds, P. (1928). Factors influencing test reliability. *The Journal of Educational Psychology, 19(2),* 73-87.

Tabachnick, B.G., & Fidell, L.S. (1989). *Using multivariate statistics (2nd ed).* Harper Collins Publishers, New York.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches.* Applied Social Research Methods Series (Vol .46). Thousand Oaks, CA: Sage.

Tavakoli, P. (2009). Assessing L2 task performance: understanding effects of task design. *System 37(3),* 482-495.

Taylor, L. (2002). Assessing learners' English: but whose/which English(es)? *Research Notes 10.* Cambridge: University of Cambridge ESOL Examinations.

Taylor, L. (2006). The changing landscape of English: implications for language assessment. *ELT Journal, 60(1),* 51-60.

Taylor, L & Jones, N (2001). 'Revising the IELTS Speaking Test'. *EFL Research Notes, 4,* 9-12.

Taylor, L & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26(3),* 325-339.

Toulmin, S. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Trafimow, D., & Sheeran, P. (1998). Some tests of the distinction between cognitive and affective beliefs. *Journal of Experimental Social Psychology, 34,* 378-397.

Velicer,W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231-251.

Walters, S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24(2),* 155-183.

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

Weir, C., & Wu, J. (2007). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing, 23(2),* 167-197.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Houndmills, England: Palgrave Macmillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing 10,*305–35

Wigglesworth, G. (2001) 'Influences on performance in task-based oral assessment.' In M.Bygate, P. Skehan, & M. Swain (Eds), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 186-209). Longman.

Wilson, J., & Bayard, D. (1992). Accent, gender, and the elderly listener: Evaluations of NZE and other English accents by rest home residents. *Te Reo, 35*, 19-56.

Witz, K. (2006). The participant as ally and essentialist portraiture. *Qualitative Inquiry, 12(2),* 246-268.

Witz, K., Goodwin, D., Hart, R., & Thomas, H. (2001). An essentialist methodology in education-related research in-depth interviews. J. *Curriculum Studies. 33(2),* 195-227.

Wolfram, W. & Schilling-Estes, N. (2006). *American English: Dialects and Variation, 2nd edition*. Malden, Massachusetts/Oxford, U.K.: Blackwell.

Worthington, R. L. & Whittaker, T.A. (2006). Scale development research. A content analysis and recommendations for best practices. *The Counseling Psychologist, 34* ,806-838.

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22(4),* 463-508.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL academic speaking test (TAST) for operational use. *Language Testing, 24(2),* 251-286.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27(2),* 147-170.

Yoshida-Morise, Y.(1998). 'The use of communication strategies in language proficiency interviews'. In R. Young & A.W. He (Eds.), *Talking and testing: discourse approaches to the assessment of oral proficiency* (pp. 205-238). Philadelphia: John Benjamins.

Young, R.F. (2000). Interactional competence: challenges for validity. Paper presented at the Language Testing Research Colloquium, Vancouver, Canada.

Zahn, C. J & Hopper, R. (1985). Measuring language attitudes: The Speech Evaluation Instrument. *Journal of Language and Social Psychology. l4(2),*113-123.

Zahn, C. J & Hopper, R. (1985). Measuring language attitudes: The Speech Evaluation Instrument. *Journal of Language and Social Psychology. l4(2),*113-123.

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

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native *and* native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28(1),* 31-50.

APPENDIX A

INVITATION EMAIL

Dear (name of the ESL program director),

I'm a doctoral student in Educational Psychology at the University of Illinois at Urbana-Champaign (UIUC) and writing to you with regards to my dissertation study that looks into ESL teacher's views of varieties of English. I'm currently recruiting ESL instructors that have teaching experience for at least half a year and would like to invite the ESL instructors in the (name of the ESL program) to my study.

This study is approved by UIUC Institute Review Board. It takes approximately one hour and can be done from participant's own computer. They'll receive $15 remuneration upon completion of the study. If you could pass this recruitment info to the eligible ESL teachers in the (name of the ESL program), I'll be greatly appreciated.

****

1. Study description: an one-hour online study. You can do it from your own computer. You will receive $15 remunerate by check upon completion of the study.

2. Study procedures:

Step 1: Listen to several speech samples.
Step 2: Assign analytic scores to each sample.
Step 3: Complete a questionnaire.
Step 4: Listen to different speech samples.
Step 5: Assign scores to each sample.

3. Contact info: If you're interested in participating, please email me at hhsu9@uiuc.edu, I'll then send you a link to the study.

****

If you have any question, please feel free to email me. I'll be happy to talk more about my dissertation with you.

Many thanks,

Tammy, Huei-Lien Hsu

PhD candidate
Educational Psychology
Research Assistant of Foreign Language Assessment Group
University of Illinois at Urbana-Champaign

APPENDIX B

VARIETAL SPEAKER EVALUATION

A. Background information

Name: ______________________

Gender: Female[ ] Male [ ] (please circle)

First language(L1): ______________________

Country of birth: ______________________

I spoke to people from different L1 backgrounds:

Never Sometimes Often Everyday
[ ] [ ] [ ] [ ]

I have ________experience listening to the following accents(please tick):

	Little/no	some	extensive
American English	[ ]	[ ]	[ ]
British English	[ ]	[ ]	[ ]
Indian English	[ ]	[ ]	[ ]
Chinese English	[ ]	[ ]	[ ]
Korean English	[ ]	[ ]	[ ]
Singaporean English	[ ]	[ ]	[ ]
Japanese English	[ ]	[ ]	[ ]
Pakistan English	[ ]	[ ]	[ ]
Others(please identify)			

B. Listening task

<u>Instructions</u>

There are two tasks in total. For the first task, you will hear eight different speakers that talk about a place where they think the tourists should visit. Each speaker talks less than 25 seconds. Your task is to write down every word that you hear. You can listen to each talk for up to three times. If the talk is still unclear to you, make your best guess.

Then please rate each speaker's comprehensibility. Remember: by 'comprehensibility', I mean that you are able to understand what the speaker says without trying to guess what they try to say. Instructions for task 2 will be given at the end of the task 1.

APPENDIX B *(continued).*

<u>Now please click on "Varieties of English" folder and listen to Speaker 1.</u>

1. Please write down every word that you hear

2. How much to you understand the speaker?
Please rate the speaker's comprehensibility on the following scale:
(Remember: by 'comprehensibility', I mean that you are able to understand what the speaker says without trying to guess what they try to say)

Easy to understand								Difficult to understand
1 [ ]	2[ ]	3[ ]	4[ ]	5[ ]	6[ ]	7[ ]	8[ ]	9 [ ]

<u>Now please click on "Varieties of English" folder and listen to Speaker 2.</u>

1. Please write down every word that you hear

2. How much to you understand the speaker?
Please rate the speaker's comprehensibility on the following scale:

Easy to understand								Difficult to understand
1 [ ]	2[ ]	3[ ]	4[ ]	5[ ]	6[ ]	7[ ]	8[ ]	9 [ ]

This is the end of Task 1.
Now we begin Task 2-accent evaluation.

B. Accent evaluation

APPENDIX B *(continued).*

Instructions

In Task 2, you will listen to eight different speakers and rate each speaker's accent on a 9-point scale. Please listen to each speaker ONCE only.

Then please write 3-5 adjectives to describe your feelings of the way the speaker speaks English.

The adjectives are to complete the sentence, "The speaker sounds. . . ".
For example, clear, intelligent, unsure, happy, not fluent.

Now please click on "Accented speech " folder and listen to Speaker A.

How strong is the speaker A's accent?

I.) Please rate the speaker's accent on the following scale:

No accent								Strong accent
1[ ]	2[ ]	3[ ]	4[ ]	5[ ]	6[ ]	7[ ]	8[ ]	9[ ]

II.) Which accent do you think the speaker has? Please tick one only.

Chinese English [ ] Indian English [ ] Singaporean English[ ]

Korean English [ ] Japanese English [ ] Pakistan English [ ]

I don't know [ ] Others (please identify) ______________________________

III.) How does the speaker sounds when he/she speaks English? Use 3-5 adjectives to answer the question.

APPENDIX C

RATER ATTITUDE INSTRUMENT

Section A. Rater Background Information

Please place one X per question.

1.	Country of current residency	U.S. _______ U.K. _______ Others (please specify):
2.	Nationality	American _______ British_______ Indian_______ Others(please specify):
3.	Native language	English _______ Others (please specify) :
4.	Gender	Female _______ Male _______
5.	Year of teaching experience	Less than 1 year _______ 1-3 years_______ 4-6 years_______ More than 6 years _______
6	Highest level of education	Bachelor's_______ Master's _______ Doctoral_______

7 If you're an ESL instructor, what is your major of highest degree? (please specify):

If you're an ESL TA, what is your current major? (please specify):

Section B. Speaker Evaluation

In this section, you will **rate six Indian speech samples**. Each is 90 second long. The speeches are obtained from an oral proficiency test that assesses test-taker's English proficiency level to survive in English-medium universities. You'll serve as **a rater** to assess speaker's proficiency level.

Below are the instructions on rating:

Appendix C *(continued).*

Step 1. Read the following four rating criteria:

Fluency (i.e. The speaker is fluent in English),

Pronunciation (i.e. The speaker's pronunciation was easily understood)

Grammar (i.e. The speaker used sentence structure correctly)

Lexical range and accuracy (i.e. The speaker effortlessly selected appropriate vocabulary to express him/herself)

Step 2. Click on the folder, **"Section B. Speaker Evaluation"**, in the link that I sent you and then listen to the first sample. The topic of the speech is listed on the next pages.

For **"Speaker 4"**, please turn up the volume before listening.

Step 3. **Assign analytic scores from 0-9** to each criteria on the next pages.

Step 4. Repeat the steps above for the remaining five speech samples.

You may listen to the sample again, if necessary.

Appendix C *(continued).*

Speaker 1: describe an elderly person you know

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

Appendix C *(continued).*

Speaker 2: describe something useful that you've learnt recently

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

Appendix C *(continued).*

Speaker 3: describe a sports event you watched at a party

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

Appendix C *(continued).*

Speaker 4: describe a photograph you've seen (p.s. turn up the volume)

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

Appendix C *(continued)*.

Speaker 5: describe an interesting story you watched on TV

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

Appendix C *(continued).*

Speaker 6: describe an activity you most enjoyed doing when you were a child

	0	1	2	3	4	5	6	7	8	9	
The speaker is not fluent in English.											The speaker is fluent in English
The speaker's pronunciation was not easily understood.											The speaker's pronunciation was easily understood.
The speaker used sentence structure incorrectly.											The speaker used sentence structure correctly.
The speaker had difficulty selecting appropriate vocabulary to express him/herself.											The speaker effortlessly selected appropriate vocabulary to express him/herself.

This is the end of the rating task. Please turn to next page for a questionnaire.

Section C. Questionnaire

Appendix C *(continued).*

C.1 Expectation of Indian English

Instruction: You've just listened to 6 speech samples spoken by Indian speakers. What are your feelings for Indian English? Please place an X next to the number indicating your response per question.

1	2	3	4	5
Strongly Disagree (SD)	Generally Disagree (GD)	Neutral (N)	Generally Agree (GA)	Strongly Agree (SA)

		SD	GD	N	GA	SA	U
1	I have no problem understanding Indian speakers in non-test situations.	1	2	3	4	5	
2	Indian English has become a steady variety that carries its own distinctive linguistic features.	1	2	3	4	5	
3	I have experience in rating Indian test-takers.	1	2	3	4	5	
4	Indian speakers may be treated as native speakers of English nowadays.	1	2	3	4	5	
5	Indian speakers should not be exempted from English tests.	1	2	3	4	5	
6	I need to make more effort to understand Indian test-takers.	1	2	3	4	5	

If you have any comment, please write here.

Please scroll down to next page.

Appendix C *(continued).*

C.2 Personal Evaluation

Instruction: **Please recall what you just did when you rated the speech samples. Please place an X <u>next to</u> the number indicating your response per question.**

1	2	3	4	5
Strongly Disagree (SD)	Generally Disagree (GD)	Neutral (N)	Generally Agree (GA)	Strongly Agree (SA)

1. Varieties mainly refer to differences in (select all apply)

_____a. accent
_____b. sentence structure
____ c. vocabulary use
_____d. pragmatic use (i.e. intended use v.s. actual meaning)
_____e. communication styles

		SD	GD	N	GA	SA	U
2.	I think the differences between standard English and varieties of English, as selected in the previous question, are creative and just as correct as standard English.	1	2	3	4	5	
3	Test-takers do not need to speak like a native speaker in order for me to assign high scores.	1	2	3	4	5	
4	When test-takers use unfamiliar expressions, it decreases their intelligibility.	1	2	3	4	5	
5	I do not grade down test-takers that speak a variety, as long as they express themselves well.	1	2	3	4	5	

Appendix C *(continued).*

6	I do not penalize examinees who use negotiation strategies (e.g. asking for clarification, rephrasing) to achieve communicative goals.	1	2	3	4	5
7	When test-takers use less familiar expressions, it suggests that they have not fully mastered English yet.	1	2	3	4	5
8	The rater is not responsible for examinees' intelligibility.	1	2	3	4	5
9	I give high scores to test-takers that use expressions/idioms such as that used by the native speakers of English.	1	2	3	4	5

If you have any comment, please write here.

C.3 Interpersonal history

1	2	3	4	5
Strongly Disagree (SD)	Generally Disagree (GD)	Neutral (N)	Generally Agree (GA)	Strongly Agree (SA)

1. I have chances to speak English with people of different ethnic backgrounds (select all apply)

________a. in my neighborhood.
________b. at home
________c. in the workplace

		SD	GD	N	GA	SA	U
2	I feel comfortable listening to varieties of English.	1	2	3	4	5	

3	I can’t communicate well with people who speak a variety different from mine.	1	2	3	4	5
4	Use of varieties can cause cross-cultural misunderstandings.	1	2	3	4	5
5	English has evolved into different steady varieties.	1	2	3	4	5
6	I think features of varieties are developed in the same way as American English developed from British English.	1	2	3	4	5

C.4 Perceived Cultural Factor

1

2 3 4 5

Strongly Disagree (SD)	Generally Disagree (GD)	Neutral (N)	Generally Agree (GA)	Strongly Agree (SA)

		SD	GD	N	GA	SA	U
1	Standard English (e.g. British English or American English) should be used to judge test-takers’ performance in the test setting.	1	2	3	4	5	
2	Varieties of English are not appropriate to use in cross-cultural communication.	1	2	3	4	5	
3	Native speakers of English do not best serve as raters of oral English test (e.g. TOEFL, IELTS).	1	2	3	4	5	
4	Varieties of English are not appropriate in everyday communication.	1	2	3	4	5	

5 In the region where I live, I think the following variety should be taught in English as a second or foreign language classes (select all that apply):

Appendix C *(continued).*

________a. local English
________b. British English
________c. American English

________d. Other (please specify)

6	Language learners should develop an awareness of the global spread of English.	1	2	3	4	5
7	Unless varieties of English are promoted via educational efforts, such as by being codified in the dictionary, they can't obtain legal status and become standard.	1	2	3	4	5
8	Language learners should be exposed to different varieties of English.	1	2	3	4	5
9	Native speakers of English do not best serve as English language teachers.	1	2	3	4	5
10	Speakers of non-standard varieties (i.e., not British or American English) currently outnumber native speakers of standard English.	1	2	3	4	5
12	Raters of speaking tests (e.g. TOEFL, IELTS) should have opportunities to be exposed to varieties of English during training.	1	2	3	4	5
13	Raters of speaking tests (e.g. TOEFL, IELTS) should develop an awareness of the global spread of English.	1	2	3	4	5

If you have any comment, please write here.

This is the end of the questionnaire. Please proceed to the last section on the next pages.

Appendix C *(continued).*

Section D. How do you feel about each speaker?

Instructions.

You will hear 6 different speech samples produced by Indian speakers. Each sample was spoken for 90 seconds. Your task is to indicate how you feel about the speaker by responding to the scales on the next pages. There are no right or wrong answers.

The scales contain seven-points, and at the ends of each scale are two adjectives which are **exact opposites**.

Respond to the scales by placing a check mark (x) at one point on each of the scales to indicate your evaluation of the speaker on that trait.

For example:

If you think the speaker sounds very clear, you would place a check mark near the word "clear' on the scale:

		1	2	3	4	5	6	7	
1	Clear	:x	:	:	:	:	:	:	Unclear

If you think the speaker sounds fairly clear, you might place a check mark towards the center:

		1	2	3	4	5	6	7	
1	Clear	:	:	:	:x	:	:	:	Unclear

Please be careful as you respond, because **the positive and negative adjectives are not all on one side of the scale.** Make sure you read each adjective carefully when you mark your response on the scale.

You may respond as you listen to each speaker. Try to complete your responses within one minute after you have heard each speaker. You may listen to the sample again, if necessary.

Read the adjectives on each scale carefully
Place one check only on each scale
Be sure you place one check mark (x) on every scale
Work quickly through the items
Do not worry about individual items. It is your first impressions that are wanted

Appendix C *(continued).*

Speaker 1. (Please click on the folder, “Section D. How do you feel about each speaker” in the link that I sent you. Then click on “Speaker 1” and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

Appendix C *(continued).*

Speaker 2. (Please click on the folder, “Section D. How do you feel about each speaker” in the link that I sent you. Then click on “Speaker 2” and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

Appendix C *(continued).*

Speaker 3. (Please click on the folder, “Section D. How do you feel about each speaker” in the link that I sent you. Then click on “Speaker 3” and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

Appendix C *(continued).*

Speaker 4. (Please click on the folder, “Section D. How do you feel about each speaker” in the link that I sent you. Then click on “Speaker 4” and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

Appendix C *(continued).*

Speaker 5. (Please click on the folder, "Section D. How do you feel about each speaker" in the link that I sent you. Then click on "Speaker 5" and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

Appendix C *(continued).*

Speaker 6. (Please click on the folder, "Section D. How do you feel about each speaker" in the link that I sent you. Then click on "Speaker 6" and listen to the speaker).

The speaker sounds. . .

		1	2	3	4	5	6	7	
1	Articulate	:	:	:	:	:	:	:	Unclear
2	Inexperienced	:	:	:	:	:	:	:	Experienced
3	Intelligent	:	:	:	:	:	:	:	Unintelligent
4	Slow	:	:	:	:	:	:	:	Quick
5	Knowledgeable	:	:	:	:	:	:	:	Uneducated
6	Unkind	:	:	:	:	:	:	:	Kind
7	Fluent	:	:	:	:	:	:	:	Not fluent
8	Good-natured	:	:	:	:	:	:	:	Hostile
9	Considerate	:	:	:	:	:	:	:	Inconsiderate
10	Shy	:	:	:	:	:	:	:	Talkative
11	Has bad pronunciation	:	:	:	:	:	:	:	Has good pronunciation
12	Hesitant	:	:	:	:	:	:	:	Sure
13	Informative	:	:	:	:	:	:	:	Unhelpful

This is the end of the study.

Are you interested in a follow-up interview for approx. an hour?
_______ Yes. I am. I'll be receiving $17 remuneration for the interview.
_______No, thanks.

APPENDIX D

CORRELATION MATRIX FOR FEELING ATTRIBUTE

	Clear	Sure	EA	FL	CD	Calm	IT	TF	HP	Quick	KG	Kind	FD
Clear	1	.455**	.125	.537**	.361**	.385**	.489**	.270**	.235**	.014	.425**	.005	.092
Sure	.455**	1	.283**	.659**	.652**	.465**	.455**	.253**	.471**	.403**	.511**	.004	.124
Enthusiastic	.125	.283**	1	.090	.272**	-.069	.123	.205*	.287**	.308**	.267**	.378**	.394**
Fluent	.537**	.659**	.090	1	.531**	.465**	.538**	.246**	.310**	.365**	.523**	-.068	.068
Confident	.361**	.652**	.272**	.531**	1	.405**	.543**	.358**	.310**	.423**	.523**	-.025	.095
Calm	.385**	.465**	-.069	.465**	.405**	1	.532**	.272**	.255**	-.061	.402**	.135	.097
Intelligent	.489**	.455**	.123	.538**	.543**	.532**	1	.545**	.195*	.246**	.612**	.167*	.152
Thoughtful	.270**	.253**	.205*	.246**	.358**	.272**	.545**	1	.290**	.111	.418**	.366**	.370**
Happy	.235**	.471**	.287**	.310**	.310**	.255**	.195*	.290**	1	.366**	.313**	.317**	.395**
Quick	.014	.403**	.308**	.365**	.423**	-.061	.246**	.111	.366**	1	.282**	.016	.199*
Knowledgeable	.425**	.511**	.267**	.523**	.523**	.402**	.612**	.418**	.313**	.282**	1	.210*	.257**
Kind	.005	.004	.378**	-.068	-.025	.135	.167*	.366**	.317**	.016	.210*	1	.715**
Friendly	.092	.124	.394**	.068	.095	.097	.152	.370**	.395**	.199*	.257**	.715**	1
Informative	.108	.361**	.327**	.299**	.258**	.122	.207*	.372**	.433**	.384**	.394**	.309**	.493**
easy	.238**	-.067	-.093	-.101	-.138	.298**	-.008	.052	.010	-.424**	-.009	.101	.076
Quiet	-.001	-.259**	-.203*	-.123	-.233**	-.046	-.088	-.135	-.296**	-.298**	-.178*	-.027	-.122
Strong	.258**	.381**	.369**	.225**	.130	.164	.280**	.276**	.419**	.280**	.238**	.396**	.429**
Organized	.305**	.430**	.198*	.295**	.278**	.354**	.354**	.261**	.354**	.338**	.395**	.248**	.311**
Experienced	.276**	.521**	.356**	.396**	.443**	.350**	.340**	.267**	.403**	.438**	.369**	.308**	.336**
good-natured	.153	.121	.296**	.051	.157	.138	.317**	.428**	.274**	.201*	.316**	.665**	.599**
Pleasant	.294**	.133	.365**	.134	.140	.210*	.369**	.414**	.317**	.176*	.416**	.630**	.594**
Considerate	.155	.097	.323**	.101	.015	.160	.236**	.413**	.326**	.170*	.273**	.628**	.636**
Talkative	.117	.522**	.311**	.363**	.405**	.223**	.303**	.226**	.438**	.515**	.444**	.199*	.252**
Aggressive	-.111	.150	.043	.013	.151	-.227**	-.178*	-.191*	.133	.153	-.044	-.375**	-.241**
GoodPro	.694**	.344**	.013	.464**	.320**	.338**	.495**	.290**	.207*	-.006	.418**	-.080	-.035

EA=Enthusiastic, FL=Fluent, CD=Confident, IT=Intelligent, TF=Thoughtful, HP=Happy, KG=Knowledgeable, FD=Friendly

Appendix D *(continued)*

	IF	Easy	Quiet	Strong	OG	EP	GN	PS	CS	TT	AG	GP
Clear	.108	.238**	-.001	.258**	.305**	.276**	.153	.294**	.155	.117	-.111	.694**
Sure	.361**	-.067	-.259**	.381**	.430**	.521**	.121	.133	.097	.522**	.150	.344**
Enthusiastic	.327**	-.093	-.203*	.369**	.198*	.356**	.296**	.365**	.323**	.311**	.043	.013
Fluent	.299**	-.101	-.123	.225**	.295**	.396**	.051	.134	.101	.363**	.013	.464**
Confident	.258**	-.138	-.233**	.130	.278**	.443**	.157	.140	.015	.405**	.151	.320**
Calm	.122	.298**	-.046	.164	.354**	.350**	.138	.210*	.160	.223**	-.227**	.338**
Intelligent	.207*	-.008	-.088	.280**	.354**	.340**	.317**	.369**	.236**	.303**	-.178*	.495**
Thoughtful	.372**	.052	-.135	.276**	.261**	.267**	.428**	.414**	.413**	.226**	-.191*	.290**
Happy	.433**	.010	-.296**	.419**	.354**	.403**	.274**	.317**	.326**	.438**	.133	.207*
Quick	.384**	-.424**	-.298**	.280**	.338**	.438**	.201*	.176*	.170*	.515**	.153	-.006
Knowledgeable	.394**	-.009	-.178*	.238**	.395**	.369**	.316**	.416**	.273**	.444**	-.044	.418**
Kind	.309**	.101	-.027	.396**	.248**	.308**	.665**	.630**	.628**	.199*	-.375**	-.080
Friendly	.493**	.076	-.122	.429**	.311**	.336**	.599**	.594**	.636**	.252**	-.241**	-.035
Informative	1	-.015	-.240**	.428**	.542**	.499**	.369**	.423**	.510**	.454**	.011	.022
easy	-.015	1	.174*	.088	.041	-.074	-.020	.141	.075	-.138	-.206*	.247**
Quiet	-.240**	.174*	1	-.262**	-.199*	-.214*	-.077	-.147	-.133	-.366**	-.219**	.001
Strong	.428**	.088	-.262**	1	.524**	.519**	.343**	.375**	.408**	.433**	-.088	.146
Organized	.542**	.041	-.199*	.524**	1	.622**	.413**	.377**	.421**	.485**	-.128	.172*
Experienced	.499**	-.074	-.214*	.519**	.622**	1	.365**	.317**	.418**	.497**	-.171*	.148
good-natured	.369**	-.020	-.077	.343**	.413**	.365**	1	.706**	.685**	.247**	-.335**	.084
Pleasant	.423**	.141	-.147	.375**	.377**	.317**	.706**	1	.697**	.317**	-.371**	.187*
Considerate	.510**	.075	-.133	.408**	.421**	.418**	.685**	.697**	1	.306**	-.347**	.015
Talkative	.454**	-.138	-.366**	.433**	.485**	.497**	.247**	.317**	.306**	1	.147	.031
Aggressive	.011	-.206*	-.219**	-.088	-.128	-.171*	-.335**	-.371**	-.347**	.147	1	-.003
GoodPro	.022	.247**	.001	.146	.172*	.148	.084	.187*	.015	.031	-.003	1

IF=Informative, OG=Organized, EP=Experienced, GN=Good-natured, PS=Pleasant, CS=Considerate, TT=Talkative, AG=Aggressive, GP=Good Pronunciation

APPENDIX E

CONSENT FORM

Dear examiner:

I'm a PhD student at University of Illinois at Urbana-Champaign and currently working on my dissertation that looks into examiner perception of varieties of English in oral test settings where examiner encounters test-takers of multiple English varieties. I'm writing to invite you to participate in this **on-line** study which will take approximately **one hour** to complete. You'll be receiving **$20** reimbursement upon completion of the study.

The specific tasks that you're asked to perform are as follows:
Step 1: Listen to six Indian speech samples.
Step 2: Respond to several questions after listening to each sample.
Step 3: Assign a holistic score to each sample.
Step 4: Read the rating scale and descriptors.
Step 5: Listen to the Indian speech samples.
Step 6: Assign analytic scores to each sample.
Step 7: Complete a questionnaire.

Your participation in this study is entirely voluntary. There are no risks associated with your participating in this study over and above those associated with everyday life. Your decision to grant or to decline permission will have no effect on your employment in, status at, or future relations with MELAB or Purdue University. All references to you as an examiner will be through a pseudonym. In addition, if you provide feedback on the rating and evaluating process and if that feedback is essential to my research analyses, that feedback will also be handled through a pseudonym.

If you should have any questions about the study, you may contact Huei-Lien (Tammy) Hsu at 217-819-8429 or by e-mail at hhsu9@uiuc.edu. If you have questions in regards to your rights as a research study participant, you may contact the University of Illinois Institutional Review Board at 217-333-2670 or by email at irb@uiuc.edu. Thank you very much for your time!

Sincerely,

Huei-Lien (Tammy) Hsu

I have read and understand the above information and voluntarily agree to participate in the research project described above. I have been given a copy of this consent form.

____

Signature Date

If you have any questions about your rights as a research participant please contact Anne Robertson, Bureau of Educational Research, 217-333-3023, or ber-irb@ed.uiuc.edu or the Institutional Review Board at 217-333-2670 or irb@uiuc.edu

APPENDIX F

CONSENT FORM FOR VERBAL PROTOCOL STUDY

Dear ESL instructors:

I would like to invite you to participate in a research project that explores your perception of varieties of English in a testing situation. This research is being carried out by Huei-Lien (Tammy) under supervision of Professor Fred Davidson in the department of Educational Psychology at the University of Illinois at Urbana-Champaign.

The purpose of the interview is to identify features of varieties of English and which part of the features affects your scoring judgment. This interview will last for approximately an hour and will be audio-recorded. The purpose of the audio-recording is for transcription only. Please place a check mark to indicate if I am allowed to transcribe our talk:

_______ Yes. You may transcribe the audio-recording.
________ No. Please do not transcribe the audio-recording.

The interview will be conducted by Huei-Lien in her office at Foreign Language Building or over Skype. You will be paid $17 an hour for the interview.

The benefits to you as a participant would be to expose varieties of English during the study and learn the potential issues of English teaching and learning. The only possibility of risk involved would be slight emotional discomfort and fatigue. You may withdraw your paricipation in the study at any point.

All the data collected in this research will be kept confidential, and a pseudomym will be used in any analysis of the data in the final research paper and discussion. Your decision to grant or to decline permission will have no effect on your employment in, grades at, status at, or future relations with UIUC or the institution that you are affiliated.

You will be given a copy of this consent form. If you have any questions about the research or the results, please feel free to contact Huei-Lien (Tammy) Hsu at hhsu9@illinois.edu and Prof. Fred Davidson at fgd@illinois.edu.

Name: ____ Date : ____/____/____/

Signature:

****
If you have any questions about your rights as a research participant please contact Anne Robertson, Bureau of Educational Research, 217-333-3023, or arobrtsn@ad.illinois.edu or the Institutional Review Board at 217-333-2670 or irb@illinois.edu