Relating Learner Culture to Performance on English Speaking Tests with Interactive and Non-Interactive Formats

by

Nicholas Travers
B.A., University of British Columbia, 1998
M.A., University of British Columbia, 2002

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF ARTS

in the Department of Linguistics

Supervisory Committee

Relating Learner Culture to Performance on English Speaking Tests with Interactive and Non-Interactive Formats

by

Nicholas Travers
B.A., University of British Columbia, 1998
M.A., University of British Columbia, 2002

Supervisory Committee

Dr. Li-Shih Huang, Department of Linguistics
Supervisor

Dr. Hua Lin, Department of Linguistics
Departmental Member

Abstract

Supervisory Committee

Dr. Li-Shih Huang, Department of Linguistics
Supervisor

Dr. Hua Lin, Department of Linguistics
Departmental Member

This thesis explores relations between learner culture, operationalized as degree of individualism/collectivism (I/C), and English-as-an-additional-language (EAL) speaking test performance with two test formats that differ in terms of interactiveness.

Seven Korean participants' speaking test performances with the two different formats were compared. Results did not differentiate the speaking test formats in terms of mean speaking test scores or gains. However, results supported the value of the interactive format – Dynamic Assessment (DA) – for discriminating between test-takers in terms of grammatical and lexical performance. This characteristic suggests DA's potential effectiveness as a component of a formal speaking test, particularly for ongoing classroom testing and/or exit testing.

I/C scores did not correlate significantly with scores on the two speaking test formats. However, qualitative analysis based on I/C scores identified differences in the ways that participants oriented themselves towards accuracy or task topics in corrective exchanges during DA tests. Participants' email survey responses supported this analysis. These findings are commiserate with reports of accuracy focus in Korean educational culture. This link points to the value of future I/C research focusing on accuracy/task-focus orientations. To more reliably demonstrate relations between I/C and EAL performance, this study's discussion describes a more direct I/C measurement approach.

TABLE OF CONTENTS

LIST OF TABLES

ACRONYMS

The following is a list of acronyms that appear in this thesis:

EAL: English as an additional language

DA: Dynamic Assessment

NI: An abbreviation for the non-interactive format used in the present study

I/C: Individualism and collectivism

HI: An abbreviation for the high individualism group in the present study

HC: An abbreviation for the high collectivism group in the present study

IELTS™: International English Language Testing System

TOEFL®: Test of English as a Foreign Language

TOEFL iBT™: Test of English as a Foreign Language Internet-based Test

FCE: First Certificate in English

CAE: Certificate in Advanced English

TOEIC®: Test of English for International Communication

CF: Corrective feedback

NS: Native speaker

NNS: Non-native speaker

ZPD: Zone of Proximal Development

SCS: Self Construal Scale

RS: Regulatory Scale

ACKNOWLEDGEMENTS

It is not possible to thank everybody who assisted and supported me in the completion of this thesis. I would like to thank Dr. Li-Shih Huang for all of her support, expertise and enthusiasm in supervising this project. I feel very fortunate that I have been able to work with Dr. Huang, and learn from her not only how to improve as a researcher, but also the many skills that go into being an academic professional. I also wish to thank Dr. Hua Lin for her encouragement, careful reading and critical feedback during the preparation of this thesis. The same applies to Dr. Ulf Schuetze.

For their generosity and support in carrying out the data collection, I would like to thank Nancy Ami and all the teachers and students at Global Village English Centre Victoria. For her help in rating the speaking tests I would like to thank my colleague Emily Story. For answering questions about the International English Language Testing System (IELTS™) test, and for directing me to practice speaking tests, I would like to thank a fellow teacher, Wayne Everett. For his help in training me to administer the speaking tests, I would like to thank my fellow grad student and tennis partner, Akitsugu Nogita.

The sacrifices involved in my MA meant that I have leaned heavily on my family for support, including my parents Heather and Tim Travers. However, Mum and Dad were typically generous and supportive of my efforts. My daughter Erika, despite utterly rejecting my requests for extra writing time, has brought her endless supply of joy to my time as an MA student, which coincided with her first two years of life. Lastly, for all of her love and loyalty, which has kept me going through good times and bad, I thank Kana.

CHAPTER ONE: INTRODUCTION

1.1 Background

A central motivation of this study is to explore the intersections of learner culture and second language learning. Once individuals are socialized into cultural groups, culture constitutes their "mode of being," or often-unconscious patterns of relating to the environment they inhabit (Kitayama, Duffy & Uchida, 2007, p. 137). Yet culture is so deeply integrated into all facets of human experience that its workings are not easy to define or even observe, and so its importance in the context of second language teaching and learning is often underappreciated. Nonetheless, a number of researchers have stressed the inseparability of language and culture, and therefore the necessity of teaching culture alongside other linguistic skills. Lantolf (2006) described the cultural relativity of lexical organization and gesture. Magnan (2008) urged administrators to embed language teaching within other target-culture coursework, as a means for students to acquire an authentic cultural voice along with linguistic tools. Savignon and Sysoyev (2002) advocated role-play as a means of deepening awareness of the cultural values that come along with learning English. The value of such teaching is clear, as Magnan (2008) argued, because too often learners use a second language only as a medium for the conceptual system of their first language. Instead such research has pointed out that the communicative competence (e.g., Canale & Swain, 1980) teachers seek to equip learners with often lacks information about such things as metaphor, body language and appropriacy, all of which contribute to cultural fluency.

Other culture-oriented studies have described difficulties in applying Western language teaching ideas to non-Western contexts, and particularly East Asian classrooms (i.e., Chinese, Japanese, Korean, Taiwanese, and Vietnamese) (e.g., Han, 2005; Ross, 1998; Song, 1994; Spratt, Humphreys & Chan, 2002; Sullivan, 2008; Wen & Clement, 2003; Young & Halleck, 1998). These studies have relied on observation, local accounts, experience teaching in foreign classrooms, and social history to support arguments about cultural differences. Such discussions have raised awareness of cultural differences, and also serve as warnings to teachers with prescriptive views about the best way to teach second languages. However, they beg the question of how culture realizes itself in actual classroom practice. How do aspects of learner culture

affect classroom behaviours? Is it possible to move from general cultural information to more specific cultural factors that affect learners' second language development?

The present study explores these questions by drawing upon a construct from cultural psychology: individualism and collectivism (I/C). I/C shows promise for second language research for a number of reasons. Firstly, over 20 years of research have lent support to its ability to distinguish between cultural groups (e.g., Hofstede, 1980; Kitayama et al., 2009; Oyserman et al., 2002; Singelis, 1994; Trafimow et al., 1991; Triandis & Gelfand, 1996). Secondly, I/C research has largely targeted the same East Asian/Western cultural differences that a number of second language teaching/learning researchers have focused on (e.g., Han, 2005; Ross, 1998; Sullivan, 2008; Spratt, Humphreys & Chan, 2002; Wen & Clement, 2003; Young & Halleck, 1998). Indeed, this body of research has often been critical of a perceived individualist bias in English teaching (e.g., Han, 2005; Schmenk, 2005; Spratt, Humphreys & Chan, 2002; Sullivan, 2008; Wen & Clement, 2003), which adds impetus to a more detailed examination of I/C, and its implications for language teaching and learning. As such, the present study represents an investigation into the feasibility of adopting I/C as a framework for investigating second language learning and teaching issues. To this end, participants' I/C orientations are measured using a questionnaire, in order to correlate I/C score differences with differences in second language speaking test scores.

The focus on speaking tests reflects their importance for many language learners, and the fact that tests are a focal point in the cross-cultural interaction that takes place between native speakers and non-native speakers. Academic placement and exit tests, as well as speaking test components of major language tests, represent doors to educational and occupational advancement. It is therefore crucially important that speaking tests are fair to all test-takers. On the other hand, there is reason to believe that the International English Language Testing System (IELTS™) speaking test, which is a focus of the present study, may not be a culturally equitable measure of oral proficiency. The test contains little authentic interaction between examiner and test-takers, and the format places high value on offering personal information and opinions (UCLES, 2005; IELTS™, 2009). Superficially these characteristics do not appear particularly unusual. However, the I/C and speaking test literature suggest that the IELTS™ speaking test may disfavour collectivist test-takers (e.g., Gudykunst et al., 1996; Gudykunst, 1998; Kim, 1994;

Kim & Suh, 1998; Oyserman et al., 2002; Ross, 1998; Young & Halleck, 1998). For this reason the present study compares two speaking test formats. A Non-Interactive (NI) format simulates the IELTS™ speaking test's administration. A second format, Dynamic Assessment (DA), which includes more examiner-test-taker interaction, and may therefore be more culturally equitable, is also used. DA is an approach that has emerged from Sociocultural Theory, and specifically the ideas of educational psychologist L.S. Vygotsky (e.g., Rieber & Carton, 1993). Of particular importance for DA are Vygotsky's notions that learning is both mediated and dynamic (e.g., Lantolf & Poehner, 2008). In the present study, DA involves the examiner intervening to assist participants with error correction at moments where they are unable to perform independently. At the same time this intervention provides assessment information regarding learners' language proficiency (e.g., Lantolf & Thorne, 2006). Additionally, as part of DA's mandate to assist learners in attaining higher proficiency levels, DA in the present study also involves affective support (e.g., Feuerstein, 1981). This study relates scores on these two speaking test formats with I/C scores using correlational analysis. To provide an additional perspective on the relation between I/C and speaking tests, this study also re-examines interaction in the DA format tests, to assess whether I/C orientation relates to communication style patterns.

Using DA for speaking tests represents a novel application of this assessment approach. Moreover, this study's standardized version of DA has not previously been used in second language learning and teaching research (see, however, Aljaafreh & Lantolf, 1994; Lantolf & Poehner, 2004; Poehner, 2007; Poehner, 2008). As such, the present study further represents an investigation into the efficacy of DA as part of a standardized speaking test format.

1.2 Purpose of the Study

One purpose of the present study is to examine the relation between learner culture and speaking test performance. Since culture is an extremely broad concept, it is operationalized here as degree of I/C as measured by a questionnaire. In particular, the study considers the relation between test format, which here relates to the amount of interaction between test-takers and examiner, and speaking test performance in terms of the learner's culture. To explore these variables, the study includes a non-interactive testing approach (NI), which simulates the administration of the IELTS™ speaking test. As a comparison, the study also employs an interactive testing approach, DA, and a complementary rating system, which have not previously been used with speaking test interviews. Therefore, a second, related purpose of this study is to consider the efficacy of DA for use in oral proficiency examinations. In this study DA takes the form of an interactive test that remains controlled, in terms of the type and amount of interaction that the examiners are permitted to engage in. This control facilitates test standardization, which helps to ensure consistent administration and rating. At the same time, DA's interactive elements allow for more naturalistic interaction than non-interactive tests permit. In terms of I/C, this study seeks to evaluate the construct's usefulness as a framework for illuminating relations between culture and EAL teaching and learning performance.

1.3 Outline

The thesis is organized as follows: Chapter 2 reviews the literature on testing second language speaking ability, as well as the literature on Dynamic Assessment, corrective feedback in second language settings, and individualism and collectivism. This literature review offers a discussion of key issues in the above fields of study, and establishes a research context for this study. Chapter 3 describes the study's methodology, including the design, participants and procedure. Chapter 4 presents the study's results and discussion, as well as its limitations and directions for future research. Finally, Chapter 5 summarizes the study's findings and their major implications.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter reviews the literature that is most relevant to this study. The main sections are elaborated by sub-sections that target key issues in the respective fields. Section 2.2 introduces a number of studies and surveys concerning speaking tests. The section raises a number of points, which often relate to the tension between developing reliable tests and tests that reflect real-life speaking contexts. Overall, the literature revealed how the interrelatedness of design, administration, scoring, and other factors makes evaluating speaking tests particularly difficult.

Section 2.3 discusses Corrective Feedback (CF). This is a key issue in the present study because the interaction in one of the speaking test formats – Dynamic Assessment (DA) – primarily involves negotiating correct forms after the examiner hears an error. Additionally, the qualitative portion of my analysis focuses on these corrective exchanges. The CF literature offers few clear conclusions, in terms of the most effective ways of correcting learner errors. Instead, CF research shows that a multitude of factors can affect learners' noticing and using correction, including the degree to which the class (and the teacher) focuses on linguistic forms or communicative taskwork, as well as learner factors such as their CF preferences and expectations.

Section 2.4 introduces the literature relating to DA. This approach to assessment and instruction is rooted in Sociocultural Theory, and stresses the importance of mediators assisting less expert learners through interaction. A key question for DA is the degree to which mediator assistance should be standardized, or should adapt itself to the learner and situation. The consistent success of the approach in distinguishing learners with similar independent test scores serves as validation for the approach, even though its typically non-stardardized administrations make the approach difficult to evaluate.

Lastly, section 2.5 discusses the literature relating to individualism/collectivism (I/C), which is a major line of research in Cultural Psychology. A large number of studies (e.g., Gudykunst et al., 1996; Kitayama et al., 2009; Oyserman et al., 2002; Oyserman & Lee, 2008; Singelis, 1994; Trafimow et al., 1991; Utz, 2004) lend support to the contruct's power to differentiate between cultural groups. The research has supported correlations between I/C and a number of variables, including values (e.g., Hofstede, 1980), notions of self (e.g., Singelis, 1994),

perceptions of others' behaviour (e.g., Oyserman et al., 2002), ideas regarding relations with others (e.g., Kim, 1994), and communication styles (e.g., Gudykunst et al, 1996). In addition, a central debate has focused on the best means of measuring individual I/C, and through individual-level responses, of tapping into underlying cultural norms. The picture that emerges is that individual variability makes cultural-level generalizations problematic, even while overall trends strongly support basic cultural differences in terms of I/C. At the same time, this variability has enabled researchers to model the ways individuals mediate the cultures that they were socialized into.

Before discussing research in related fields, however, it is necessary to define some key terminology. Therefore each part of the literature review begins with a number of definitions.

2.2 *Speaking Tests*

2.2.1 *Speaking Test Terminology*

- *Speaking test interview*: A formal test in which an examiner elicits spoken language from one or more test-takers. Test-takers may be asked to answer questions, engage in more open-ended discussions, speak at length, or complete a speaking task. Test-takers are evaluated based on their performance in the test.
- *Test-taker:* Often a language learner, this is the individual who takes the speaking test.
- *Examiner*: This is the individual who administers the speaking test.
- *Interlocutor*: The examiner and interlocutor may be the same person, but in speaking test contexts, the term interlocutor denotes an individual who interacts with test-takers during a speaking test.
- *Administration*: Carrying out the speaking test tasks with the test-taker. Often this involves the examiner reading a set script of questions and prompts, using (and refraining from using) certain language, and following a strict time limit for the test sections.
- *Rater*: An individual who scores a speaking test. This person may also be an examiner and/or interlocutor, or may exclusively listen to and score the test.
- *Rating Scale*: Typically a ranked scale, often subdivided into linguistic categories (e.g., grammar, pronunciation, etc.), that raters use to assign scores to test-takers.
- *Criterion Situation*: The real-life correlate to the criteria by which test-takers are rated. Often this denotes the performance of a native speaker, but could also reflect situation-specific functional requirements, such as talking on the telephone.
- *End-users*: The interested parties who use the test scores, including employers and academic institutions.
- *Standardized Test*: A speaking test with the same length, format, level of difficulty, task types, and administration from test-taker to test-taker.

2.2.2 *Introduction to Speaking Tests: Standardization versus Authenticity*

Second language oral proficiency tests abound. There are job-specific tests and locally developed academic placement and achievement tests. There are also commercially available exams devoted to testing speaking abilities, such as the Test of Spoken English (TSE®). Other

commercial speaking tests form part of influential language proficiency exams such as the Test of English as a Foreign Language (TOEFL®), and the International English Language Testing System (IELTS™). The above measures differ more or less widely in terms of format, length, rating methods and rating scales; whether test-takers are alone, or in pairs or groups; in terms of interaction between test-takers, and/or interaction between test-takers and the examiner; with regard to task types, and other factors. This variety reflects differing conceptions of communicative competence, different purposes for testing, and economic and logistical constraints. It also reflects a tension between efforts to ensure test reliability, and efforts to ensure that tests can legitimately claim to measure the speaking abilities that test-makers, test-takers and other end-users are interested in.

Reliability demands controls and standardization, while a desire to simulate real-world interaction leads, for example, to incorporating interlocutors and/or a degree of spontaneity into test design and administration. This is not to say that tests seeking authenticity necessarily involve unscripted casual conversation. This is only one of many types of spoken interaction. Instead authenticity means tests that simulate as far as possible the "criterion situation" against which test-takers will be measured (McNamara, 1996, p. 16). These choices also affect (and are affected by) test content, and the criteria used to define oral proficiency and score the tests. As a result, demands for authenticity and reliability tend to pull tests in opposite directions, creating a dilemma for test-makers that is not easy to resolve.

Standardized tests allow for greater reliabilities in terms of administration and scoring. The ease with which scores may be quantified and compared is not the only advantage of standardization, though it is certainly primary (Ellis, 2003). An additional benefit, from a test-taker's point of view, is that standardization reduces sources of error from variability in content, administration and scoring, and therefore increases test fairness (Brooks, 2009; Fulcher and Davidson, 2007). Learners with equal speaking abilities may earn very different results on non-standardized tests, with serious ramifications for their educational and occupational careers. It is necessary to ensure that all test-takers have an equal opportunity to produce their best possible performance during the test. Seen from this angle, controlling test variables benefits all test end-users, including the test-takers themselves. On the other hand, in an effort to eliminate sources of measurement error, tests may become so controlled that they actually fail in their primary

purpose of assessing oral proficiency (Hughes, 2002). To put it another way, the kind of talk that happens during a highly standardized test may not bear much resemblance to the kind of talk in the real-life situations for which the test-taker is being assessed.

2.2.3 *Speaking Test Interviews as Rule-Governed Interaction*

Many speaking test interviews rest on an assumption that the speech samples they elicit are representative of test-takers' general speaking abilities, and can predict their success at speaking in a variety of real-world settings. This is somewhat disingenuous if we take the view that talk is affected by the contextual factors that surround it. For speaking tests these factors include the interlocutors, setting, and conversational conventions that mark interview tests as both interviews and tests. It is important to remember that interviews, and even more specifically speaking test interviews are not naturalistic conversations. This may seem obvious, but it has important ramifications for test design and assessment. Situation-specific roles and statuses, as well as generic conventions that define interviews mean that certain types of language, conversation management or interactive strategies are more or less likely to occur. Thus one area of test interview research has looked at interview interaction as social events with distinctive turn-taking and discursive conventions (e.g., He & Young, 1998; Riggenbach, 1998). A number of researchers, through use of Conversation Analysis or Discourse Analysis, have identified unique features of language interviews that distinguish such talk from casual conversation (e.g., He & Young, 1998; Lazaraton, 1992; Riggenbach, 1998; Ross, 1998). Comparing this type of interaction to a conversation with a friend, Weir (2005) pointed out that reciprocity, or sharing responsibility for the conversation, does not really apply to interviews. In interview talk, initiation, topic management and concluding responsibilities are firmly in the hands of the interviewer. Most saliently these claims are centred on the examiner's right to control the conversation through an (nearly) exclusive right to ask questions (e.g., He & Young, 1998). In addition, interview tests are characterized by examiners prompting test-takers for greater accuracy or elaboration, even when they have communicated ideas successfully, moves which are unlikely to occur in casual conversations (e.g., He & Young, 1998).

Ross (1998) argued that an underemphasized fact about speaking test interviews is that they implicitly test pragmatic competence. He offered an example of an examiner who asks a test-taker to explain how he can use a public telephone. The test-taker may be taken aback by a

request to give information that the examiner surely already knows, unless the test-taker is aware of (and willing to follow) the following key English speaking test interview schema: in most speaking test interviews, truthfulness or even topical relevance are less important than the linguistic skills evidenced in the test-taker's responses. This claim is supported by instances where examiners do not reformulate questions after non-relevant answers, but move on to subsequent topics, something which is unlikely to happen in conversations elsewhere (Ross, 1998). Conversely, test-takers who understandably focus on the literal content of questions, and are embarrassed by them, may display self-defensive or non-cooperative attitudes towards examiners. Such behaviour may surprise examiners who are focused less on test content than simply eliciting a sufficient sample of the test-taker's talk. An example is test-takers who choose not to divulge casual personal information, or offer cursory answers because of their lack of intimacy with the examiner. Ross (1998) calls this the "minimalist response strategy" that he observed in Japanese learners, by which test-takers provided no more than yes or no answers to questions (p. 340). This strategy can be understood in relation to Japanese culture, in which superfluous responses tend not to be highly valued. From a Japanese cultural perspective, the unequal power relations in an interview may further restrain test-takers from speaking, as the examiner is seen as possessing "speaking rights," while a Japanese interview test-taker typically provides "exact responses to questions – no more, no less" (Ross, 1998, p. 341). In an English speaking test, such behaviour may satisfy basic communicative requirements by precisely answering examiner questions. Yet the behaviour fails pragmatically, by revealing a lack of awareness of – or resistance to – the English interview schema, which prioritizes a sufficient sample of talk to assess linguistic abilities.

To elicit a wider range of interactional features, test designers may adjust speaker roles, as in role-play tasks, or add a test-taker to allow for paired interaction. Such designs gain validity by integrating more naturalistic real-world interaction. However, depending on the purpose of the test, more rigid interviews may be equally valid. To assess proficiency for work or academic settings, interview tests gain validity by replicating the power asymmetry test-takers may later encounter in the targeted criterion settings (Weir, 2005). This is an important point, since many communicatively-oriented language teachers and researchers tend to privilege unstructured conversation over other spoken genres. More accurately, casual conversation is only one type of

talk, and may not be appropriate as an indication of communicative skills in, for example, a test of English for academic purposes (Riggenbach, 1998). With regard to language interview tests, raters and rating schema need to show an awareness of the constraints that this type of talk puts on conversation. In short, it is important to acknowledge the rules defining speaking test interviews as a genre, including pragmatic expectations and topic management. With this understanding we can limit our expectations to the sorts of test-taker talk such tests can reasonably elicit.

2.2.4 *Variability in Interactive Speaking Tests*

The reliability/authenticity dilemma affects every aspect of speaking tests, from design to scoring. However, a particular flashpoint is interaction. Many indices of oral proficiency, such as the ability to claim turns, maintain and yield turns, backchannel, and ensure comprehension, and engage in effective repair exchanges, are necessarily interactive (Riggenbach, 1998). There is an increasing awareness that such skills are an essential component of oral proficiency; however, there is debate regarding where such abilities are located. One model sees successful interaction as reflective of individual competency (e.g., Hymes, 1972). Another model emphasizes the essentially social nature of interaction, and argues that interpretations of successful communication therefore must take into account both speakers' contributions, as well as other contextual factors (e.g., Kramsch, 1986; Jacoby & Ochs, 1995). For the purposes of most speaking tests, interactive test formats certainly seem to increase validity. For one thing, interaction corresponds closely to the sort of authentic speaking tasks prevalent in communicative language classrooms. These, in turn, attempt to simulate the kinds of meaning- and goal-focused communication learners are likely to engage in outside the classroom (Ellis, 2003). Yet interaction adds an additional variable to the challenge of designing and administering reliable speaking tests.

Test formats differ in their inclusion of interactive elements. Even within interview formats, which is the present study's focus, there are degrees of interactiveness. Some oral proficiency measures, such as the IELTS™ test, include no real examiner-test-taker interaction, in terms of the negotiation and repair skills promoted by Riggenbach (1998). The examiner has a set frame and may not stray from the script, with the exception of repeating questions and possibly prompting test-takers for more information. Other tests, such as Cambridge ESOL First

Certificate in English (FCE), include no more examiner-test-taker interaction than IELTS™, but do contain limited paired (between-test-taker) interaction. At the other end of the spectrum, in-school placement, progress or exit tests may be loosely structured interviews that attempt to elicit more naturalistic, spontaneous conversation.

In one study focused on speaking test interaction, Brooks (2009) found that paired test-takers received higher scores than individual test-takers in an interactive speaking test. Brooks' pairs also produced a greater variety of conversational features than individuals. This suggests that paired tests may lend themselves to greater conversational complexity; however, the targeted features were interactional, and it was not clear whether, in the individual tests, the examiner/interlocutor played a highly interactive role in the discussion or not. This would notably affect the number of interactive features that emerged in the tests. Brooks did not develop an argument showing that the increased number of interactional features led to the paired test-takers receiving higher scores. However, Brooks pointed out that the rating scale used by her raters did not take into account many of the interactive features that occurred in the test conversations. Whether this related to the difference in scores between pairs or individuals, or whether raters were impressed by the richer interaction evidenced in the paired tests, is not clear. As Brooks (2009) herself suggested, raters may have found it difficult to give individual scores for interactive achievements, and may have compensated for this uncertainty by awarding higher scores to both test-takers. However, because raters were not interviewed to elicit their decision-making rationales, these possibilities remain speculative.

Wigglesworth (2001) similarly found that non-native speaker (NNS) pairs earned higher scores than NNS-native-speaker (NS) dyads, and in a possibly related finding, that pairs who were familiar with each other produced more discursive features than stranger pairs. These features included sentence completions, interruptions, overlaps, as well as increased comprehension checks, and clarification requests. On the other hand, the data from stranger pairs revealed more "echoic" repetitions that served to avoid misunderstandings (2001, p. 187). The results from this and Brooks' studies are intriguing, and suggest that interactive tests may be able to measure greater discursive skills than non-interactive tests. However, the studies also revealed the difficulties involved in rating interaction, since reliable rating scales must be developed alongside interactive test content. The apparently greater success of paired test-takers is also

difficult to account for. It is unclear whether pairs earned higher scores because the paired format allowed the test-takers to display communicative abilities that individual testing did not elicit; whether the paired format reduced the test-takers' anxiety; or whether an inadequate rating scale, or rater biases somehow favoured the paired test.

2.2.5 *Variability in Rating Interactive Speaking Tests*

It is apparent that if interaction is considered vital for assessing spoken proficiency, rater training and scoring criteria must also reflect this emphasis. This is not always the case, however, in that an increasing awareness of the differences between writing and speaking has not extensively affected speaking test rating criteria, which traditionally focused on the same features (i.e., grammar, vocabulary, coherence) that are assessed in writing (Riggenbach, 1998). Hughes (2002) pointed out that this has created a strange disjunction between the talk that happens in speaking tests, and the ratings used to judge it. While ratings tend to focus on discrete elements such as grammatical or lexical items, the speakers themselves focus on information, meanings, and each other. Even when rating scales include more communicatively-focused criteria, raters may simply ignore them, being unable (or unwilling) to forego grammatical-lexical criteria in their judgments (McNamara, 1996). A further problem with rating criteria is that they have often posited an ideal native speaker as a benchmark against which learners can be judged. Yet this choice contradicts the evidence of actual native speaker conversations, which are often full of hesitations, fragments and ungrammatical language (e.g., Hughes, 2002; McNamara, 2006; Riggenbach, 1998).

It is also clear that adding interactive elements to speaking tests complicates (and thus affects the reliability of) test scoring. There are both theoretical and practical difficulties. For one thing, a theoretically sound understanding of what constitutes "good" spoken language is necessary, and is a step that will have important positive effects on test content, rating criteria and rater training. Both Galaczi (2008) and May (2009) found, for example, that raters were unsure how to score individuals in some paired tests. Though this is problematic for test fairness, these studies also provided insights that may improve/standardize interactive test rating. Thus even as these studies pointed to difficulties in evaluating spoken interaction, they contribute to better understanding it, a need that Hughes (2002) has stressed. As with Brooks' (2009) study, May (2009) argued that current rating scales were inadequate for capturing skills relating to

interactive features. May identified conversational features that affected rater perceptions, and which could be incorporated into scoring criteria, including body language, assertiveness, managing conversation and cooperation. Galaczi's (2008) study identified four principle patterns of test interaction and their relative valuation. Patterns evidencing mutuality (responding to and developing a partner's topics) and equality (neither dominance nor passivity) were positively evaluated by raters. These findings point to test-taking strategies that teachers can pass on to students during test preparation. The high valuation of cooperation found in both these studies suggests that giving one score for two test-takers is a reasonable choice for improving the rating of interactive tests. Added support for this move is Galaczi's (2008) finding that raters had difficulty awarding individual scores for asymmetric conversations, in which one partner dominated the other. May's (2009) results concurred with Galaczi's, both in terms of positive evaluations of mutually supportive interaction, and uncertainty regarding how to rate asymmetric interactions.

2.2.6 *Examiner Variability in Interactive Speaking Tests*

Rater inconsistency is not the only obstacle to reliable interactive testing. Wigglesworth (2001) found that examiners who framed tasks more extensively for test-takers affected scores positively for them, as opposed to examiners who provided less explanations about how to carry out the task. The same study also looked at the effects of test-takers having a NS or NNS examiner. Results showed that test-takers fared better with NNS examiners. This was perhaps because test-takers felt more at ease with someone who was a fellow (albeit high-level) learner, as opposed to a NS. The study also evaluated interactive versus non-interactive administrative formats, finding that the negotiated talk that occurred in interactive tests supported the test-takers as they tried to accomplish speaking tasks. Both McNamara (1997) and Brown (2003) found that different interviewers (examiners) administering the same interactive speaking test varied significantly in terms of the results their test-takers earned. Yet this variability also had an effect on raters, so that test-takers tested by severe interviewers received higher scores than those tested by interviewers judged to be easier.

These findings point to the importance of contextualizing evaluations of speaking test performances. In other words, it is important to recognize that test-takers' scores may be affected by a number of variables beyond their speaking proficiency levels. Brown (2003) showed that

examiners who engaged in "supportive, scaffolding behaviour" in order to assist test-takers inadvertently hurt them with regard to scores (p. 9). Ironically, then, supportive behaviour, which has been perceived favourably by judges in *paired* interactive tests (Galaczi, 2008; May, 2009), was apparently judged to be an indication of a learner's inability to carry out the test tasks *without* examiner support. McNamara (1997) similarly looked at the effects of examiner variability, finding that different examiners caused tests to become more or less difficult for test-takers. Yet McNamara did not advocate strict control over interviewer turns, pointing out that even highly controlled speaking tests have evidenced notable differences between examiners. Instead, McNamara (1997), as well as Brown (2003), stressed that examiners can have both positive and negative effects on the interaction that takes place. With this in mind, Brown (2003) emphasized the need for both improved interviewer training and clearer criteria by which interactional speaking abilities can be judged.

With regard to interaction in speaking tests, the issue comes back to a choice between controlling a test's scope for interaction, or accommodating interactive variability in a comprehensive model of effective spoken communication. As an example of the former, the IELTS™ speaking test rating system (IELTS™, 2009) does not include interactional elements. Hughes (2002) for one has been critical of this approach. As she pointed out, excessive control is deeply ironic, since we understand that, in fact, "good oral communication is founded on one speaker actually having an effect on another, and on the reactions and responses which take place between interlocutors" (p. 79). A reasonable alternative to excessive control is to limit the scope of interaction by defining task contexts clearly, and to place emphasis on achieving global success, in terms of task and functional goals, rather than defining success in discrete grammatical and lexical terms. Yet even here questions arise, such as how to allocate scores in paired tests, or when the interlocutor is an examiner. Skills such as topic management and turn management, if evaluated by the overall communicative success of an exchange, cannot easily be ascribed to only one of the speakers (e.g., He and Young, 1998).

O'Sullivan et al. (2002) proposed a checklist of interactive features as a means of integrating interactive skills into a reliable scoring system. This checklist has the advantage of not being overly time-consuming to use. On the other hand, counting features seems to be an unrealistic way of determining proficiency (e.g., Ellis, 2003; Fulcher & Davidson, 2007). In

addition, raters using the checklist were inconsistent in finding interactional features (O'Sullivan et al., 2002), suggesting that the instrument is difficult to use, or that more rater training is required. Still, the checklist remains a plausible non-time-consuming instrument for rating interaction in speaking tests, and its shortcomings apply to other scoring systems as well.

Conversation Analysis (CA) (e.g., Sacks, Schegloff & Jefferson, 1974; Schegloff, Jefferson & Sacks, 1977) has the advantage of looking at conversation in great detail, which can allow us to assess the effectiveness of speakers' contributions during speaking tests. Riggenbach (1998), through CA of casual NS-NNS conversations, isolated a number of interactional skills that contributed to successful communication. Because CA is very time-consuming, however, it is unrealistic to attempt to use it as a scoring tool. One important contribution that CA can make to improving speaking tests is that CA's findings can assist in developing scoring instruments like the one produced by O'Sullivan et al (2002).

The impression that emerges from the literature on interactive speaking tests is that such tests' emphasis on simulating real-world interaction often comes with reliability shortcomings. Interactive speaking tests are theoretically driven, grounded in conceptions of talk as a social event in which all speakers contribute to successful communication. Test content and format appear to satisfy a demand for a test that elicits interactional skills in addition to more purely linguistic ones. However, relevant studies suggest that other essential testing procedures have not caught up with theory, or have been downplayed. McNamara (1997) makes this point when he reminds us that test validity does not end with task content and a format that closely align the measure with interactional models of language use. Subsequent testing must show that the test does indeed capture differences in interactional competence. In order for tests to be fair and reliable, interviewers need to be consistent, though there is no reason why this cannot include providing positive (i.e., scaffolded) support to test takers. Similarly, assessment criteria need to incorporate clear definitions of interactional proficiency, and raters must be able to understand these criteria and apply them consistently.

Weir (2005) stressed a "multifaceted" approach to test development (p. 13), an orientation that other researchers share (e.g., Fulcher & Davidson, 2007; McNamara, 1996; O'Sullivan et al., 2002). Such an approach needs to integrate mutually reinforcing (and ongoing) validity checks, which provide cumulative support for a test's soundness (Weir, 2005). While

satisfying every type of validity in this interlocking system may not always be possible, such a rigorous mode of test development represents a "mandate" for designers to follow (Weir, 2005, p. 20).

2.2.7 *Test-taker Variables in Speaking Tests*

The research on speaking test variability discussed above has focused on shortcomings of test design, administration and scoring. Test-taker variables have not featured prominently in studies, despite calls for a more emic perspective in second language interaction studies (e.g., Firth & Wagner, 1997). An implication seems to be that it is the test-takers' responsibility to adapt themselves to suit tests, rather than tests adapting to accommodate the people who take them. This is partly understandable, since it is difficult to design language tests that elicit only targeted skills, while controlling all other test-taker variables. Yet it is important to research the effects of non-targeted test-taker performance variables, since their presence raises doubts about a test's validity. Weir (2005) suggested that physical, psychological and experiential differences may all affect test performance. McNamara (1996), too, reminded us that learner variables affect test performance, and reported a study that found that level of (post-secondary) education significantly affected test performance. This suggested, problematically, that general knowledge, and not specifically language ability was being tested. Brooks' (2009) findings that pairs scored higher than individuals in speaking tests raises the possibility that anxiety in one-on-one interviews with NS examiners prevented test-takers from performing at their best. Hughes (2002) raised this point as well, reminding us that conversing with someone we know and are comfortable with is easier than talking to a stranger. Ultimately, it is unreasonable to expect test designs to account for all possible learner variables that may affect speaking test performance. Yet Weir (2005) stresses that, at the very least, administrators have a responsibility to make test task expectations clear, as well as make the scoring criteria and the scoring system transparent for test-takers. In addition to this, it does seem possible at least to evaluate test equitability with certain test-taker variables, including content knowledge and learner anxiety relating to the test format and administration. A novel aspect of the present study will be evaluating this equitability in terms of another test-taker variable: degree of I/C.

2.2.8 *Cultural Issues in Speaking Test Interviews*

Hughes (2002) listed cultural differences amongst variables that may affect test performance. In terms of the tests themselves, studies have shown that language interview tests vary cross-culturally, in terms of speaker roles and expectations (e.g., He & Young, 1998; Kim & Suh, 1998; Young & Halleck, 1998). Specifically, He and Young (1998) pointed to amount of talk, turn lengths, speaking rate, and talkativeness or taciturnity as communicative features that differ across cultures, and are also likely to affect test performance. For Japanese learners, this may include responding literally, and briefly, to display questions that were meant to elicit extended talk, a behaviour which is likely transferred from Japanese (Ross, 1998). Similarly, verbosity is generally not highly valued in Japanese, and if this estimation is transferred to English speaking tests, learner silence may be incorrectly interpreted as a lack of communicative ability (Young & Halleck, 1998). Along these lines, researchers have focused on the cultural relativity of orientations to speaking test interviews (e.g., He & Young, 1998; Ross, 1998). Ross (1998) showed that test-takers may have quite different ideas about appropriacy in such contexts. English NS examiners have tended to focus narrowly on eliciting a sufficient quantity of talk from test-takers to evaluate linguistic skills. The test-takers, on the other hand, may reasonably be more focused on other factors, such as not wishing to reveal personal information to a relative stranger, not wishing to offer opinions that they perceive as counter to the examiner's beliefs or expectations, and not wishing to challenge the examiner's status by speaking too much. Whereas learners in English speaking tests tend to be positively evaluated when they initiate topics, offer opinions and talk more than the examiner, these features may not be universally shared, and have to be learned by many second language speakers (e.g., Young & Halleck, 1998). Kim and Suh (1998), for example, showed how successful Korean-as-an-additional-language test-takers were careful to ratify unequal power relations by deferring to the examiner for evaluative remarks and concluding and topic-initiating turns. Contrary to instances of English speaking test interviewers using newsmarks or similar utterances (e.g., "Is that so?") to encourage further talk (Brown, 2003), Kim and Suh (1998) showed that in Korean language interviews such moves are pivotal points at which examiners reclaim control of conversation management. Moreover, only less proficient learners misinterpreted the interviewer's turn by adding more talk about the previous topic. Yet elaborating in such a manner would likely be evaluated positively by English speaking

test raters. While it is fair and necessary to evaluate learners based on such culturally relative sociolinguistic competencies, it also then becomes crucial to teach the generic conventions of interviewing in the target language. Otherwise, test-takers who are quite capable of developing topics may receive low scores, when a perceived lack of competency is more accurately a cultural misunderstanding.

While the studies cited above highlighted cross-cultural variability in speaking test interviews, a recent study suggested that there is a great deal of universality in basic conversational management (Stivers et al., 2009). Focusing on questions and answers in ten diverse languages, the researchers overall found many similarities. Overwhelmingly the data indicated that speakers across languages avoided overlap, and minimized silences between turns. In addition, response times were uniformly faster when answers were given than cases when answers were not given, suggesting that delays in responses occur when they run counter to the asker's agenda (i.e., not following a question turn with a preferred answer turn). Finally, across tested languages requests resulted in the slowest response times. The results are interesting, because they question ideas that some languages endorse silences, or endorse more "aggressive" turn-taking talk than others. With relevance to East Asian communication styles, the study's results contradicted anecdotal impressions that Japanese speakers respond more slowly than others. According to Stivers et al.'s (2009) data, Japanese speakers responded, on average, earlier than all nine other languages' participants. On the other hand, the study's fairly narrow focus on questions and answers means that results may not apply to other transition points in conversation. In addition, research that took into account wider contextual factors, such as the type of talk, and the relationships between the interlocutors, might have generated different results.

2.3 Corrective Feedback

2.3.1. Corrective Feedback Terminology

- *Corrective Feedback*: Providing information about the incorrectness of an item of language. It can be spoken or written. The feedback can be negative (indicating that a form was not correct), positive (offering a correct alternative), or both.
- *Feedback Types* (from Lyster & Ranta, 1997): (a) Explicit Correction: the correct form is given, and the corrective intention is made clear; (b) Recasts: Reformulating a learner's utterance in a correct form; (c) Clarification request: A question indicating that comprehension or accuracy was not achieved (e.g., 'Pardon me?'); (d) Metalinguistic Feedback: A description of the error's form is given, usually with reference to rules or grammatical terms. (e) Elicitation: The learner is prompted to correct an item with elision (e.g., 'No, it's a ...'), a question (e.g., 'How do we say that?'), or a direction (e.g., 'Please say that again.'). (f) Repetition: The error is repeated, often with rising intonation to focus attention on the target for correction. (e.g., 'You *go* to a movie last night?').
- *Uptake:* The learner responds to the correction, showing that she/he has noticed it. This may or may not include repair (i.e., the learner offers a correct form),
- *Communicative Orientation*: This refers here to the degree to which a classroom focuses on language forms (discrete grammatical, phonological, syntactic items) or on meaning-focused taskwork (i.e., *using* language to achieve communicative ends).

2.3.2 Types of Corrective Feedback and Learner Responses

CF studies have been concerned with documenting types of feedback that occur in NS-NNS interactions, their potential effectiveness, and the effects of variables such as communicative orientation and nationality on types of feedback and learner responses to them. This area of research is important, for one thing, because in task-based language classrooms corrective feedback represents a means for teachers to draw learners' attention to linguistic form, while maintaining communicative taskwork as the primary activity type. Moreover, as Lyster and Mori (2006) have pointed out, addressing errors during meaning-focused interaction (as

opposed to offering a prescriptive grammar lesson) has the advantage of targeting language that learners themselves attempted to produce.

Lyster and Mori (2006) divided corrective feedback into three categories: explicit (including both negative and positive feedback), prompts (which cue the learner to self-correct), and recasts (reformulations of learner utterances, but without the error that the learner produced). Teacher-led corrective sequences begin with an indication that an error occurred, and end with some form of learner response (i.e., uptake), or with one or both speakers abandoning the corrective framework to resume the interrupted task. Lyster and Mori divided learner responses to corrective feedback into either uptake that includes repair (i.e., a correct form), or uptake with no correct form. Evidence exists, from post-tests and delayed post-tests, that providing corrective feedback (rather than ignoring errors that occur) leads to improvements in the targeted structures. It is not clear, however, whether learner responses to correction are enough to claim that language development has taken place. Even subsequent utterances containing correct forms are no guarantee that a learner has internalized the new form (Sheen, 2004).

2.3.3 *Prevalence of Feedback Types and Learner Responses to Them*

A highly consistent finding is that recasts are the most frequently occurring type of CF (e.g., Long, Inagaki & Ortega, 1998; Lyster & Ranta, 1997; Lyster & Mori, 2006; Sheen, 2004; Yoshida, 2008). Their prevalence reflects their position between a conversational turn and a correction. In other words, they offer a correct form without disrupting the flow of conversation (Lyster & Mori, 2006). A difficulty with recasts, however, is that learners engaged in meaning-focused activity may not notice a teacher/interlocutor's shift to a corrective mode. Likely for this reason recasts are not necessarily the most successful form of corrective feedback, if measured by the frequency that they are noticed and/or elicit repairs (e.g., Lyster & Ranta, 1997; Lyster, 1998; Lyster & Mori, 2006; Nabei & Swain, 2002). The situation is further complicated by the fact that different researchers have not always been consistent in defining CF types (Nassaji, 2007). Nabei and Swain (2002), in a heuristic case study of one Japanese learner's awareness of recasts, concluded that a wide range of factors complicate our understanding of recasts' effectiveness. Nabei and Swain's holistic approach revealed a richly contextualized picture of recast effectiveness. They found that such variables as the teacher's orientation towards correcting errors, the type of language point that the teacher targeted, the student's learning style,

and importantly, the learner's interest in attending to the correction, all affected the degree to which recasts led to uptake.

Studies have shown that certain CF types, such as elicitation, generate 100% uptake (e.g., Lyster & Ranta, 1997; Sheen, 2004). Still another type, explicit correction, led to very low uptake in Sheen's (2004) survey of 4 English classrooms in 3 countries. These results are likely related to the conversational implications of these feedback types. In other words, elicitation (e.g., "Please say that again") demands a response, whereas explicit correction (e.g., "No, I would say, I *went* to the store yesterday") does not. Likewise, clarification requests (e.g., "What was that?") require a response, and so unsurprisingly generate high uptake. On the other hand, metalinguistic feedback (e.g., "There's a problem with the verb tense") and repetition also produced high uptake (Sheen, 2006), though in conversational terms neither type requires a response, in the same way as a clarification request or elicitation. Overall, however, in Sheen's survey the conversational implication of the feedback type seemed to predict high or low uptake. This is supported by the example of clarification requests, which generated high uptake but low repair rates, since learners tended to respond by repeating content, rather than reconsidering the linguistic form of the utterances. This led Sheen to stress that uptake may be a deceptive indicator of feedback effectiveness.

Along the same lines, Seedhouse (1997) focused on the contradiction between feedback types and the pedagogical function of classroom interaction. He criticized teachers for applying outside-classroom norms of avoiding giving negative feedback, as evidenced by teachers using indirect forms of correction, or refraining from correcting errors altogether. This finding was echoed by Lyster (1998), who reported that teachers' implicit negative feedback often was not noticed by learners, and that teachers often gave positive feedback to erroneous utterances. This is despite a prevailing overt instructional message that making errors is acceptable, and indeed a necessary stage in language development. Moreover, studies have shown that learners share this position (Yoshida, 2008) and consistently requested correction from their teachers (Nunan, 1988). Brown (2009), for example, found that learners wanted their errors corrected immediately, and wanted immediate explanations for errors, significantly more than teachers were willing to provide such support.

To complicate the issue somewhat, Yoshida (2008) found that learners of Japanese, while wishing their teacher to correct them, also expressed a preference for CF types that prompted them to self-correct. In other words, the learners in Yoshida's study did not necessarily prefer to be explicitly corrected. This suggests another reason why teachers might avoid giving explicit, negative correction. Offering more implicit correction might not simply be an attempt to avoid embarrassing learners, but might also reflect a methodological preference for allowing learners to self-correct. Yoshida's study partly concurred with Seedhouse's (1997), in that there was an apparent contradiction between the CF types learners preferred (i.e., types that elicited self-correction) and the type that teachers mostly provided (recasts). Recasts, which contain a corrected form, did not match learner preferences for self-correction. However, the teachers in Yoshida's study defended using recasts, partly because they were an efficient CF type in time-constrained lessons, but also because they felt that CF types that prompted self-correction were potentially face-threatening. Teachers suggested that learners might lose face if they were targeted in front of peers to correct themselves, and might lose even more face if they were unable to correct themselves satisfactorily.

Seedhouse (1997) stressed that clear negative correction is face-threatening in outside-classroom contexts, but not in the classroom, where it responds to learner demands, and fills an important pedagogical need. Seedhouse's discussion importantly resituated corrective feedback within the social context of the language classroom, and considered the particular roles and expectations that define this context. Yet there appears to be a disjunction between teachers' and learners' perceptions of what teachers' roles in the classroom should be, with regard to the amount that correcting errors is face-threatening. Ultimately, the prevalence of recasts suggests that it is not easy for teachers to achieve a balance between providing form-focused correction, while at the same time promoting communicative, meaning-focused taskwork.

2.3.4 *Relations between Contextual Factors and Corrective Feedback*

Sheen's (2004) survey of feedback in four countries' classrooms found that recasts resulted in markedly differing amounts of uptake and repair across contexts. Such inconsistent results involving recasts led Lyster and Mori (2006) to evaluate the effects of a classroom's communicative orientation on different types of feedback. They found that in a Japanese immersion classroom, where there was a focus on *accuracy* in oral production, corrective

feedback correlated with higher learner uptake than in a French immersion classroom, which displayed a greater communicative focus. A higher percentage of learner responses also included repair in the Japanese than in the French classrooms. To illustrate the contrast between the teachers' feedback approaches, Lyster and Mori reported that the Japanese teacher often provided recasts for learners, but then followed repaired learner responses with further formal explanation of the language point. On the other hand, the French teacher, after offering recasts, was observed not stopping learners from continuing telling stories. In fact the teacher's subsequent turn often focused learner attention on story content, rather than on the target form in question. The study speculated that the syntactic and orthographic similarities of English and French made meaning-focused lessons more viable, whereas the relatively greater cognitive demands of English L1 speakers learning Japanese favoured a more form-focused approach to teaching. Lyster and Mori did not suggest cultural differences as a possible explanation for the contrasting orientations towards corrective feedback. Yet their description of highly controlled, teacher-centred speaking activities in the Japanese classroom, with an emphasis on formal correctness, certainly typifies the language teaching approach in many East Asian classrooms (e.g., Han, 2005; Lee, 2001; Sullivan, 2008; Wen & Clement, 2003).

Sheen (2004) reported a similar pattern. In Korean classrooms, the native English teachers consciously avoided explicit types of corrections, so as not to disrupt the flow of meaning-focused conversation. The prevailing feedback type was recasts. Results showed that the Korean learners provided significantly more instances of uptake, and uptake plus repair, than did learners in Canadian classrooms. This raises the possibility that the Korean learners were more focused on producing accurate forms, and were more attuned to the corrective intention of teachers' responses, than the learners in Canadian classrooms. As with Lyster and Mori's (2006) description of a Japanese teacher's CF style, Sheen's Korean findings seem to reflect a prevalent East Asian emphasis on accuracy, which extends to an expectation that teachers will focus on linguistic form rather than meanings. Sheen did not make such a claim, but more cautiously suggested that learners in different settings may develop familiarity with corrective feedback styles, and expectations of repair associated with those styles.

2.4 Dynamic Assessment

2.4.1. Dynamic Assessment Terminology

- *Mediation*: The term used in Sociocultural Theory to describe assisted learning, where an expert intervenes to assist a novice in completing a task.
- *Mediatior*: The expert who intervenes to assist a learner in completing a task. Like an examiner, this role permits assessment of a learner's level of progress; unlike an examiner, a mediator also actively helps the learner achieve progress.
- *Static test/ Independent test/ Achievement test*: Terms that are often used interchangeably in the literature as a means of highlighting DA's unique qualities. Static tests are those that do not involve assistance from a mediator, and do not involve multiple tests over time. Achievement tests are those that measure learning that has taken place, but are not designed to assess a learner's proximity to higher levels of proficiency. An independent test is one that does not involve a mediator assisting a learner to complete tasks.
- *Graduated Assistance*: This refers to graded levels of mediator assistance, so that learners are only given just enough assistance to complete the task at hand.

2.4.2 Dynamic Assessment and Sociocultural Theory

DA is an approach to both assessment and language development which has emerged from the ideas of Vygotsky (e.g., Rieber & Carton, 1993) and later researchers within Sociocultural Theory. DA is relatively new to second language contexts, but has been used with people with mental disabilities or injuries, the elderly, learning-disabled individuals, pentitentiary inmates, preschoolers, and test-takers for university admission, among others (Sternberg & Grigorenko, 2002). Briefly, the approach involves examiners actively assisting learners to accomplish task goals, rather than only assuming a detached evaluative role (e.g., Anton, 2009; Poehner, 2007). This emphasizes tests' potential as learning activities, but interaction in DA also serves an evaluative purpose. By intervening with appropriate assistance, examiners can ascertain both skills that learners are able to control independently, and also how close learners are to controlling skills that the expert assisted them with (e.g., Poehner, 2008). Fundamental to DA, and to Sociocultural Theory, is the genetic model of mental development, which describes an ongoing interplay between already-developed biological and mental features,

and dynamic socio-historical forces (e.g., Lantolf & Thorne, 2006; Rieber & Carton, 1993). Moreover, this developmental interplay is facilitated by increasing mastery of tools, with the primary one being language. From an educational perspective, this genetic model translates into a view of learning as originally externalized, or mediated through experienced cultural insiders. Independent functioning only happens later, as we are able to internalize functions, through tools such as language, which we had previously only been using with others' assistance (Rieber & Carton, 1993). Thus Wertsch (1985) calls personality an "aggregate of internalized social relations" that have been transformed into functional practices and mental forms (p. 58).

More specifically, DA is one application of a developmental space conceptualized by Vygotsky, called the Zone of Proximal Development (ZPD). For Vygostky, the ZPD allowed for improvements to assessments that only measured current levels of development, but could not capture "those processes that are in the period of maturation" (Rieber & Carton, 1993, p. 200). Thus the ZPD space is defined by an individual's independent performance, at one end, and at the other end by what the individual is only able to accomplish with the assistance of a more experienced guide. At the same time, it is important to see the ZPD as created and fluidly redefined through a learner's interaction with a more experienced individual (e.g., Nassaji & Swain, 2000). As such, the ZPD activates a key concept in Sociocultural Theory – that new skills first emerge socially, and only later can be used independently. By offering a glimpse of an individual's proximity to higher levels of proficiency, interaction within the ZPD can provide information about the source of learner difficulties, and assist in developing appropriate future instruction (e.g., Anton, 2009; Lantolf & Poehner, 2008). In addition, practitioners emphasize that DA interaction should be more than simply an evaluation. DA advocates encourage examiners to abandon their evaluative role in order to assist learners towards attaining higher-level functioning (e.g., Guthke & Beckmann, 2000).

Because of DA's potential to reveal more information about an individual's proximity to higher developmental levels, and because some DA approaches allow examiners to individualize the mediation involved, supporters have suggested that it offers a fairer means of assessment crossculturally (e.g., Guthke & Beckmann, 2000; Sternberg & Grigorenko, 2002). This claim comes in the context of suspicions regarding cultural (and other) biases with achievement tests, and particularly standardized intelligent tests (e.g., Feuerstein, 1981; Guthke & Beckmann, 2000).

Feuerstein et al. (1981) argued that many tests designed to measure cognitive processes actually measure content familiarity, which has benefited cultural insiders over outsiders. The claim that DA may be more culturally equitable also has a theoretical basis. Generally the emergence of the ZPD responded to a concern that standard testing approaches often failed to reveal many learners' educational capacities (e.g., Campione et al., 1984). In contrast, Vygotsky stressed a holistic, contextualized model of mental development, seeing it as an intersection not only of individual factors but also social and historical ones, all of which are in constant flux (Rieber & Carton, 1993). From this perspective, DA is to some extent mandated to accommodate contextual factors such as learner culture. A difficult question, though, is how to transform this theoretical model of development into everyday practice. This is one of the challenges that DA and other socioculturally grounded approaches to learning face. For Sternberg and Grigorenko (2002), the research to date has not reliably demonstrated connections between DA test procedures and a Sociocultural model of learning. Despite DA's impressive theoretical basis, if we are unable to reliably evaluate the test procedures we use, it is difficult to understand their efficacy (e.g., Anton, 2009).

2.4.3 Dynamic Assessment Approaches

An important difference between DA and many achievement tests is that, through interaction with experts or more advanced peers, DA can provide tangible information about the learner's "developing expertise" (Sternberg & Grigorenko, 2002, p. 3), or proximity to more advanced developmental levels. Support for this claim comes from instances where learners with similar achievement test scores performed very differently in jointly accomplished tasks with experts or advanced peers (e.g., Aljaafreh & Lantolf, 1994; Poehner, 2007; Rieber & Carton, 1993). This contrasts with independent tests, which may claim to predict future potential, but are not designed to tap into that potential through the assessment itself. There is also the *dynamic* component of DA, which is based on the idea that social interaction reveals developmental needs, and also assists in integrating new information into extant concepts and processes. This transformation then creates a new point of departure, which initiates a new cycle of development, and the process continues (e.g., Lantolf & Poehner, 2004; Rieber, 1993). For this reason, another component of DA is that it is ongoing, and in practical terms typically involves a number of tasks over time, with subsequent tasks evaluating whether apparent learning has been retained, and

also can be applied to new situations (e.g., Campione et al., 1984; Lantolf & Thorne, 2006; Poehner, 2007).

DA researchers have developed a number of approaches. The differences between them tend to relate to the question of reliability, and more specifically the problem of how much to control expert assistance. It is easier to administer tests that are standardized (e.g., Sternberg & Grigorenko, 2002), and such tests are also easier to score (e.g., Lidz, 2000). Yet flexible interaction with the learner can provide key information for mediators, which can inform their choices for subsequent interventions (Feuerstein, 1981). Such interaction can also make the assessment experience less stressful for test-takers, and thus elicit their best performances (Lidz, 2000). Thus researchers have differed in their valuation either of intervention that is flexible as individual needs or circumstances vary, or intervention that is uniform across individuals.

DA approaches have also differed in the nature of mediator intervention. According to Sternberg & Grigorenko (2002), researchers have tended to use either a "sandwich" or a "cake" approach (p. 27). In the sandwich version, individuals are tested, followed by intervention, and then a post-test. The intervention may or may not be individualized. The cake version involves intervention *during* testing, and typically involves a graded series of hints that an examiner offers to an individual. Again, researchers must decide whether to strictly follow a list of hints, or use them as broad guidelines that can be altered to suit individual needs (Sternberg & Grigorenko, 2002).

Influential DA approaches include Feuerstein's (1981), which resulted in the Learning Potential Assessment Device (LPAD). The purpose of this instrument was to go beyond static psychometric testing to identify the processes involved in cognitive change, and then to help to bring about positive change in those processes. To achieve this, Feuerstein emphasized exposing individuals to cognitive processes and suitable tasks to develop them. As a result, the LPAD constitutes a battery of tasks targeting basic mental functions, with mediators assisting learners by intervening to correct errors. In addition, the process involves progressively more complex tasks, which build upon previous ones in terms of required skills. Feuerstein et al. (1981) argued that the LPAD allows mediators to ascertain learners' capacities to comprehend and apply underlying principles, their ability to respond to assistance, and their preferred learning styles. The mediator must also provide positive feedback, as Feuerstein considered affect, and

specifically motivation an integral factor in our capacity to learn. Feuerstein et al. (1981) reported several LPAD studies, which focused on both individuals and groups, and which persuasively supported the argument that the LPAD process can facilitate learning with children who scored poorly on static, one-time intelligence tests. While such results are impressive, the adaptability that is central to administering the LPAD also makes training mediators difficult, and makes it difficult to evaluate the approach's efficacy. This, in turn, complicates the important process of relating results to the cognitive functions that the instrument targets, and therefore validating specific components of the LPAD itself.

As with other DA proponents, Budoff did not accept that IQ tests or other static measures sufficiently revealed individuals' learning potential (e.g., Sternberg & Grigorenko, 2002). Unlike Feuerstein, however, Budoff sought to standardize practices to the point that researchers could reliably measure the effects of interventions. To that end, his Learning Potential Measures (LPM) came with detailed instructions, and mediator assistance was in the form of prescribed hints. Positively, post-test results from Budoff's LPM revealed greater learning potential than IQ tests (e.g., Sternberg & Grigorenko, 2002). On the other hand, Sternberg and Grigorenko (2002) suggested that, despite greater efforts towards reliability, Budoff did not rigorously explain or assess the LPM's connection to results.

A third influential approach is the Graduated Prompt Approach (GPA) (e.g., Campione et al., 1984), which attempts to measure learning ability, and also see whether learners can apply new skills to future tasks. The GPA operationalized the ZPD by using a fixed list of hints, moving from less to more specific help, in order to assist learners in completing various tasks. Its purpose was to assess whether learners (and groups of learners) distinguished themselves in terms of their facility at processing new learning. This was measured in terms of the number of hints that learners required to answer correctly. In addition, the idea of "transfer" was important, so learners also received increasingly novel tasks involving the same skills they had required in earlier tasks (Campione et al., 1984, p. 81). Results pointed to transfer ability being an important criterion, as "strong" and "weak" groups diverged most noticeably when they were required to use skills in tasks that were most dissimilar to earlier mastered tasks (Campione et al., 1984, p. 82). These results also lend support to DA's general emphasis on multiple assessment sessions, rather than single tests. A further advantage of the GPA is that it is standardized, which allows

for replication, and thus evaluations of the approach's validity. On the other hand, it is important to note that Campione et al. were looking at children with low academic success, and using relatively simple cognitive tasks (i.e., patterns of blocks and letters). This made it easier to devise, in advance, an effective list of hints. It is not possible to assume that the approach will work in other more complicated settings, such as evaluating adult second language learners. For more open-ended tasks, such as speaking test interviews, it may not be possible to produce graded hints that are suitable to all errors that may emerge.

2.4.4 Dynamic Assessment in Second Language Contexts

A relatively small number of studies of Dynamic Assessment (DA) with second language learners have been carried out. Within this small body of research, however, a variety of second language skills have been targeted, such as writing (Aljaafreh & Lantolf, 1994; Anton, 2009; Erben, Ban, & Summers, 2008; Nassaji & Swain, 2000), listening (Ableeva, 2008), and grammar use in speaking (Anton, 2009; Poehner, 2007). To date, studies have not targeted overall speaking proficiency, though spoken interaction has typically been the medium for negotiating correctness between participants and mediators. Many of these second language studies have compared DA results with independent tests, and persuasively presented case studies of learners with similar unassisted performances who revealed notable differences in language ability with mediator assistance.

Aljaafreh and Lantolf (1994) produced an influential study involving spoken corrective feedback of written errors. Their argument regarding corrective feedback was that both implicit and explicit types of instructor intervention can facilitate language development. This claim gains tacit support from corrective feedback research *without* conclusive results regarding which types of feedback best facilitate language development (e.g., Lyster & Ranta, 1997; Lyster, 1998; Sheen, 2004). Aljaafreh and Lantolf's position is grounded in Sociocultural Theory, which sees the aim of intervention as helping learners to move from assisted to independent performance. As such, they stress that feedback should only be provided if necessary, and only offered until learners are able to function independently. Because of this, the mediators in the study offered graduated assistance, if necessary, beginning with implicit prompts and adding more explicit information if the learners were not able to correct their errors.

In the study, tutors and learners talked about writing assignments, with a focus on a limited range of grammatical errors. Tutors did not follow a fixed list of feedback moves, since Aljaafreh and Lantolf emphasized the importance of the interaction being dialogic and negotiated. At the same time, tutors did follow *guidelines* for giving feedback, being sure to begin with implicit hints, and only adding more explicit information if learners required it. This stance is theoretically rooted, following Vygotsky's rejection of a simplistic transfer model of learning, which imagines knowledge passing from teacher to pupil. According to Vygotsky (Rieber & Carton, 1993), at best teacher/mediators can create circumstances that *facilitate* learning, which amounts to a complex interplay of social, historical and psychological forces. At the same time, Aljaafreh and Lantolf suggested that, on a small scale, DA may allow researchers to see "microgenetic" growth (p. 468). In concrete terms, this equates to evidence of decreasing reliance on mediator support in producing correct forms of targeted language points.

In the study, Aljaafreh and Lantolf's description of corrective interaction between tutors and pupils led them to develop a "Regulatory Scale" (p. 470; see Appendix A). This scale, which is used in the present study, amounts to a framework for analyzing corrective exchanges, beginning with more implicit mediator intervention, and moving towards increasingly explicit assistance as required. In their study, Aljaafreh and Lantolf suggested that measuring the degree of corrective support that mediators needed to provide against the Regulatory Scale could provide an indice of microgenetic growth. As such, the Regulatory Scale, which is based on the authors' data, and which is framed by Sociocultural Theory, laid the groundwork for formalizing DA as a procedure within second language contexts, including mediator training, administering sessions, and even allocating scores in terms of required assistance. On the other hand, Aljaafreh and Lantolf themselves only used the instrument as an *a posteriori* means of evaluating mediated performance. Using the Regulatory Scale as a script for mediator intervention runs up against Aljaafreh and Lantolf's emphasis on expert-novice interactions being negotiated and tailored to the learner and the situation, rather than prescribed. Indeed, the study's presentation of results is by means of protocols and qualitative analysis, rather than scoring. Therefore, it is not clear how Aljaafreh and Lantolf envisaged how future researchers would use the Regulatory Scale.

Overall, through protocols of conversations between tutors and learners, Aljaafreh and Lantolf persuasively showed DA's efficacy in discriminating between learners with similar

independent performances, in terms of the amount of assistance they require to produce correct forms. One problematic aspect of the study's presentation, however, is that the presented conversations appeared to be selected to illustrate the successfulness of the procedure. It is not clear whether, for example, any attempts at correction were abandoned for various reasons, were rejected by pupils, or revealed any shortcomings in the approach. In particular, Aljaafreh and Lantolf did not explore the central question of whether the degree of explicitness of feedback that a learner responds to is a fair indication of their language level and/or a sign of language development.

One subsequent study has re-examined Aljaafreh and Lantolf's Regulatory Scale. Nassaji and Swain (2000) addressed the question of whether following an implicit to explicit order of feedback best facilitated language development. In their small-scale project, Nassaji and Swain focused on correcting errors with written article use, either by following the Regulatory Scale's order, or by offering random feedback moves from the scale. Quantitative analyses counted instances of correct article use in compositions and a study-ending cloze test. Both this and qualitative analysis indicated that feedback using the Regulatory Scale's order facilitated article development better than using randomized moves. These results support an argument that the Regulatory Scale represents an effective framework for offering corrective feedback. The study's findings gain additional validity due to the fact that Nassaji and Swain regulated mediator interaction with participants, thus limiting one possible source of error (i.e., variability in mediator interaction with participants). Interestingly, the study also found that, with randomly given prompts, explicit feedback types led to almost double the number of correct article productions than implicit types. This is consistent with another corrective feedback study (Lyster & Mori, 2006), but possibly not consistent with the Regulatory Scale, which favours giving implicit prompts, and only adding explicitness if the learner is unable to produce a correct form (Aljaafreh & Lantolf, 1994). However, this apparent contrast may be misleading. DA does not eschew explicit assistance, but rather makes the explicitness dependent upon perceived learner need (e.g., Aljaafreh & Lantolf, 1994). Therefore, the DA learner's article use gains compared to the non-DA learner suggest that implicit prompts as part of a coherent approach may be a highly effective means of providing corrective support to learners.

With regard to limitations, both Nassaji and Swain, and also Aljaafreh and Lantolf's studies targeted a limited number of grammatical points, and involved sessions with a clear focus on correcting errors. This is understandable, since a narrow focus makes it easier to measure possible language development across sessions. This specific target may also facilitate learning, as learners can concentrate on a single language point, rather than simultaneously struggling with a number of forms. Finally, though, it is not possible to generalize the Regulatory Scale's advantages beyond the studies' targets, so more research is needed to determine whether the scale could facilitate similar gains in other communicative skills, and in other types of interaction, such as speaking tests.

Ultimately, a central issue with DA is whether it is possible at least partly to standardize the mediator's intervention. This step would allow test developers to improve DA's reliability through mediator training, task evaluation, and the development of fair rating schemes. On the other hand, standardization reduces the mediators' capacities to adapt their assistance to particular situations and learner needs. Poehner (2007) supported the idea that mediators need to adapt their interventions to the learners they are dealing with. He called for assistance that is "developmentally appropriate" for the learner in question (p. 324), indicating that mediators are responsible for determining the kind and amount of help that they offer. Correspondingly, Poehner's study used qualitative means of evaluating learner abilities, and he presented his findings through transcriptions of mediator-learner conversations. Poehner did suggest that scores could also be used to evaluate learner performance, though he did not explain how adaptable mediator interventions could be combined with a reliable scoring system.

As with Campione et al. (1984), Poehner (2007) also argued that for DA to be successful, mediators need to challenge learners to apply skills to increasingly difficult tasks. He demonstrated this point by presenting examples of two learners doing spoken story reconstruction tasks. While both learners showed the ability to transfer verb tense learning from one task to another identical task, only one learner could apply previous learning to a more difficult task. The other learner reverted to earlier, non-grammatical forms when faced with reconstructing the more difficult story. These findings persuasively supported Poehner's call for graded tasks, which in his study discriminated between learners with otherwise similar performances. However, the notion of developmentally appropriate assistance is rather vague.

Poehner claimed that "in DA, one continually alters both tasks and mediation in order to work successfully within a learner's ZPD, because individuals' abilities and corresponding developmental needs are always emergent" (2007, p. 333). This stance is understandable, but it limits our ability to evaluate whether any demonstrable connection exists between DA methods and learning outcomes.

Central to Poehner's (2008) subsequent DA study is the idea of reciprocity, or the learner's engagement in the mediation process. Like Lidz (2000), Poehner argued that if DA mediators wish to attune their interventions to the learners they are working with, they must increase their "sensitivity to learners' reciprocating acts" during DA sessions (p. 36). He pointed out that for many DA approaches, learner responses fall into only two reductive categories: incorrect (in need of more assistance) and correct (no more assistance required). In addition, Poehner questioned assessments that claim causality between learner responses to fixed prompts, and learners' proximity to higher levels of performance. He doubted that such an oversimplified model of mediator-learner interaction sufficiently captures the variability of explanations for learner behaviour, with consequences for both evaluating learners and understanding their difficulties. To enrich interpretations of the learner's contributions to DA, Poehner illustrated five forms of learner response to mediation, not including attempts at providing correct forms. The following list emerged from DA sessions focused on French verb tenses: (1) negotiating mediation (cases where learners and mediators negotiate the type of support that is needed); (2) use of mediator as a resource (cases where learners request support from the mediator); (3) creating opportunities to develop (cases where DA allows learners unexpected opportunities to learn); (4) seeking mediator approval (cases where learners ask mediators to evaluate their performance); (5) rejecting mediation (cases where learners refuse mediator support for various reasons). By drawing attention to often-ignored learner contributions to DA, the study successfully calls for a broader interpretation of how learners orient themselves to tasks and mediators. At the same time, by backing a fluid and situation-dependent approach both to administering DA, and also to interpreting DA interactions, Poehner somewhat undermines DA's possibilities as a formal assessment tool. This is unfortunate, since it seems possible that a compromise can be achieved, in which mediator assistance is not inflexible to the learner and

situation, but remains consistent enough to support a fair evaluation of learner responses to assistance.

Anton (2009) looked at DA with placement speaking and writing examinations. The study was part of a process of improving assessment procedures for undergraduate Spanish majors. In line with the present study's aims, I will focus on the speaking component. The study divided a picture narration task into two or three phases. In the first phase, learners attempted to tell a story with no help. In phase two, the learners re-attempted the task, but this time examiners intervened with "hints, direct instructions, or appropriate vocabulary that might improve the student's performance of the task" (2009, p. 584). In some cases, examiners added a third phase, which involved the examiner telling the story, before asking learners to attempt it again. The tests were scored based on generic rubrics targeting grammar, vocabulary, content, pronunciation and fluency. This scheme provided a numerical score; a qualitative assessment of the recorded test provided the rater's observations, an assessment of the learner's strengths and weaknesses, and recommendations for future study. Anton, like Aljaafreh and Lantolf (1994) and Poehner (2007; 2008), presented protocols of the narration task, contrasting two learners with allegedly similar independent performances, but notable differences in language ability during mediation. However, looking at the two learners' phase one protocols raises doubts about Anton's claim that, in a non-dynamic test, the two learners "may have been perceived as being at similar levels of language development" (2009, p. 591). Already in the first phase the coherence and accuracy of one learner's production appears notably superior to the other. On the other hand, the additional information that emerged in the mediated narration does lend support to the argument that DA can provide richer diagnostic details about learners.

2.5 Individualism and Collectivism

2.5.1 Individualism and Collectivism Terminology

- *Face (Saving/Losing face)*: Gudykunst (1998) nicely defines this idea as "the projected image of the self in relations with others," and describes individuals as continuously negotiating between threats to our own and others' face, and attempts to protect our and others' face (p. 122).
- *Attributions*: The explanations individuals give for others' behaviours. Some people attribute behaviours to individual traits; others attribute behaviours to contextual factors such as social role and situation.
- *Self-concept:* The way that we conceive of ourselves.
- *Cognitive Style*: This refers to different ways of conceptualizing reality, and of processing information, which appear to differ across cultures.
- *Relationality*: In this case the term refers to value judgments, perceptions and behaviours relating to individuals and others in their social context.
- *Communication Style:* Gudykunst et al. (1996) describe communication style as the preference for certain types of verbal and non-verbal coding that signal how messages should be interpreted/understood.
- *Direct/Indirect Communication*: These opposites differentiate direct communicators, who value clarity, factual accuracy and self-expression, on the one hand, and indirect communicators, whose style of communication is contingent upon social role, situation and other contextual factors.
- *Priming*: In I/C terms, this is a technique which stimulates individuals' latent individualism or collectivism, in order to evaluate their behaviours on subsequent tasks.
- *Ingroup*: An important group in a person's life, such as a company or family, to which they are deeply connected, and which often guides their social behaviour.
- *Self-construal*: Closely related to self-concept, this is the way that individuals present themselves in relation to the social environment that surrounds them.

2.5.2 Individualism and Collectivism Research: An Overview

This section reviews I/C studies in order to establish the construct's relevance to second language studies, and to describe the ways that research into the construct has progressed up to the present. Hofstede (1980) is generally accepted as the catalyst for contemporary research into I/C. Hofstede's research involved value surveys of multi-national corporate employees from forty countries. Results indicated that four overarching factors best discriminated between the survey respondents: power distance, individualism, uncertainty avoidance, and masculinity. Hofstede defined individualism and collectivism as two ends of a spectrum. At the one end, individualists are inner-oriented, or conceive of themselves and value themselves as distinct from a collectivity. At the other end, collectivists conceive of themselves and value themselves as indistinct from their social environment, a connection that "makes their existence meaningful" (1980, p. 215). Based on a survey of work goals, with items stressing independence from a company considered individualist, and items stressing a company's benefits to the individual considered collectivist, Hofstede ranked nationalities in terms of individualism.

The most influential aspect of Hofstede's study was also its principal limitation. The results established an argument for cultural differences at the level of nations, but in doing so raised questions about generalizability. Primarily, the validity of claiming national I/C levels based on 14 survey items which focused on work issues, and which were completed by respondents from a limited social class (i.e., white-collar corporate employees) is questionable. By generalizing from individuals, and even more specifically from individual employees of a multinational corporation, Hofstede's study did not address gaps in causality in linking respondents' personal value choices with national generalizations (Kitayama, Duffy & Uchida, 2007). In addition, in subsequent research nationalities that have similarly evidenced individualism or collectivism have differed in I/C sub-factors (e.g., Oyserman et al., 2002; Triandis & Gelfand, 1998), which points to the necessity of testing I/C across a broad range of perceptions, values and behaviours. Finally, on an individual level, succeeding studies have shown that individuals contain and can access both individualist and collectivist ways of thinking (e.g., Singelis, 1994; Trafimow, Triandis & Goto, 1991; Triandis, 1989), challenging Hofstede's model of I/C, which assumed high/low individualism to be the opposite of high/low collectivism.

Studies following Hofstede have sought to extend and retest his findings with different populations, and through more diverse conceptual frames, such as values, self-concepts, cognitions (mental processes) and relationality (social roles and relationships). Another major task in subsequent research has been an attempt to develop reliable instruments for measuring I/C. Finally, research has focused on another challenge created by Hofstede's study, which is the culture-individual interface, or how I/C generalizations at a cultural level can be related to individual-level attitudes and behaviours.

It is important to stress that, despite validity questions relating to Hofstede's results, a wide number of studies, typically comparing Euro-American and East Asian populations, have led to highly consistent support for Hofstede's original divisions of individualist and collectivist cultures (e.g., Gudykunst et al., 1996; Kitayama et al., 2007; Oyserman, Koon & Kemmelmeier, 2002; Singelis, 1994; Trafimow et al., 1991). A particularly strong endorsement for Hofstede's divisions came from Oyserman et al. (2002), who surveyed nearly twenty years of I/C research, and found consistent support for I/C's power to discriminate between cultures, as well as support for I/C's associations with values, communication style, attributions, and self-concepts. Still further endorsement for Hofstede's cultural divisions came from Kitayama et al. (2009), who found similar results from measuring I/C directly, through culture-related behavioural tasks, rather than indirectly accessing I/C through the traditional method of questionnaires.

The defining feature of individualism is that individuals consider themselves unique, or distinct from their social environment. For collectivists, meanwhile, ingroups are the primary social units, creating interdependent identities for its members (e.g., Oyserman & Lee, 2008). In varying formulations, these central components appear in all descriptions of the construct. Markus and Kitayama (1991) produced a coherent definition of I/C that integrates both ideas of self, predictions for behaviour, and how individuals will generally perceive others. They contrasted individualism and collectivism, following Hofstede (1980), by focusing on external and internal conceptions. They claimed that collectivists focus on external notions of self, meaning an emphasis on social role, statuses and relationships, which relate to group-oriented goals of belonging and fitting in. On the other hand, individualists are inner-focused, conceiving of themselves in terms of personal uniqueness, which is grounded in ideas, feelings and abilities. Individualists evidence a desire to express, or realize internal feelings and attributes, and to

achieve personal goals. In terms of regulating behaviour, Markus and Kitayama (1991) distinguished individualists using internal beliefs and feelings as a guide, with collectivists using perceptions of others' thoughts and actions as a guide.

These I/C formulations have been widely cited and adopted by subsequent studies (e.g., Singelis, 1994; Singelis & Brown, 1995; Triandis & Gelfand, 1998). Other studies have also enriched our understanding of the construct. In terms of relationships, for example, Triandis and Gelfand (1998) described collectivists as interdependent with ingroup members; however, individualists' relationships tend to be contractual, or exchange connections, meaning that they are valued in terms of the mutual benefits they offer the individuals taking part in them. In their study, however, Triandis and Gelfand's principal goal was to test the validity of an additional dimension – horizontal or vertical I/C – to account for the fact that broadly collectivist or individualist cultures have shown significant differences between them. According to Triandis and Gelfand, these differences may partly be explained by different attitudes towards power. While some cultures accept or endorse unequal power relations (vertical), others expect more equality in this regard (horizontal). Three studies' results supported the horizontal/vertical dimension's usefulness as an additional variable in I/C studies. Results also supported associations between these more detailed sub-divisions (such as horizontal individualism) and individual attitudes or behaviours. A benefit of the study was to suggest an explanation for I/C variation between and within cultures. Specifically the explanator was power relations, which could separate factors such as "competition" from direct I/C analysis. This then allows for a more detailed (four-part) assessment of cultural differences, which could (among other things) avoid low scale reliabilities caused by content validity problems. However, measuring I/C in terms of these four dimensions is time-consuming, and may not be feasible for all researchers.

Oyserman et al.'s (2002) meta-analysis of I/C research generated strong support for claims that Euro-Americans are indeed significantly more individualistic, and lower in collectivism, than many other nationalities. In terms of Euro-American and East Asian comparisons, reliable scales correlated with higher Euro-American individualism levels, and lower Euro-American collectivism levels, though not uniformly so when compared with Japan

and sometimes Korea.[1] Consistent support was evidenced in Oyserman et al.'s survey for claims that collectivists are more influenced by appeals to groups, and by contextualized information, than individualists. Preferences for context or situation-related attributions also described collectivists, but not individualists. Collectivists claimed much greater comfort interacting with ingroup members rather than strangers, which was not the case with individualists. Additionally, collectivists were shown to favour a more indirect communication style than individualists. In each of these cases, the opposite tendency was prevalent for individualists.

To sum up, cross-cultural research has consistently validated I/C as a powerful discriminant between cultures. Moreover, studies have reaffirmed this validity across a variety of factors, including values, self-concepts, relationality, and cognitions. A more complex conception of I/C continues to emerge, including an awareness that individuals appear to contain both sides of the construct, though socialization means that either individualism or collectivism tends to prevail. At the same time, Hofstede's original division of nationalities according to I/C has largely been ratified by succeeding research, not only through value surveys (e.g., Hui & Yee, 1998; Oyserman et al., 2002; Singelis, 1994), but also through priming studies (e.g., Oyserman & Lee, 2008; Trafimow et al., 1991; Utz, 2004; see 2.5.5 for a discussion of priming) and direct testing (Kitayama et al., 2009). Findings across cultures and using multiple methods have reinforced the central components of I/C, namely a sense of personal uniqueness and independence for individualists, and a sense of interdependence with ingroups for collectivists.

2.5.3 Individualism/Collectivism and Koreans

Oyserman et al.'s (2002) survey of I/C studies strongly disputed the validity of research that *assumed* I/C orientation based on nationality, or region (i.e., "East Asian"), rather than testing participants directly. For one thing, analyses of Japanese and Korean compared with Euro-American participants sometimes revealed Japanese and Korean participants to be higher in individualism than Euro-Americans, which was contrary to expectations (Oyserman et al., 2002). Such findings also indicate that I/C is likely to differ in its emphases from culture to culture. As my study deals with Korean participants, this section outlines I/C findings relating specifically to that cultural group.

[1] One doubt raised about this finding focused on the samples used. As with most studies, only university student samples responded. A number of researchers have associated this socio-economic group with higher individualism than the culture as a whole (Hofstede, 1980; Triandis, 1998; Oyserman et al., 2002).

Despite some between-study variability in findings, Oyserman et al.'s (2002) survey suggested that Koreans tended to be lower in individualism than Euro-Americans. Contrary to Hofstede's findings (1980), however, no difference existed between the two groups in terms of collectivism (Oyserman et al., 2002). Somewhat strangely, analyses indicated that when scale reliabilities were high, effect sizes were low, and vice versa. One possible explanation for this, and for the unexpected collectivism findings, is content validity problems, meaning that the scale items used may not have tapped into central I/C components for Koreans. Where scale content comparisons were possible, Oyserman et al. showed that the inclusion of items about *relating* to others significantly affected collectivism directions. When such items were included, Koreans rated significantly higher in collectivism than Euro-Americans. On the other hand, items focused on *belonging* to groups did not differentiate Koreans and Euro-Americans with regard to collectivism. It seems likely that both groups equally value belonging to groups, but differ with respect to the nature of relations between members.

Overall, Oyserman et al.'s (2002) meta-analysis did not show Koreans to be more collectivist than Euro-Americans. However, in terms of particular associations Koreans certainly fit the model of collectivists. Where collectivism was operationalized in terms of ingroup relations, Koreans were found to value collectivist correlates such as affiliation and sensitivity to rejection, but not personal uniqueness, which correlated negatively with collectivism. Similarly, when shown examples of advertising, Koreans found group-oriented appeals more convincing than individual-oriented ones. In terms of relationality, too, Koreans evidenced significantly more of a disjunction in comfort levels when dealing with ingroup members (high comfort) versus strangers (low comfort), than did Euro-Americans. In a related finding, Koreans differed significantly from Euro-Americans with regard to conflict resolution styles, preferring compromise as a strategy, while Euro-Americans showed preference either for avoidance or confrontation. With regard to attributions, when contextual information was available, Koreans used this information for proposing explanations for behaviours significantly more than Euro-Americans, who were more likely to use dispositional explanations. Taken together, these group-focused values and behaviours portray Koreans as more rather than less collectivist, and in ways that align them with central collectivist tendencies.

While Koreans generally followed typically collectivist patterns in the communication style studies that this study surveyed (see 2.5.4 below), one exception did occur. In Gudykunst et al.'s (1996) study, collectivist values did not predict a preference for using indirect forms of communication among Korean (or Japanese) participants. It was suggested, however, that the values survey used was too broad, and in fact narrower value categories did predict the use of indirect communication among Koreans. It appears that subtle individual-level variation may account for some communication style differences among Koreans.

2.5.4 I/C and Communication Style

One area of particular focus in I/C research has been communication style. Assessments of communicative preferences and behaviours have consistently discriminated between cultural groups in terms of I/C (Gudykunst et al., 1996; Gudykunst, 1998; Kim, 1994; Oyserman et al., 2002; Singelis & Brown, 1995). Gudykunst (1998) summarized research into I/C and communication, and presented a number of generalizations. At an individual level, most basically individualists prefer and evidence direct communication. In contrast, collectivists tend to communicate more indirectly. These terms are closely related to Hall's (1976) distinction of high and low-context communication. For high-context communicators (indirect), a relatively small amount of propositional content is encoded in the spoken message, so that meaning is largely embedded in, and must be inferred from the interactional context. On the other hand, low-context communicators (direct) mainly encode the propositional content in the verbal message, and depend very little on the context for inferring meaning. These differences reflect basic I/C modes of being. Individualists' behaviours seem primarily guided by achieving personal goals. These may be instrumental but also related to realizing individuality through expressing inner thoughts and feelings. In contrast, collectivists' indirect style relates to their other-orientation, meaning that behaviours are likely to be primarily motivated and regulated by others' perceived needs and expectations (Gudykunst, 1998). At the same time, in line with the argument that aspects of both individualism and collectivism are present in all groups, so too high-context communication occurs among individualists, and low-context communication among collectivists, usually in relation to the degree of social intimacy of the speakers (Gudykunst, 1998).

In addition to communication that seeks to express inner ideas and feelings, individualist talk is characterized by a valuation of clarity and precision. This accords with two of Grice's (1975) maxims for spoken behaviour, namely quantity (say no more than is necessary) and manner (speak as clearly as possible). Grice's other two maxims are quality (be truthful) and relevance (be pertinent to the context). Interestingly, Gudykunst (1998) suggested that good communication skills for collectivists may involve breaking at least three of these maxims, since truthfulness, precision and clarity tend to be subordinate to the demands of situation and social role. Thus Grice's maxims appear not to be universal, as some might believe. For instance, communication involving pauses, silence, hesitations, avoidance or concealing strategies appears to be more common for collectivists than for individualists. In addition, such actions, which are likely to be evaluated negatively by individualists, are more likely to be evaluated favourably by collectivists. Other communicative features differentiate individualists and collectivists as well. For example, collectivists generally show attention by reaffirming the interaction itself, offering short complementary and non-verbal expressions. On the other hand, individualists are more likely to show attention by content-focused comments and questions (Gudykunst, 1998). In addition, collectivists appear to be more sensitive to speaking-turn equality than individualists. While collectivists will carefully regulate turn length and distribution to ensure evenness, individualists have a tendency to take longer, monologic turns, and to accept more unequal turn distribution (Gudykunst, 1998).

Gudykunst's (1998) elicidation of I/C communication styles followed Gudykunst et al.'s (1996) study, which assessed correlations between communicative features and cultural frame. The study proposed that self-concept and values mediate cultural I/C at an individual level, and thus guide certain communicative behaviours. The first stage results indicated, as expected, that individualists preferred using inner feelings as a guide for communication, and preferred openness in communicating, all more than collectivists to a significant degree. Not all factors differentiated between I/C cultures, however, which Gudykunst et al. ascribed to individual variability. Second-stage results, though, after directly assessing participants' I/C, strongly supported predictions between self-concept differences and communication styles, across evaluated features. The same was largely true with values differences as well. Specifically, both values and self-concept predicted, for individualists, tendencies toward openness, using feelings

as a guide when communicating, and interpreting communicative behaviours in terms of personal attributes. For collectivists, interpersonal sensitivity was also predicted by self-concept and value surveys.

Kim (1994) assessed the relation between I/C and certain conversational constraints, meaning conceptual frameworks that guide communicative behaviour. Kim expected that other-oriented constraints, such as avoiding imposition, avoiding a negative evaluation by the interlocutor, and avoiding hurting others' feelings would be predicted by collectivism. Individualism, on the other hand, was expected to predict a concern for clarity, and a concern for effectiveness (i.e., achieving one's communicative goal). Kim asked participants to rate the importance of these constraints in six different request situations. Results showed only three constraints significantly discriminated between participants along I/C lines: concern for clarity (individualism), avoiding hurting others' feelings (collectivism), and minimizing imposition (collectivism). Of these, only concern for clarity produced a large effect size. The non-significance of concern for effectiveness and avoiding negative evaluation suggested either that the constraints influence both individualists and collectivists, or that the constraints did not clearly tap into basic I/C differences. On the other hand, the significant results supported expectations that, for individualists, achieving task goals guides communicative behaviour, preceding relational concerns. For collectivists, though, goal-orientation also exists, but appears to be regulated by, or contingent upon relational sensitivity (Kim, 1994).

To sum up, individualists, grounding communication in personal ideas and feelings, and motivated by a desire to achieve personal goals, typically value directness, clarity and consistency of style across contexts. Collectivist communication, though, while also motivated by personal or instrumental goals, is complicated by the additional goal of meeting others' expectations. This other-orientation necessarily affects collectivists' communication style, since interactive behaviours tend to adjust to the situation and interlocutor(s).

2.5.5 *Measuring Individualism/Collectivism*

Oyserman et al. (2002) criticized a lack of consistency in operationalizing I/C, pointing out that no less than 27 different scales had been developed since Hofstede's (1980) original questionnaire. In addition, very little replication had taken place, in terms of populations and methodologies. Where Oyserman et al. were able to compare studies, high variability was seen

between scales, combined with often low reliability ratings, and in some cases content validity problems. Meta-analyses revealed greatest reliability when basic elements of I/C were considered, rather than sub-elements of these constructs. For individualists, this basic feature was valuing personal independence; for collectivists, it was a sense of duty/obligation to the in-group.

Hui and Yee (1994) developed a compact I/C scale which did not achieve convincing reliability scores, and subcategory factor loadings were not uniformly satisfactory. One possible reason for this is that items did not adequately tap into underlying I/C differences. Subtle differences in the way Hui and Yee defined collectivism (i.e., in terms of attitudes towards *sharing*), compared to typical definitions of the construct (i.e., Markus & Kitayama, 1991), possibly weakened the scale. In addition, the target-specific design of the questionnaire also seemed ineffective. Hui and Yee argued that I/C judgments differ depending on context (i.e., at work, at home, etc.), and attempted to define those situational categories in terms of I/C components (e.g., "Colleagues and Friends" was combined with "Supportive Exchanges"). However, factor analysis did not generally support Hui and Yee's categories, suggesting that using situation as the primary explanator of individual I/C variability was not demonstrable.

According to Oyserman et al.'s (2002) I/C study survey, not only instrument concerns, but also variability between nationalities considered to be individualist or collectivist, both in degree of I/C, and also in factors correlated to I/C, as well as variability within populations contributed to the inconsistency of results. A clear conclusion was that between-scale differences in defining major I/C correlates such as attribution style, relationality, values, and communication style significantly affected results. This raised doubts about numerous studies that had generalized from small samples to whole nations, equated nations with regional blocks (such as "East Asia"), and assumed I/C based on nationality without testing it directly.

To control some of these confounds, Triandis and Gelfand (1998) divided an I/C scale into four sub-scales, adding vertical and horizontal dimensions to accommodate differences in perceptions of power relations. This approach aimed to differentiate cultures (or individuals) that share individualist or collectivist tendencies, but which differ in other respects. Findings supported the existence of these extra categories within typically individualist and also collectivist cultures. A second study also associated psychological features with the four categories, so that certain behaviours and attitudes predicted either vertical or horizontal I/C.

While this instrument obtained acceptable reliability ratings, adding two additional dimensions created a long instrument, which does not lend itself to small-scale research. In addition, it is questionable whether adding dimensions to measurements is a satisfactory shift in I/C measurement design. In this case, I/C and power relations were combined, but other potential mediating variables (i.e., gender, uncertainty avoidance, social class, age) could also add discriminating power. Yet combining constructs in this way adds methodological complexity, thus raising the possibility of measurement error. Ultimately this approach also weakens I/C's stand-alone explanatory power.

Similar to Triandis and Gelfand (1998), Gudykunst et al. (1996) combined several features (communication style, values and self-construals) into a single questionnaire. The large number of items and multi-dimensional design produced acceptable reliabilities, in terms of cultural-level I/C, but also allowed the researchers to test the associations between attitudes and communication styles. However, as with Triandis and Gelfand's (1998) instrument, the sheer number of items (nearly 350) makes the questionnaire unwieldy for a small-scale study.

Oyserman et al.'s (2002) meta-analysis pointed to the importance of reliably modeling the interfaces between low-level cultural frames and surface-level individual attitudes and behaviours. To validate the link between cultural I/C and individual-level measurements, some researchers have used a psychological technique called priming (e.g., Oyserman & Lee, 2008). According to this approach, researchers first engage, or make salient, an individual's individualism or collectivism, by a task that draws attention to one half of the construct. Then the effect of salient I/C is tested on subsequent tasks (e.g., Oyserman & Lee, 2008). An advantage of priming is the ability to demonstrate (through a control group design) strong links between I/C and correlated individual perceptions or behaviours. A second advantage is the ability to test the relative prevalence of one half of the I/C construct over the other.

An influential I/C priming study was carried out by Trafinow, Triandis and Goto (1991). This study tested the hypothesis that both individualist ("private") and collectivist ("public") cognitions are retrievable within individuals. The study also sought to ascertain how individualist or collectivist judgments were made. Chinese (typically collectivist) and Euro-American (typically individualist) participants were primed either for individualism or collectivism. Differences between priming groups, but also between nationalities were both significant,

reaffirming I/C expectations based on nationality. The priming results also indicated that both individualist and collectivist cognitions were accessible, but that relative I/C prevalence led to greater accessibility of cognitions from one half of the construct. Results confirmed that within nationality groups, priming significantly affected whether individualist or collectivist cognitions were accessed. Thus priming appeared to increase the salience of I/C for participants, leading to more consistent results in subsequent tasks.

A shortcoming of the study was that it assumed I/C based on nationality, yet found that, in the administered tasks, both Chinese and American participants produced a significantly higher number of *individualist* than collectivist statements. This suggests either that the I/C scale used did not measure I/C accurately, and/or that the Chinese participants (who were living in the U.S.) had developed more individualist perceptions through adapting to their American surroundings. It seems clear that measuring participants' I/C directly, rather than assuming I/C based on nationality, results in a more reliable baseline from which comparisons can be made.

Furthermore, the I/C instrument that was used is somewhat problematic. While the "20 Statements Task" (Kuhn & McPartland, 1954) is simple, it demands that raters code "idiocentric" (individualist) structures separately from "group cognitions" (collectivist) while excluding "allocentric" structures, which relate to qualities such as "interdependence, friendship, and responsiveness to others" (Trafinow et al., 1991, p. 650). Yet it is unclear how "group cognitions" and "allocentric" structures differ. Indeed, the two categories seem deeply related, and posited "allocentric" indices such as "responsiveness to others" appear to be aspects of collectivism. Due to the ambiguities related to determining what is, and what is not an individualist or collectivist structure in participant responses, this instrument's validity is questionable.

A priming study by Utz (2004) tested the relations between I/C primes and relative cooperation and sensitivity to a partner's behaviour during social dilemmas. This context was selected because according to Utz, social dilemmas are likely to cue more individualist or collectivist decisions, depending on whether personal or group goals are more dominant. Utz primed German university students either to focus on individualist terms (such as "individual," "self-contained," etc.) or collectivist terms ("group," "friendships," "together"). The social dilemma involved making decisions about giving money to a virtual partner during a computer game. Utz found that priming significantly affected decision-making shifts towards the half of

the I/C construct that was made salient (i.e., participants became more or less cooperative). A related finding was that participants' primes also shifted their responsiveness to a partner's behaviours. Overall, the study supports claims for priming's usefulness in testing associations between I/C and individual-level correlates. Another novel element of the study was that it tested actual behaviours, through a decision-making game, rather than attitudes towards behaviour through questionnaires. This adds strength to Utz's claims for the connection between I/C and cooperation. The study suggested that individuals' adherence to their cultural frame may be relatively flexible, and may shift towards one side of the construct or the other depending on situational factors, as other researchers have argued (e.g., Triandis & Trafimow, 2001).

Similar to Oyserman et al.'s (2002) meta-analysis of I/C studies, Oyserman and Lee (2008) carried out a systematic review of 67 I/C priming studies, and their relation to key psychological traits. Key findings were that priming both increases the salience of individualism or collectivism, and also supports clear links between cultural frame (I/C) and expected individual psychological differences in terms of values, self-concepts, relationality, and cognitions. The consistency of results has provided convergent support for the I/C construct itself and its correlates. With regard to types of primes, the study found that a wide variety of I/C stimulations produced consistent effects, but that the most reliable ones focused attention on values, self-concepts and relations with others. In addition, Oyserman and Lee indicated the usefulness of priming as a means of increasing the reliability of I/C measurement. The approach limits confounds from different sides of the I/C construct emerging in judgments regarding particular scale items and situations.[2] One widespread problem noted by Oyserman and Lee was that most studies surveyed in this article still assumed participants' relative I/C based on nationality, rather than measuring it directly. Moreover, by priming one half of the I/C construct, and then focusing on discrete I/C correlates, priming techniques do not lend themselves to ascertaining a baseline I/C orientation for individuals.

Kitayama, Duffy & Uchida (2007) presented a great deal of supportive evidence for the validity of I/C across a number of individual-level variables, including "style of action" (behavioural goals), self-concept, relationality, and cognition. Their survey focused on cross-cultural studies that have elicited actual behaviours and cognitions, rather than indirect

[2] Unfortunately, Oyserman & Lee (2008) did not publish internal reliability scores for I/C questionnaires used in conjunction with priming, which prohibits comparisons with the scales survey in Oyserman et al. (2002).

measurements through value questionnaires. This perspective allowed Kitayama et al. to evaluate how I/C reveals itself through often unconscious behaviours, and thus to describe the culture/individual interface in increasing detail. Although Kitayama et al. reported studies that accessed I/C through unconscious behaviours, they reaffirmed central conceptions of I/C (e.g., Markus & Kitayama, 1991; Oyserman et al., 2002). They confirmed that while individualists' actions tended to be directed toward personal goals and expressions of personal uniqueness and independence, collectivists' actions tended to be directed toward reaffirming relationships, maintaining group harmony, and meeting ingroup others' expectations. The survey argued that priming is not merely an experimental technique to stimulate I/C, but an embedded fact of culture. In other words, individuals may be socialized to develop a number of cultural frameworks that regulate and/or mediate individual experiences of typical life situations. Such frameworks may be linguistic (such as the use or non-use of subject pronouns), iconic (such as the Statue of Liberty, or Great Wall of China), or historical narratives (such as the story of voluntary settlement of the American West). This proposal suggests a way that nationalities with the same general cultural orientation may vary in its manifestations. It also suggests a way to predict individual behaviours and cognitions in terms of major cultural themes, which themselves mediate overarching cultural I/C. Finally, Kitayama et al. suggested that "implicit" measures of I/C, which elicit unconscious behaviours, may tap into the construct more reliably than questionnaires, which require explicit value choices.

A subsequent study by Kitayama et al. (2009) tested a related model of cultural-individual I/C, in which individual variation within cultures could be predicted by a mediating set of value themes. This means that individuals from the same broad culture may realize I/C differently on particular tasks, but these differences can be traced to a limited number of themes for that culture. Moreover, the study predicted that certain "cultural tasks," or largely unconscious behaviours that reinforce respective I/C themes, engage (as with priming) I/C for individuals. Results supported predictions, in that between-group responses to a battery of tasks effectively discriminated between cultural groups. This points to the effectiveness of the cultural task approach in revealing individuals' latent I/C. Certainly this shift to a more direct measurement practice seems sound. A difficulty in applying this approach, however, is that the tasks are necessarily highly specific, or limited to discrete features such as attention focus,

explaining behaviours, and the relative importance of the self in relation to others. To ascertain a coherent picture of an individual's relative I/C, this approach demands a battery of such tasks. For a small-scale project, carrying out such a wide range of tests with each individual participant may not be feasible.

According to Oyserman et al. (2002), the most widely used compact scale is Singelis' (1994) Self-Construal Scale (SCS). Singelis (1994), after Triandis (1989), and similar to Trafinow et al. (1991), argued for the individual retrievability of both individualist and collectivist perceptions, so that the SCS measures individualism and collectivism separately and provides separate scores for both. Singelis (1994, 1995), Singelis and Brown (1995) and Singelis et al. (1999) found that the scale effectively discriminated between cultures in terms of I/C self-construals.

Singelis followed Markus and Kitayama's I/C definitions (1991) in developing items for the SCS (1994, 1995). This is in line with Oyserman et al.'s (2002) conclusions that the most reliable questionnaires contain items that remain close to basic I/C elements, so as to limit confounds. Moreover, self-construals are considered a key mediating point between cultural-level I/C and individual-level ways of being (Gudykunst & Lee, 2003). While such features help to explain the SCS's popularity, the relatively small number of items may partly account for the SCS's generally low reliability scores (i.e., not consistently reaching a widely accepted Cronbach's Alpha threshold of .7 or higher). Singelis et al. (1999) acknowledged the need for ongoing development of the scale in order to obtain higher reliabilities. Yet Singelis (Personal communication, August 20, 2009) also argued that reliabilities above >.60 should be considered adequate given the broadness of the I/C construct, and the range of ideas and feelings being considered. Similarly, Gudykunst and Lee (2003) have argued that due to individual-level variability, or the varying extents to which individuals have adopted cultural norms, self-construal scales may not consistently register expected cultural-level I/C. However, this does not invalidate the scale. Instead, such variability points to attitudes where individuals may run counter to cultural expectations.

2.5.6 Summary of Individualsim/Collectivism Measurements

Researchers have developed a number of instruments for measuring I/C. According to Oyserman et al. (2002), a characteristic of the most reliable questionnaires is that items remain

close to I/C's basic elements: individualists' sense of uniqueness and independence, and collectivists' sense of duty or obligation to important ingroups. However, within-culture and individual-level variability may still affect the reliabilities of such questionnaires. Some researchers have argued that other factors, such as power relations or social class, are likely to predict this variability (e.g., Hofstede, 1980, Triandis & Gelfand, 1998), and have developed longer, more complex scales that can incorporate possible mediating variables (e.g., Gudykunst et al., 1996; Triandis & Gelfand, 1998). Although strong reliabilities point to these longer instruments as effective I/C measures, the sheer size of the scales makes them unfeasible for small-scale research.

The strength of results of another measurement technique, priming, indicates that this method could be used in conjunction with other tasks or questionnaires. However, certain difficulties argue against using priming. First, as Kitayama et al. (2009) pointed out, priming makes specific aspects of I/C salient, not the construct as a whole. This narrow focus means that ascertaining a general picture of an individual's I/C requires administering a large battery of tasks. This is not necessarily workable in a small-scale study. Related to this difficulty is the fact that priming normally makes salient either individualism or collectivism, but not both. This technique lends itself to control-group designs, which can demonstrate the difference that priming makes on subsequent tasks. This, in turn, can demonstrate relations between I/C and the attitudes or behaviours tested in the follow-up tasks. However, priming may not be suitable for studies with few participants and no control group, or where an aim is not to confirm cultural-level I/C similarities, but to find correlations between I/C differences and subsequent task performance. For the same reason, priming only one half of the construct may not be desirable, since there is wide support for the claim individuals contain both individualist and collectivist tendencies (e.g., Singelis, 1994; Trafimow et al., 1991; Utz, 2004). Therefore focusing attention on only one half on the construct may mask key differences between individuals from the same culture.

Kitayama et al. (2009) have reported the success, in terms of reliabilities, of priming culture through tasks that elicit unconscious cultural orientations. While this approach appears to be the most direct means of eliciting I/C differences, each culture task targets a single factor related to I/C. As such, ascertaining a picture of an individual's general I/C orientation involves a

wide battery of tasks. As with explicit primes, therefore, this approach is not suitable for small-scale research.

One reason that Singelis' SCS (1994) has been widely used is that it is relatively short, and thus manageable for small-scale researchers. It has the added advantage that its items relate closely to basic I/C elements. On the other hand, it has not always demonstrated Cronbach's Alpha reliabilities above a standard of .70. It is worth pointing out that the SCS is not unique in this regard. No compact, widely-used questionnaire has consistently showed Cronbach ratings of >.70 (Oyserman et al., 2002). These results raise legitimate concerns about the usefulness of these measures, suggesting that questionnaire items may not consistently tap into I/C's core distinctions, and/or that simply more items are needed (e.g., Gudykunst et al., 1996; Triandis & Gelfand, 1998). Another possible explanation is that scale items relating to I/C (i.e., values, self-construals, attributions, relationality and communication style) may also tap into individual variability. Singelis argued that Cronbach Alpha reliabilities above .60 should be considered acceptable, given the breadth of the I/C construct (Personal communication, August 20, 2009), and Gudykunst and Lee (2003) likewise stressed that self-construal scales serve both to indicate cultural orientation, but also to point to places where individuals differ from cultural norms. This type of model was proposed by Kitayama et al. (2009), so that variation within cultural groups can be predicted by the prevalence of certain I/C sub-features (but not others) for individuals. The difficulty with this claim is that it becomes difficult to know whether to interpret a scale's reliability scores as legitimate indications of unreliability, or as revealing points where individual and their underlying cultures diverge.

At present, even while widespread agreement exists about I/C, its manifestations at an individual level, and its importance in explaining cultural differences, a sure-fire method of capturing the dynamic of cultural frame and individual variation remains a work in progress.

2.6 Conclusions: Connecting Individualism/Collectivism to Speaking Tests

The present study is motivated by a concern that English oral proficiency interviews[3], and particularly the IELTS™ speaking test, may disadvantage test-takers with collectivist orientations. In this section, I will synthesize key points from the preceding literature review that support my concern about a cultural bias in the IELTS™ speaking test.

- Evidence from the literature suggests that collectivist test-takers will be less comfortable revealing personal information and offering opinions to the examiner than individualist test-takers. This claim finds support in both the I/C literature (e.g., Gudykunst et al., 1996; Gudykunst, 1998; Oyserman et al., 2002), and also the literature on speaking tests. In the latter, several studies stressed the cultural relativity of speaking test interviews (e.g., He & Young, 1998; Kim & Suh, 1998; Ross, 1998; Young & Halleck, 1998). Collectivist test-takers may not understand the overriding schema in English speaking tests, which is that tasks and in-test interaction primarily serve to elicit sufficient talk for rating the test-taker's linguistic proficiency (Ross, 1998). In other words, examiners and raters ascribe relatively little importance to the social context of the interaction, which is seen simply as a frame for eliciting a language sample. Due to source-culture differences, even in cases where this schema *is* understood, collectivist test-takers may not be willing or able to behave in status-challenging ways, such as talking more than the examiner, or providing lengthy status-sensitive types of talk, such as providing opinions and personal information (e.g., Kim & Suh, 1998; Ross, 1998; Young & Halleck, 1998). IELTS™ speaking test questions, however, focus heavily on offering personal information and opinions (UCLES, 2005). I/C research, too, has indicated that individuals' willingness to express themselves openly, and the people to whom individuals are comfortable expressing themselves, are points of divergence for individualists and collectivists (e.g., Gudykunst et al., 1996; Gudykunst, 1998; Kim, 1994; Oyserman et al., 2002). Individualists tend to value open self-expression regardless of interlocutor. Collectivists, on the other hand, perceive appropriate behaviour, including communicative behaviour, as contingent upon the expectations and roles implicit in speakers' interpersonal contexts

[3] By traditional oral proficiency interviews, I mean one-on-one interviews in which examiners do not interact with participants in a meaning-focused manner, but simply prompt them to provide information and opinions on a variety of topics.

(Gudykunst, 1998). Kim (1994), for example, found that "minimizing imposition" was a primary guide for communicative behaviour among collectivists, over and above concerns for clarity and achieving communicative goals (p. 143). In a related finding, unlike individualists, collectivists' levels of comfort in communicating tend to vary greatly depending on whether the interlocutor is an ingroup or outgroup member (Gudykunst, 1998). For these reasons, and allowing for individual differences, it appears collectivists may be less willing to offer opinions and personal information in the IELTS™ speaking test, particularly to a stranger examiner.

- Related evidence suggests that collectivist test-takers may be less likely to provide extended answers, and possibly be more likely to display silences and hesitations, both of which are likely to negatively affect scores generated by the IELTS™ scoring descriptors (IELTS™, 2009). For one thing, verbosity does not tend to be valued in East Asian collectivist cultures (e.g., He & Young, 1998; Ross, 1998; Young & Halleck). In Japan, for example, in interview contexts, lower-status test-takers typically are expected to be concise in their responses to higher-status interviewers (Ross, 1998). Devalued behaviours for individualists, including silences, pauses or hesitations, are typically valued more highly in collectivist than individualist cultures (Gudykunst, 1998). Indeed, silences and hesitations may serve subtle communicative ends in collectivist cultures, though such behaviour has tended to be regarded simply as an indication of linguistic incompetency when transferred to English (e.g., Young & Halleck, 1998). Moreover, this interpretation has stubbornly been applied even to advanced EAL users (e.g., Ross, 1998; Young & Halleck, 1998). Although there is no question that silences and hesitations, which feature prominently in the IELTS™ descriptors (IELTS™, 2009), may indicate a lack of communicative fluency, collectivists' general valuation of concision and devaluation of verbosity appear to disadvantage them for the IELTS™ speaking test.
- Related evidence points to the importance of contextual features, including the interlocutor's status, expectations and behaviour, for regulating collectivist speakers' talk (e.g., Gudykunst, 1998; Holtgraves, 1992; Kim, 1994). In contrast, a typical individualist's communication style is characterized by an emphasis on clarity and self-expression, and consistency across interpersonal contexts (e.g., Gudykunst, 1998; Kim,

1994). These features reflect individualists' primary valuation of the self as unique, and a desire to express that uniqueness (e.g., Gudykunst, 1998; Markus & Kitayama, 1991; Oyserman et al., 2002). For collectivists, meanwhile, the general conclusion from a number of studies was that interpersonal sensitivity was the primary guide for communicative behaviour (Gudykunst et al., 1996; Gudykunst, 1998; Kim, 1994), including avoiding imposition (Kim, 1994), and in another (non-I/C) study, showing awareness through politeness behaviour of power relations between speakers (Holtgraves, 1992). Given these differences, low-interaction speaking tests such as IELTS™, which emphasize independent production, and devalue contextual sensitivity, both in terms of the test format, and in terms of scoring criteria, appear to disfavour collectivists and favour individualists. To increase standardization and improve reliabilities, such speaking tests strictly limit examiner contributions as an interlocutor. This removes an apparently vital source of guidance for collectivist spoken behaviour. In addition, by encouraging extensive talk from the test-taker, often concerning personal information and opinions, the schema for speaking tests such as IELTS™ subverts, or downplays the social context of the interaction. The interview context is formal, with status differences separating examiner and test-taker, who are not familiar with each other. Despite this, test-takers are expected to speak more than examiners, are rewarded for taking initiative in extending their responses, and are expected to speak at length about topics that may be, from a collectivist point of view, socially inappropriate given the status differences and lack of intimacy with the examiner. In contrast, in a collectivist (Korean) speaking test interview, Kim and Suh (1998) showed how participants' subtle reinforcements of status difference (with examiners) and role expectations represented the most important criteria for test-taker success. In the IELTS™ test, this skill is neither integrated into the test format, nor recognized by the scoring criteria. Overall, the IELTS™ speaking test's emphasis on providing extended personal information and opinions, its restrictions on examiner involvement in interview talk, and its status-challenging schema all appear to favour individualist and disadvantage collectivist test-takers.

2.9 Research Questions

Specific factors related to I/C, and aspects of the format of English speaking tests such as the IELTS™ exam, suggest that such tests may favour learners with individualist rather than collectivist orientations. On the other hand, DA may represent a more culturally equitable approach to administering speaking tests, because the examiner interacts more freely with the participant. Comparing results of DA and a Non-Interactive (NI) speaking test format in terms of degree of participant I/C, the study addressed the following questions:

1. Is there a difference between participants' NI and DA scores, as measured by the IELTS™ scoring descriptors?
2. Is there a difference between participants' NI and DA scores, in terms of gains on successive tests?
3. Is there a difference between participants' DA scores, as measured by the IELTS™ scoring descriptors, and their DA scores measured by the Regulatory Scale?
4. What is the relation between participants' culture, as measured by degree of I/C, and their DA and NI scores?
5. What is the relation between variability in I/C scores and characteristics of DA corrective exchanges, as realized in test data recordings?

CHAPTER 3 – METHODOLOGY

3.1 Participants

In total, seven participants took part in the main study. At the time of the study, the participants were all enrolled in intermediate/high-intermediate-level second language classes at a private language school in British Columbia. This level was sufficiently high for the participants to complete the Self-construal Scale (SCS) in English, which was a condition of carrying out the data collection at the language school. After initial recruitment did not generate enough volunteers, it was necessary to widen the targeted levels to include high-intermediate students.[4] All participants completed a background questionnaire (Appendix B).

All participants were Korean. Choosing an East Asian nationality follows studies that have questioned individualist language teaching practices in East Asian contexts (Spratt et al., 2002; Sullivan, 2008; Wen & Clement, 2003). One reason for choosing Korean participants in particular was that, although Koreans have generally evidenced higher collectivism and lower individualism, they have also shown notable variability in scores (Oyserman et al., 2002). Therefore there was good reason to expect individualism/collectivism (I/C) differences between participants who share the same cultural background. This would allow me to compare I/C differences with speaking test score differences, and also to compare Dynamic Assessment (DA) corrective exchanges in terms of I/C differences. Choosing only one nationality also controlled a potential source of variability. Finally, a practical consideration was that the private language school's enrollment at the time of the study was predominantly Korean, thus increasing the likelihood of recruiting a satisfactory number of volunteers.

Six participants were female, and one was male. The mean age was 21.9 years old. On average the participants had been living in Canada for 2.1 months before the beginning of the project. None of the students had studied abroad before, and only three of seven had been outside of Korea before, on short sightseeing holidays to English speaking countries. Six of the seven participants had started studying English in Korea from the first year of Junior High School, resulting in an average of 10.6 years of formal English education. None of the participants had any experience with the International English Language Testing System (IELTS™) exam, or

[4] The private language school divides classes into 8 levels, based on entrance and monthly placement tests. The participants in this study came from the school's levels 5 and 6.

with IELTS™ preparation courses, although three students had previously taken the Test of English for International Communication (TOEIC®) exam, and one had taken the Test of English as a Foreign Language (TOEFL®) test. Table 1 summarizes pertinent information about the seven participants.

Table 1

Main Study Participants' Information.

Participant	Age	Length of Residence	Formal English Schooling	Previous Trips Abroad	Previous English Tests
1	26	3 months	6 years	None	TOEIC® (2005); 830 pts.
2	21	1.5 months	10 years	4 weeks (Phillipines)	TOEFL®; did not remember score
3	20	3 months	10 years	None	TOEIC® (2009); 700 pts.
4	22	1.5 months	12 years	1 month (Singapore, Malaysia, Australia)	None
5	23	2 months	13 years	None	None
6	20	1 month	10 years	1 week (U.S.)	None
7	21	1 month	10 years	None	TOEIC® (2009); 835 pts.

Note: TOEFL®: Test of English as a Foreign Language; TOEIC®: Test of English for International Communication

The participants were recruited on a voluntary basis. I posted notices in the private language school's intermediate classrooms, which briefly explained the study, encouraged students who wanted extra speaking practice to join, and provided contact details. I organized individual schedules for the six speaking tests with each participant. All tests took place within a four-week period.

3.2 Instruments

3.2.1 Self-Construal Scale

Participant culture, operationalized as degree of I/C, was measured using Singelis' SCS (1994). The questionnaire is reproduced in Appendix C. This instrument contains 30 items written as declarative statements, and uses a seven-point Likert-type scale. Half of the randomized items relate to individualism, and the other half relate to collectivism. Dividing the scale this way is in line with the widely-shared claim that individuals contain and can access both individualist and collectivist perceptions, though one tends to predominate (e.g., Singelis, 1994; Trafinow et al., 1991; Triandis, 1989). By adding the respective I/C items' scores, then dividing the result by the number of items (15), raters arrive at a mean score for individualism, and another for collectivism.

Advantages of this instrument included its compact size, which was suitable to the small scale of this project. Its items also appear to tap into basic elements of individualism (a valuation of personal independence and uniqueness) and collectivism (a sense of duty/obligation to ingroups), a focus that was endorsed by Oyserman et al. (2002). In addition, Gudykunst and Lee (2003) argued for the importance of self-construal scales for pointing to the ways that individuals mediate cultural norms. In other words, with a sufficiently large sample, self-construal scales will tend to ratify expected cultural orientations, while surface item-by-item variability reveals the different emphases that individuals place on culturally relevant factors. Therefore he argued that low scale reliabilities do not necessarily reveal construct validity problems. Instead low reliabilities may indicate simply that individuals variously realize shared cultural norms. Following this argument, the present study used the SCS as a means of differentiating individuals within the same culture, in order to compare I/C differences with speaking test performances.

3.2.2 Simulated IELTS™ Speaking Tests

To control for between-test difficulty, the tests were six different IELTS™ practice speaking exams (UCLES, 2005; for an example see Appendix D). This exam claims 1.4 million test-takers annually and recognition as a language qualification by over 6,000 higher-educational institutions and employers worldwide (IELTS™, 2009). The speaking test is one of four test

components – the others being reading, listening and writing – and takes between 11 and 14 minutes to complete. It contains three sections. In the first section, the examiner asks test-takers questions about familiar topics. In the second, the examiner gives a topic to the test-taker, who has one minute to prepare a talk, and then should speak about the topic for two minutes. In the third section, the examiner asks the test-taker more general questions, which tend to relate to the topic in part two.

The principal reason for using IELTS™ practice speaking tests relates to examiner-participant interaction. During the IELTS™ speaking test, examiner interaction is limited to asking questions, repeating questions if necessary, prompting the test-taker for more information, and giving encouraging gestures such as smiling and nodding (W. Everett, personal communication, March 4, 2010; C. Ebert, personal communication, June 6, 2010). This appears to be the limit to the examiner's contributions to the test interview, despite the fact that online promotional material emphasized the "interaction" and "discussion" in part three of the speaking test (UCLES, 2009). None of the three speaking test sections requires the test-taker to jointly accomplish a task with the examiner. This emphasis on independent production is also reflected in the nine-point speaking band descriptors that IELTS™ uses for speaking test scoring (IELTS™, 2009). The descriptors do not mention skills that Riggenbach (1998) described as interactive, in the sense of communicating successfully with an interlocutor: the ability to claim turns, maintain and yield turns, backchannel, ensure comprehension, and engage in effective repair exchanges.

3.2.3 Regulatory Scale

The six speaking tests were divided into two formats (i.e., three tests for each format). One format (NI) involved little interaction between the examiner and participants. The other format (DA) involved greater examiner-participant interaction, in the form of examiner encouragement and corrective feedback. The corrective feedback followed Aljaafreh and Lantolf's (1994) Regulatory Scale (RS; see Appendix A), which showed its effectiveness (compared with random feedback) when used as a frame for assisting learners in correcting writing errors (Nassaji & Swain, 2000). The RS amounts to a sequence of correcting moves which increase in explicitness if the learner is unable to self-correct. Aljaafreh and Lantolf (1994) did not prescribe specific utterances for each level of the scale – the determining characteristic of

each prompt was that it should be slightly more explicit than the previous one. In this way the RS is guided by its emphasis on learner self-correction, rather than examiners explicitly providing correct forms.

Unlike previous RS studies (e.g., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000), and unlike other DA studies involving spoken corrective exchanges (e.g., Anton, 2009; Poehner, 2007), the present study used the RS to generate scores for each participant (see section 3.4.1.3). The RS targets DA corrective exchanges, which the IELTS™ band descriptors (IELTS™, 2009) does not account for. This novel application of the RS represents a means of standardizing DA for oral proficiency testing, while retaining its interactive characteristics. Additionally, in terms of analysis, by using the RS to generate quantitative data, it became possible to compare RS scores with scores generated using the IELTS™ descriptors. This comparison is comparable to ones that previous DA studies have made between independent and assisted performance, although those studies used qualitative data (e.g., Ableeva, 2008; Aljaafreh & Lantolf, 1994; Anton, 2009; Poehner, 2007). Moreover, scoring the RS allowed me to compare participants' scores with their degree of I/C, which was a primary purpose of the study. I did not consider the RS to be an alternative to the IELTS™ descriptors, as the RS focused only on corrective exchanges, which themselves targeted only grammatical or lexical items. As such, the RS is not a global measure of oral proficiency. However, I was interested in assessing whether the RS scoring system could reveal differences in lexical and grammatical performance between students. This could point to the efficacy of the DA test format over the NI test format, and also point to the benefit of including the RS as a component of (or supplement to) an improved system for scoring speaking tests.

In terms of examiners using the RS for offering correction, while it originally offered a corrective feedback sequence for *written* errors (Nassaji & Swain, 2000), with very minor modifications it was adaptable to a speaking test. In the original RS, the writing tutor begins by focusing attention on a sentence with an error, then more narrowly on a clause or phrase, and adds specificity until the learner is able (or not able) to correct the error (Aljaafreh & Lantolf, 1994). On the other hand, with the modified RS, the examiner focuses the participant's attention on a short stretch of speech that contained an error.

3.2.4 Email Survey of Participants' Perceptions of DA Tests

As an additional support (or challenge) to quantitative and qualitative analyses of DA tests, I developed a short Email Survey (see Appendix E). The questions elicit participant perceptions of the DA format, and particularly the corrective assistance that the examiner provided. As part of an assessment of DA's efficacy for formal speaking tests, the questions focused on the DA format and examiner's perceived effects on participants' accuracy orientation (i.e., their focus on producing correct sentences). An additional focus was participants' perceived comfort with the format, as well as the causes of that comfort (or lack thereof).

3.3 Data Collection Procedures

3.3.1 Individualism/Collectivism Measurement

To test for reliability before the main study, the SCS was piloted with 33 English language students at a private language school in British Columbia. The participants were volunteers, and were recruited from intermediate-level classes. Therefore the pilot group had approximately the same language level as the main study group. An intermediate or higher level was necessary as the school's English-only policy did not allow me to translate the questionnaire into the participants' first languages. With regard to nationality, it was not possible to find 30 volunteers who were the same nationality as the main study group (i.e., Korean). Therefore, the participants came from a number of different nationalities. The piloting took place during regular class times, with the permission of the classes' instructors. I briefly explained my research to the students, who then completed a consent form before doing the SCS. If any language questions arose relating to one of the items on the scale, either the class teacher or I explained the meaning of the statement as clearly as possible. Participants asked very few such questions, and after they received an answer, they expressed confidence in their understanding. For these reasons, I determined that administering the questionnaire in English was acceptable.

After assessing the internal reliability of the piloted SCS data, each main study participant completed the SCS before they carried out the first speaking test. The participants' SCS forms were then used to calculate I/C scores.

3.3.2 Administering NI and DA Speaking Tests

In the present study, I took on the role of examiner, administering six speaking tests to each participant (n=7). Although the exam tasks were practice IELTS™ speaking tests, I am not a certified IELTS™ examiner. The instructor of the private language school's IELTS™ exam preparation class coached me on how to administer the speaking tests. However, I have been teaching EAL classes for ten years, and have administered many speaking placement tests, exit speaking tests, and am trained to administer and rate two other standardized speaking tests, the First Certificate in English (FCE) and Certificate in Advanced English (CAE) exams. This experience leads me to believe that I administered the tests to an acceptable standard. I am also currently an instructor at the same language school, but at the time of data collection had not

previously taught, and did not know the seven participants in the study. Finally, while I have my own opinion about what constitutes a good speaking test, I made every attempt not to express any bias towards either of the two test formats (NI and DA) used in this study, both during test administration and also while I rated the participants' tests.

The speaking test administrations took place in a spare classroom at the participants' private language school, and lasted approximately 20 minutes each. All tests took place within a four-week period, and were recorded using a small digital voice recorder. These recordings were used later to score the tests and for qualitative analysis.

The six practice IELTS™ speaking tests were divided into two test formats (i.e., three tests for each format), NI and DA. The formats differed in terms of the amount that the examiner interacted with the participants during tests. To control for practice effects, I randomized the tests in the following manner:

Participant 1: NI1 --> NI 2 --> NI 3 --> DA1 --> DA2 --> DA3
Participant 2: NI 2 --> NI 3 --> DA1 --> DA2 --> DA3 --> NI 1
Participant 3: NI 3 --> DA1 --> DA2 --> DA3 --> NI 1 --> NI 2, etc...

During the 3 NI tests, the examiner's interaction was limited to the following moves: (a) Greeting the participant (i.e., "Good afternoon"); (b) Reading the test paper's frame for the test questions (e.g., "In the first part I'm going to ask you some questions about some familiar topics"); (c) Reading the test questions; (d) Acknowledging the participant's response (e.g., "Thanks;" "OK"); (e) If necessary, repeating or rephrasing test questions[5]; (f) Offering encouraging gestures, such as smiling and nodding; and (e) Concluding the test interview (e.g., "OK, that's the end of the test. Thanks").

The other three tests used the DA test format, which involved more interaction on the part of the examiner. Specifically, in addition to the NI style's interactive moves, the examiner provided corrective support. The "support" component included the same friendly examiner gestures (e.g., smiling and nodding) as the NI tests, while adding encouraging responses to

[5] My colleague at the private language school, who teaches the school's IELTS™ preparation course, indicated (W. Everett, personal communication, Mar. 4, 2010) that examiners could repeat questions if necessary. He did not mention rephrasing questions to aid comprehension. A number of test questions were incomprehensible to the participants, due to the level of vocabulary. In such cases, I rephrased the questions to allow for comprehension. It felt unnatural not to rephrase questions in this way.

participant turns (i.e., "That's interesting;" "Wow;" "I didn't know that"). However, these did not include evaluative comments (i.e., "great;" "good") that participants might have interpreted in terms of their oral proficiency level. The intention of the support during the DA tests was to encourage further talk, reduce nervousness and frustration, and to show interest. Such support follows Wen and Clement's (2003) recommendations for improving collectivist learners' communication skills. They argued that learners' motivation to communicate will be enhanced by (a) teacher support, meaning high involvement and "closeness" with students; and (b) a feeling that goals can be accomplished together, even at the micro level of task completion. Feuerstein (1981), who developed DA methodologies and instruments, also stressed the importance of affective support as a component of expert assistance.

The "corrective" component of DA corrective support involved myself intervening during tests to assist participants in correcting errors. This approach follows previous DA applications in second language teaching and learning studies (e.g., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000; Poehner, 2007). Corrective support followed the RS, which reflects DA's principal purpose of facilitating independent performance, and only supporting learners to the extent that they require assistance (Aljaafreh & Lantolf, 1994). These corrections were extensive, meaning that that the examiner did not target a discrete, pre-determined type of error, such as verb tenses or article use, as was the case in previous second language DA studies (Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000; Poehner, 2007). Instead, corrections in the present study addressed errors as they arose. However, corrections were also limited to lexical and grammatical items that occurred during the participants' speaking turns[6]. Not determining a priori which errors to target has the advantage of tailoring correction to individual needs. DA's flexibility in this regard is considered an important characteristic of the approach (Aljaafreh & Lantolf, 1994; Poehner, 2008). The fact that previous studies have limited the focus of examiner assistance to specific types of errors seems to run counter to this emphasis on individualizing assistance. On the other hand, this limited focus is understandable in order to ensure the validity

[6] Originally my intention was to target errors across the four categories in the IELTS™ speaking test band descriptors (i.e., Fluency and Coherence; Lexical Resource; Grammatical Range and Accuracy; Pronunciation). In the end, however, I determined it would be difficult to correct Fluency and Coherence errors (e.g., errors with connecting words) and Pronunciation errors. Because it is difficult quickly to detect discursive or organizational errors, and because it is difficult quickly to offer help with the mechanics of producing individual sounds in English, I decided to limit corrections to lexical and grammatical items.

of claims for DA's efficacy for language development. In the present study, the somewhat broader range of error correction was also in line with a purpose of this study, which was to consider the merits of DA as a means of evaluating general oral proficiency.

In practice, the key difference between previous RS interaction (Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000) and interaction in the present study was this study's application of the RS to meaning-focused speaking tasks. As a result, it was necessary to take advantage of short pauses in streams of task-focused talk to interrupt and point out an error. To this end, I used a short utterance ("Sorry?") that quickly alerted the participant to an error in the preceding utterance, but which was also maximally implicit. A second, related characteristic of RS use in the present study was that I did not correct all participant errors. Persistently interrupting participants could negatively affect the coherence of their responses, and could also negatively affect their confidence during the test, both of which could impact their speaking test scores. Lastly, for the same reason, I decided not to follow the RS through all ten available corrective moves (see Appendix A). If participants did not respond to my first prompt with a correct form, I offered a maximum of two additional prompts before providing the correct form. As a result, my RS corrective moves were limited to the following:

1. Examiner indicates that something may be wrong in a speaking turn ("Sorry?").
2. Examiner narrows down the location of the error (i.e., examiner repeats the specific speaking turn that contained the error).
3. Examiner indicates the nature of the error, but does not identify the error (e.g., "There was something wrong with the verb tense there").
4. Examiner provides correct form.

It is important to note that not all corrective exchanges in the present study followed these steps. In some instances, I abandoned my attempts to correct participants because they resisted the interruption or did not notice my corrective intention. In other cases, I skipped additional prompts because participants' responses to my preceding prompt(s) clearly indicated that they did not know the correct form. In those cases, in order not to lengthen my interruption unnecessarily, I provided correct forms.

I provided participants with their speaking test scores as soon as I obtained them. After all tests were rated, I emailed participants a summary of their scores, in addition to a brief written

evaluation. The latter included my impression of the participants' strengths as speakers, the areas in which they had improved, and areas where they needed further improvement. I also thanked the participants for volunteering for the present research project. All participants expressed satisfaction at having volunteered for the study.

3.4 Data Analysis

3.4.1 Preliminary Analysis

The first stage of analysis involved scoring the SCS, as well as scoring the speaking tests. For the latter, raters used both the IELTS™ descriptors, and also the Regulatory Scale to determine participants' scores.

3.4.1.1 Scoring the I/C Questionnaires

I entered SCS scores from pilot participants into a spreadsheet using the Statistical Package for the Social Sciences (SPSS, Version 16) software, and then assessed internal reliability using Cronbach's Alpha. Cronbach's Alpha scores for the piloted SCS are presented in Table 2. For the 15 individualism items, the score was .63, which was lower than a generally accepted threshold of .70. On the other hand, the score falls within the .60 to .70 range that Singelis (Personal communication, August 20, 2009) considered acceptable, given the breadth of the I/C construct, and the fact that scale items dealt with a variety of ideas and feelings. For the 15 collectivist items, the score was .43, which was far below the .70 standard for acceptable reliability. However, further analysis revealed that only two collectivist items had large negative correlations with other items[7]. Therefore I recalculated Cronbach's Alpha without these two items. The remaining 13 items' score was .61. This remains below .70, but falls within Singelis' acceptable range. Another reason for considering these scores adequate is the fact that a variety of nationalities took part in the pilot study. Other studies have stressed that nationalities which are predominantly individualist or collectivist often differ sharply in terms of the I/C sub-factors that they emphasize (e.g., Oyserman et al., 2002; Triandis & Gelfand, 1998). For this reason, pilot study participants who obtained similar I/C scores may have responded differently to particular items, which would affect internal consistency results. As a result, I determined that the individualism scores and the scores of the 13 highly-correlated collectivist items showed adequate reliabilities. Therefore, I proceeded to administer the questionnaire, minus the two negatively-correlated collectivist items, with the participants in the main study group.

[7] It is unclear why these two items (see Appendix C) correlated negatively with other collectivist items. Both items appear to relate closely to a primary collectivist valuation of duty and obligation to the ingroup (e.g., Oyserman et al., 2002).

Table 2

Self Construal Scale Internal Reliability Scores

Cultural Orientation	Number of Items	Cronbach's Alpha
Individualism	15	.634
Collectivism	15	.433
Collectivism (without item 17 and 23)	13	.612

For each main study participant, mean scores for both individualism and collectivism were calculated. Scoring the SCS questionnaires involved adding the scores for individualist items (n=15) and collectivist items (n=13) and dividing the total by the number of items. This produced a mean score for each participant for both individualism and collectivism. These scores allowed me to carry out correlational analysis to assess whether relations existed between degree of I/C and scores in the two speaking test types. The SCS scores were also used to examine interaction patterns in DA tests in terms of participants' I/C differences.

3.4.1.2 Scoring with the IELTS™ Descriptors

For the speaking tests, two raters (including myself) independently scored six tests for each participant using the IELTS™ speaking test descriptors (IELTS™, 2009). The IELTS™ descriptors distinguish speakers with nine bands, with a 9 being the highest level and score (IELTS™, 2009). These bands are subdivided into four categories (Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, Pronunciation). In terms of each category, the 9 bands describe how a speaker of that band level is likely to perform. For example, in the "Lexical Resource" category, band 6 states the following:

• Has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies.
• Generally paraphrases successfully.

(IELTS™, 2009).

The raters listened to the digital files on computers using audio software. Before all speaking tests were scored, the raters held a training session with the IELTS™ preparation course instructor at the instructor's private language school. We read the descriptors, discussed the criteria for achieving particular band scores, then together rated two speaking test recordings. We compared our scores and discussed discrepancies in order to establish a working standard. In line with oral examiner training for other speaking tests, I determined that a one-point discrepancy between raters was acceptable[8]. During the training session the raters agreed to re-rate tests with a discrepancy greater than one.

In the present study, for each test and for each of the four categories, raters decided upon a full or half score (i.e., 5 or 5.5). The raters then added the scores and divided them by four to achieve a mean score for that test. The participants' final scores were the mean scores of the two raters. Rater reliability was 100%, meaning that all scores fell within one point of each other. This indicates that the two raters were consistent in using the IELTS™ descriptors.

3.4.1.3 Scoring with the Regulatory Scale

The three DA tests were scored a second time using the Regulatory Scale (RS). The scoring system targeted the principal interactive component in the DA tests, which was corrective exchanges. Firstly, scoring with the RS entailed identifying Corrective Exchanges in the DA tests, which I defined as containing the following elements:

1. A participant makes an error in his/her speech.
2. The examiner intervenes to indicate that an error has occurred (i.e., "Sorry?"; Level 1 on the RS); OR the participant self-corrects without examiner intervention (Level 0 on the RS).
3. If the participant is unable to self-correct, the examiner offers up to 2 additional prompts, following the RS.
4. Either the participant, OR, if she/he is unable to, the examiner offers a correct form.

Once I had identified a corrective exchange, I counted the number of corrective moves that the examiner used before the participant was able to self-correct. This became the participant's score for that exchange. At the end of the test, I added the scores from all corrective exchanges, and

[8] A one-point variation between raters was described to me as acceptable during a training session for oral examiners for the Cambridge First Certificate and Certificate of Advanced English speaking tests (2009, Personal Communication).

divided the total by the number of exchanges that took place. This mean score became the participant's RS score for that test. Scores ranged from 0 (self-correction before examiner intervention) to 4 (unable to self-correct; the examiner offered the correct form). The following are examples of each score from the DA test recordings:

(1) RS score 0

Participant 2: Koreans usually use … uh drive a car.

(2) RS score 1

Participant 2: So we … celebrate our New Year's day on February usually.
Examiner: Sorry?
Participant 2: In February [usually.
Examiner: [good.

(3) RS score 2

Participant 4: They choose one day and they went – go out – goes out.
Examiner: Sorry?
Participant 4: They, uh, famous movie stars or singers goes out.
Examiner: Goes out?
Participant 4: Uh, yes … go out.
Examiner: Go out. Yeah, OK, good.

(4) RS Score 3

Participant 3: But my mom nowadays usually does stock on the computer computer.
Examiner: Sorry?
Participant 3: Stock [on the computer.
Examiner: [She she <u>does</u> stocks?
Participant 3: Yes.
Examiner: Can you think of another verb? Not not does.
Participant 3: Mmm (…) Invest invest.
Examiner: Yeah, I could say she <u>invests</u> on the computer.

(5) RS Score 4

Participant 3:	In the middle of Seoul there is mountain, and [there is
Examiner:	[So – Sorry?
Participant 3:	In the middle of Seoul there is mountain.
Examiner:	In the middle of Seoul there is mountain?
Participant 3:	Yes.
Examiner:	You missed an article.
Partcipant 3:	Ah. There is a middle – a middle of mountain?
Examiner:	Almost. A mountain.

I did not give a score for certain corrective exchanges. One such instance was when the examiner did not begin a corrective exchange with "Sorry?" but with a more explicit prompt. In other cases the examiner determined, based on participants' responses to the initial "Sorry?", or to a subsequent corrective move, that participants did not know a correct form, and would not be able to self-correct. In these cases the examiner provided a correct form before offering the maximum three prompts. In terms of scoring, I gave such corrective exchanges a 4, which is equivalent to example (5), above, where the examiner used 3 prompts, and then corrected the participant.

The highest score for each corrective exchange, 0 (zero), may appear to confuse self-correction, which implies forms that a participant knows but has not completely integrated into their second language, with more routine slip-repair sequences that often occur during connected speech. Moreover, self-correction does not appear to be a result of interaction, which is the key variable distinguishing the two test formats in the present study. However, Aljaafreh and Lantolf (1994) persuasively argued that the "collaborative frame" that an expert and a learner create together is fundamentally different from the orientation learners have to their language when they are alone or with non-experts. In other words, the presence alone of the examiner, in a speaking context oriented towards formal accuracy, can focus a learner on monitoring the accuracy of learners' speech. It is likely that some self-corrections may have been slip-repair sequences. Nevertheless, my impression during test administration and while listening to the recordings was that, in the context of corrective support, including self-correction in the RS

scoring system was justified. Similarly, I limited the number of corrective moves in order to increase score reliability. This limited the effect of single errors that participants could not correct on their overall score.

3.4.1.4 Speaking Test Scores over Successive Tests

Gains or losses in scores were calculated for both the non-interactive (NI) tests and Dynamic Assessment (DA) tests. This was calculated as the positive (or negative) difference in score between the first and last NI tests, and between the first and last DA tests. These included scores on NI tests that the IELTS™ descriptors generated, scores on DA tests that the IELTS™ descriptors generated, and RS scores for DA tests. Finally, I also calculated gains from the first to last test regardless of test format as a comparison for gains with a specific format (i.e., NI or DA). For all six IELTS™ practice tests the tasks were identical, raising the possibility that participants would become more comfortable with, and perhaps more proficient at taking the tests. On the other hand, different content in the 6 tests may have mitigated practice effects.

DA approaches emphasize administering several tasks over time (Aljaafreh & Lantolf, 1994; Lantolf & Poehner, 2004; Lantolf & Thorne, 2006). From an evaluative standpoint, this allows examiners to ascertain whether learners have consolidated gains. From a pedagogical standpoint, this allows learners to apply new learning to subsequent tasks, with an ultimate goal of independent (unassisted) performance. I was interested in assessing whether the corrective support in DA tests related to increased scoring gains compared NI tests, which did not feature corrective support.

3.4.2 Second-Stage Analysis

3.4.2.1 Correlating Individualism and Collectivism with Speaking Test Scores

Correlational analyses were carried out to determine whether relations existed between degree of I/C and participant performance on the NI and DA speaking tests, participant RS scores, and participant gains with RS scores. Due to the small sample size, I determined that nonparametric statistics were suitable (e.g., Pett, 1997), and so I used Spearman's rank correlation coefficient (Spearman's Rho) for the analysis. In addition, I created a high individualism group (HI), and a high collectivism group (HC). The resulting groups contained two participants each. The four participants demonstrated the greatest *difference* between

individualism and collectivism in their questionnaire results. Qualitative analysis then explored the question of whether differences existed between the HI and HC groups in speaking test corrective exchanges, which represented the principal interactive component of the tests.

3.4.2.2 Analyzing Corrective Exchanges in Terms of Participant Individualism and Collectivism

Two participants' SCS scores showed notably higher individualism than collectivism scores (HI group). Two other participants' SCS scores were the opposite: the collectivism score was notably higher than that for individualism (HC group). Using these two groups' DA test recordings, I carried out qualitative analysis that looked at different types of responses during corrective exchanges, in order to assess whether I/C differences related to communication style differences. This analysis followed Poehner's (2008) analysis of DA and learner reciprocity, or the ways that learners varied in their responses to corrective support. In the present study, the first stage involved listening to the speaking tests and identifying types of responses to the examiner's corrective interruptions. I included interaction types that were outside of corrective exchanges, but that focused on correctness, or diverged from expected routines involving examiner questions followed by participant responses. In the second stage, I counted instances of response types, to determine if any types occurred more frequently with the HI or HC groups. The following examples from the DA speaking tests illustrate the response types:

1. <u>Responds as correction:</u>

The response indicated that the participant understood that the examiner's first prompt ("Sorry?") was corrective:

(6)

Participant 2:	So we … celebrate our New Year's day on February usually.
Examiner:	Sorry?
Participant 2:	In February [usually.
Examiner:	[good.

2. <u>Attempts self-correction</u>:

These were instances outside of corrective exchanges where participants attempted to correct themselves, whether or not the original utterance was erroneous, and whether or not the form they replaced it with was correct. I did not include slips, where participants produced an erroneous form and quickly (i.e., without a reflective pause) corrected that form, in this category. I defined attempted self-corrections as instances where participants deliberately stopped their flow of speech to reconsider their selection (i.e., with a reflective pause), and then offer a new selection:

(7)

Examiner:	Do you think they will get more popular?
Participant 3:	Yes, they will get more popular because tech- techniques…technology will improve and develop more.

3. Initiates accuracy check:

These were instances outside of corrective exchanges where the participants themselves either requested a correct form, or requested (often through interrogative intonation) confirmation that a form was correct:

(8)

Participant 4:	Actually I have never seen the Peking Opera before, but I saw some pictures. On the picture…in the picture?
Examiner:	Oh, in the picture.

4. Attempts correction after minimal prompt:

These were instances where the participants began to correct an erroneous form before the examiner could finish uttering his first corrective prompt ("Sorry?"):

(9)

Partcipant 2:	The Korean team went to the semi-finals, so it- it was huge- it would be huge– it-
Examiner:	So-
Participant 2:	A:::h, no, it is- it was huge for Koreans.
Examiner:	Good.

5. Ambiguous corrective exchange:

These were instances where one (or more) participant responses inside corrective exchanges were other than expected attempts at self-correction. In other words, there was evidence of misunderstanding in the corrective exchange. In some cases, the responses took the form of defending the correctness of the targeted utterance. However, these responses may also have been affirmative answers to perceived clarification requests, since the first two corrective moves from the RS that the examiner used ("Sorry?", and repetition of the erroneous utterance) may be interpreted in that way:

(10)

Participant 3:	If you go at night, it's awesome. You can see whole Seoul, and if weather-
Examiner:	Sorry?
Participant 3:	Whole- whole Seoul.
Examiner:	Ah. You can see whole Seoul?
Participant 3:	Yeah.
Examiner:	I would say, you can see the whole of Seoul.

In other cases participants, at some point during the corrective exchange, chose not to attempt self-correction, but instead requested a correct form. I identified such instances as "Ambiguous corrective exchange" rather than "Initiates accuracy check" because the former occurred within examiner-initiated corrective exchanges. Such instances also included unexpected breakdowns of corrective exchanges' routine of the examiner providing prompts, and the participant attempting to self-correct:

(11)

Participant 4:	They, uh, make-up. They get make-up.
Examiner:	Sorry?
Participant 4:	Make-up.
Examiner:	Ah, you you [you need a you need a different verb.
Participant 4:	[Even though- What's that?
Examiner:	They they get make-up?

Participant 4: They get make-up? That's right? They make-up, like a …
Ugh!

Examiner: I- I understand. I would say, they they wear make-up.

Participant 4: Oh yes, they wear make-up, even though they are a man.

6. Participant Takes Initiative:

These were instances where participants temporarily took control of managing speaking test events from the examiner. In most cases this took the form of participants reversing expected examiner-question/participant-response routines in the speaking tests by asking the examiner a topic-related question or a question relating to the test format. I did not include requests to clarify test questions in this category. Such questions related critically to participants' ability to provide a relevant answer, and, therefore, I determined that they were in line with expected examiner-participant routines. The following is an example of a temporary control-taking move:

(12)

Participant 3: We passed the line of country.

Examiner: Mm hmm.

Participant 3: The person who was from Mexico checked my passport and my family's passports, but they didn't, uh… They usually give stamps, right?

Examiner: Yeah, that's right. I think so.

Participant: But they didn't give stamps, they just allowed to pass the line of country.

After counting the types of corrective (and other) exchanges that took place between participants and the examiner, I looked for similarities and differences within and between the HI and HC groups. The subsequent stage of qualitative analysis involved assessing these differences and similarities in terms of I/C.

As an additional means of supporting (or challenging) the findings from this qualitative analysis, participants responded to an Email Survey (see Appendix E) which focused on differences that emerged between the HI and HC groups. I used the survey responses to compare participants' perceptions of the DA tests with my analysis of the recorded data.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Results

Results will be presented in five sub-sections, organized in terms of this study's research questions.

4.1.1 Is There a Difference Between Participants' NI and DA Scores, as Measured by the IELTS™ Scoring Descriptors?

Tables 3 and 4 present participants' scores on six practice speaking tests, which were rated using the International English Language Testing System (IELTS™) descriptors. The two tables distinguish the test format. Table 3 presents scores for the three Non-Interactive (NI) format tests, and Table 4 for the Dynamic Assessment (DA) format tests. The tables also include, for the two test formats, mean scores as well as the participants' (positive or negative) gains from the first to the last test.

Table 3

Participants' Non-Interactive (NI) test scores, including mean scores, +/- change from first to last test, group mean scores and group mean +/-.

Participant	NI Test 1	NI Test 2	NI Test 3	Mean	+/- Change
1	6	6	5.8	5.9	-.2
2	4.9	5	5.3	5.1	+.4
3	5.4	6.1	6.3	5.9	+.9
4	6.2	6.1	6	6.1	-.2
5	4.8	5	5.5	5.1	+.7
6	5.4	6.3	6.4	6	+1
7	6	5.7	6	5.9	0
Group	5.5	5.7	5.9	5.7	+.4

Note. Because the 3 tests were randomized, the above order represents the order the participants received the tests. NI= Non-Interactive

Table 4

Participants' Dynamic Assessment (DA) test scores, including mean scores, +/- change from first to last test, group mean scores and group mean +/-.

Participant	DA Test 1	DA Test 2	DA Test 3	Mean	+/- Change
1	5.1	5.7	6.2	5.7	+1.1
2	4.8	5	5.9	5.2	+1.1
3	5.2	5.9	6	5.7	+.8
4	5.6	6	6.5	6	+.9
5	5.8	5.7	5.9	5.8	+.1
6	5.7	6	6.2	6	+.5
7	6	5.7	5.6	5.8	-.4
Group	5.5	5.7	6.0	5.7	+.6

Note. Because the 3 tests were randomized, the above order represents the order the participants received the tests. DA= Dynamic Assessment

Table 5 presents the participants' scores on all six practice speaking tests, regardless of format. This description provides a comparison for specific format (i.e., NI and DA) gains.

Table 5

Participants' practice speaking test scores, plus +/- change from first to last test, regardless of format, and group mean +/- change

Participant	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	+/- Change
1	6	6	5.8	5.1	5.7	6.2	+.2
2	4.9	5	4.8	5	5.9	5.3	+.4
3	5.4	5.2	5.9	6	6.1	6.3	+.9
4	5.6	6	6.5	6.2	6.1	6	+.4
5	4.8	5	5.5	5.8	5.7	5.9	+1.1
6	5.7	6	5.4	6.3	6.4	6.2	+.5
7	6	6	5.7	6	5.7	5.6	-.4
Group	5.5	5.6	5.7	5.8	5.9	5.9	+.4

Note. Because the 6 tests were randomized, the above order represents the order the participants received the tests.

The speaking test results indicate that, overall, test format (NI or DA) had little effect on mean speaking test scores. The mean score of tests from both formats was the same (5.7).[9] Individually, too, six of seven participants' mean scores for the NI and DA tests were within .2 points of each other.

4.1.2 Is There a Difference Between Participants' NI and DA Scores, in Terms of Gains on Successive Tests?

In terms of gains between the first and last tests, the mean for NI tests (+.4 points) was the same as the mean for all tests, regardless of format (see Tables 3 & 5). Based on this result, the gains for NI tests are just as likely to be a result of familiarity with the test, including becoming accustomed to the test format, as they are to be a result of administrative style. For DA tests the average gain was slightly higher (+.6), though the gains may also be a result of test familiarity or chance.

4.1.3 Is There a Difference Between Participants' DA Scores, as Measured by the IELTS™ Scoring Descriptors, and Their DA Scores Measured by the Regulatory Scale?

Table 6 presents Regulatory Scale (RS) scores for participants, as well as gains (or losses) from the first to last DA test. With regard to mean scores, five of seven participants improved overall from the first to last tests, in terms of the number of prompts that they required in order to self-correct.

[9] Due to the fact that the participants' average scores for the 2 test formats were identical, I did not carry out quantitative analysis to determine statistical significance.

Table 6

Participants' Regulatory Scale scores on Dynamic Assessment (DA) tests, plus +/- change from first to last test, and group mean +/- change.

Participant	DA 1	DA 2	DA 3	Mean	+/- Change
1	1.7	1.4	1.6	1.6	- .1
2	2.1	1.6	1.3	1.7	- .8
3	2.5	3	1.5	2.3	- 1
4	2.7	2.3	.7	1.9	- 2
5	1.7	2.8	2.3	2.3	+ .6
6	1.7	1.2	1.1	1.3	-. 6
7	1.6	1.8	2.2	1.9	+ .6
Group	2	2	1.5	1.9	- .5

Note. Negative scores are favourable, as scores indicate the mean number of examiner prompts required for participants to self-correct. DA= Dynamic Assessment

The four participants (P2, P3, P4 and P6) who improved the most in terms of RS scores also showed strong improvements in terms of DA test scores using the IELTS™ descriptors (See Table 4). On the other hand, Participants 5 and 7, whose gains on DA test scores using the IELTS™ descriptors were below group average, correspondingly worsened (rather than improved) in RS scoring change. Participant 1's RS gains and IELTS™ descriptor scores did not match each other as closely. Her IELTS™ descriptor scores also strongly improved from DA test 1 to 3, while her RS scores did not vary a great deal: she scored uniformly well from DA test 1 to 3.

4.1.4 What is the Relation Between Participants' Culture, as Measured by Degree of Individualism/Collectivism, and Their DA and NI Scores?

Table 7 presents the I/C mean scores from the Self-construal Scale (SCS) for the seven participants in the study. The SCS measures individualism and collectivism separately, based on a widely-supported model of the culture-individual interface in which individuals access both sides of the I/C construct, though one tends to predominate (e.g., Singelis, 1994; Trafimow, Triandis & Goto, 1991; Triandis, 1989).

Table 7

Individualism/Collectivism Mean Scores from Self Construal Scale (SCS), and mean scores for each category, with Standard Deviation (SD)

Participant	Individualism	Collectivism
1	4.6	5.2
2	4.2	5
3	5.7	4.1
4	6.1	5.2
5	3.1	4.8
6	3.9	4.1
7	5.2	4.9
Mean	4.7 (SD = 1.1)	4.8 (SD = .47)

Note. Mean scores were calculated based on a Likert-type 7-point scale.

Results indicate that, on average for the seven participants, neither individualism nor collectivism predominated. Individually, despite sharing the same nationality (Korean), the participants differed considerably in their individualism and collectivism scores. It is important to note that the sample size is too small to generalize to other Korean groups. However, the results support this study's single-nationality design, which anticipated (following Oyserman et al., 2002) that there would be notable between-individual differences in terms of I/C. This within-nationality variability allowed me to carry out the following analyses: to ascertain whether correlations exist between I/C differences and speaking test score differences, as well as to carry out qualitative analysis with speaking test data for High Individualism (HI) and High Collectivism (HC) groups.

Table 8 presents the results of correlational analyses comparing participant I/C scores (SCS scores) and speaking test scores.

Table 8

Spearman's Rho Correlations Between Individualism/Collectivism Scores, Non-Interactive (NI) Test Scores, Dynamic Assessment (DA) Test Scores, Regulatory Scale (RS) Scores and Regulatory Scale Gains.

		Indiv.	Coll.	NI	DA	RS	RS Gains
Individualism	Correlation	1.000	.346	.561	.055	.236	.613
	Sig. (2-tailed)	.	.448	.190	.907	.610	.144
Collectivism	Correlation	.346	1.000	.095	-.131	-.222	.128
	Sig. (2-tailed)	.448	.	.839	.780	.632	.784
NI	Correlation	.561	.095	1.000	.721	-.314	.481
	Sig. (2-tailed)	.190	.839	.	.067	.492	.274
DA	Correlation	.055	-.131	.721	1.000	-.093	.019
	Sig. (2-tailed)	.907	.780	.067	.	.842	.969
RS	Correlation	.236	-.222	-.314	-.093	1.000	.000
	Sig. (2-tailed)	.610	.632	.492	.842	.	1.000
RS Gains	Correlation	.613	.128	.481	.019	.000	1.000
	Sig. (2-tailed)	.144	.784	.274	.969	1.000	.

Note. n=7. Indiv.= Individualism; Coll.= Collectivism; RS= Regulatory Scale; NI= Non-Interactive; DA= Dynamic Assessment

Although some correlations approached statistical significance, none of them reached an acceptable threshold of <.05. With regard to correlations that neared <.05, the two test formats (NI and DA) were highly correlated. This corresponds to the overall mean scores for the two tests (see Tables 3 & 4 above), which were identical. Some variables correlated negatively with each other, though none of those instances neared significance.

4.1.5 What is the Relation Between Variability in Individualism/Collectivism Scores, and Characteristics of DA Corrective Exchanges, as Realized in Test Data Recordings?

Table 9

Instances of Participant Response types in Dynamic Assessment (DA) Test Corrective Interactions

	HC Group							HI Group						
	P2			P5			Tot.	P3			P4			Tot.
	T1	T2	T3	T1	T2	T3		T1	T2	T3	T1	T2	T3	
Responds as correction	7	8	4	7	5	11	42	10	6	8	7	7	8	45
Attempts self-correction	7	4	4	8	5	7	35	1	5	2	3	4	4	19
Initiates accuracy check	1	4	0	2	2	0	9	3	1	1	4	3	4	16
Attempts correction after minimal prompt	3	5	3	0	1	3	15	0	0	0	0	0	0	0
Ambiguous corrective exchange	1	0	0	0	1	0	2	3	1	1	4	3	0	12
Participant takes initiative	0	0	1	0	0	0	1	2	0	0	4	3	3	12

Note. HC= High Collectivism; HI= High Individualism; P= Participant; Tot.= Total instances for each group

I re-examined the speaking test recordings to assess whether HI and HC I/C differences corresponded to patterns in corrective interactions that took place during DA speaking tests. Table 9 shows instances of response types by participant and according to DA test number. I will compare HI and HC responses for each type.

4.1.5.1 Responds as Correction and Ambiguous Corrective Exchange

"Responds as correction" and "Ambiguous corrective exchange" together accounted for all participant responses inside corrective exchanges. As Table 9 shows, overwhelmingly participants across both groups treated my prompts as corrective, and responded in line with that interpretation. However, even though there were only 14 instances of "Ambiguous corrective exchange," the HI group generated 86% of these. Examples (13) and (14) illustrate this type:

(13) HI corrective exchange (DA test 1)

Participant 4:	They came from uh south part north part south part of China, so they can speak Chinese very well.
Examiner:	Sorry?
Participant 4:	Mm hmm?
Examiner:	They came from south part of China?
Participant 4:	Yeah, south.
Examiner:	You missed an article.
Participant 4:	Oh, the south [part of China.
Examiner:	[Good.

(14) HI corrective exchange (DA test 3)

Participant 3:	But my mom nowadays usually does stock on the computer computer.
Examiner:	Sorry?
Participant 3:	Stock [on the computer.
Examiner:	[She she does stocks?
Participant 3:	Yes.
Examiner:	Can you think of another verb? Not not does.
Participant 3:	Mmm (…) Invest invest.
Examiner:	Yeah, I could say she invests on the computer.

In both cases, the corrective exchange ended with the participants producing correct forms, so at least after the third prompts, P3 and P4 recognized my corrective intention. However, in (13) and (14) it is possible that one or both of my first 2 prompts ("Sorry?", and repeating the erroneous utterance with interrogative intonation) were not understood as corrective, but as clarification requests. This is a plausible interpretation of P4's response, "Yeah, south," in (13). The same is true for P3 in (14), who responded to my second prompt with "Yes." However, in P3's case this is somewhat more surprising, given the fact that this is P3's third DA test, and I had corrected her with identical prompts many times previously. It is noteworthy that example (14) is the only instance of "Ambiguous corrective exchange" for either HI participant from DA test 3. Based on the corrective interaction data, by the third test the HI group (like the HC group) were more

accustomed to my corrective interruptions, and so the exchanges progressed in an expected manner.

4.1.5.2 Attempts Self-Correction and Attempts Correction after Minimal Prompt

For both "Attempts self-correction" and "Attempts correction after minimal prompts," as Table 9 reveals, the HC group evidenced notably more instances than the HI group. The HC group produced 65% of "Attempts self-correction" instances, and 100% of "Attempts correction after minimal prompts." In the latter case, it is important to point out that 73% were produced by P2, so instances were not evenly distributed between the two members of the group.

4.1.5.3 Initiates Accuracy Check and Participant Takes Initiative

Table 9 shows that both groups carried out instances of "Initiates Accuracy Check," though overall the HI group (64% of instances) did so more often than the HC group (36% of instances). With regard to "Participant Takes Initiative," the HI group (92%) accounted for notably more instances than the HC group (8%).

4.1.6 Email Survey Asking for Participants' Perceptions of DA Format Tests

All participants responded to the Email Survey concerning participants' perceptions of DA format tests. Responses for the HI and HC groups are presented in Table 10. Responses for all participants are reproduced in Appendix E.

Table 10

High Individualism (HI) and High Collectivism (HC) Participants' responses to an email survey eliciting perceptions of Dynamic Assessment (DA) tests

	HC Group		HI Group	
	P2	P5	P3	P4
1.Participant primarily focused on (A) accuracy, or (B) task topics.	A	A	B	B
2.Participant perceived examiner expectations to be focused on (A) accuracy, or (B) task topics.	B	B	B	B
3. Due to corrective interruptions, participant perceived examiner expectations to be focused on (A) accuracy or (B) task topics.	A	A	A	B
4. Participant understood examiner's prompt ("Sorry?") to be corrective (A) all of the time, or (B) some of the time.	A	B	A	B
5. Examiner's corrective interruptions made participant feel (A) nervous in the 1st test, but comfortable by test 3; (B) nervous for all 3 tests; (C) not nervous in any test.	A	A	C	C
6. Participant was comfortable by test 3 because (A) s/he was familiar with examiner; (B) s/he was familiar with the test format; (C) s/he was not comfortable by test 3.	A	A	C	A

Note. HC= High Collectivism; HI= High Individualism; P=Participant

Based on the participants' perceptions of the DA tests, the HC and HI groups differed in terms of their own reported orientations towards accuracy or task topics (Question 1). The HC group claimed to have focused more on accuracy, while the HI group claimed to have focused more on the tests' tasks topics. These responses are in line with an argument that the HC group focused more on accuracy than the HI group. Additionally, the HI group's reported focus on task topics rather than accuracy is commiserate with the higher number of "Ambiguous corrective exchange" instances (12 of 14 instances) that they evidenced. Question 2 elicited participants' perceptions of my expectations, as examiner, with regard to accuracy-focused or topic-focused talk during DA format tests. All participants perceived that I expected them to focus on task topics. Specifically, with corrective exchanges (Question 3), however, both HC participants and

one HI participant perceived a shift in my expectations towards accuracy-focused talk. P4, however, perceived my expectations to be remaining focused on task-topic talk, even with corrective exchanges. This response, as well as P4's response to Question 4, in which she reported that she did not always understand "Sorry?" to be corrective, are in line with her evidencing the most "Ambiguous corrective exchanges" (7 of 14 instances; see Table 9). P5 also reported that he did not always understand "Sorry?" to be corrective, though this response is more surprising, given that fact that P5 only evidenced one instance of "Ambiguous corrective exchange," compared to 23 instances of "Responds as Correction." In terms of comfort with corrective interruptions (Question 5), all participants reported that, by DA test 3, they were comfortable with the interruptions, which concurs with my impressions as examiner. However, the HC group reported that they were not comfortable with interruptions in test 1, whereas the HI group reported themselves to be comfortable through all three tests. Finally, three of four participants reported (in Question 6) that their comfort related to familiarity with me as examiner. P3's response was surprising, declaring that she was not comfortable with the DA format, even by test 3. This seems to run counter to P3's response to Question 5, in which she reported that she was comfortable with corrective interruptions throughout the DA tests.

4.2 Discussion of Results

The present study pursued two principal aims. The first was to assess the relation between learner culture and English speaking test performance. Learner culture was operationalized as degree of I/C. This construct was used for a number of reasons. I/C has consistently discriminated between cultures (e.g., Oyserman et al., 2002; Oyserman & Lee, 2008) and across a number of variables with high relevance to second language teaching and learning research, and to speaking tests in particular. Such variables include relationality and communication style (e.g., Oyserman et al., 2002; Gudykunst, 1998). I/C research has largely contrasted Euro-American and East Asian groups, which match the present study's interaction between a Euro-Canadian examiner and Korean participants. Moreover, this study is motivated by previous EAL teaching and learning research that has questioned the efficacy of individualist English teaching practices with East Asian learners (e.g., Spratt et al., 2002; Sullivan, 2008, Wen & Clement, 2003). This previous research suggested the pertinence of I/C as a framework for understanding how cultural differences can manifest themselves in second language teaching and learning contexts. For these reasons, I/C represented both a highly relevant and highly promising construct to apply to English speaking test research.

I evaluated the relation between I/C and participant performance on simulated IELTS™ speaking tests (UCLES, 2005). Additionally, because individualists and collectivists often differ in terms of communication style, I administered the tests with two different formats. The first format (NI) involved very little examiner-participant interaction. The second format (DA) included both encouragement and corrective support from the examiner. Therefore, the second, related aim of the present study was to assess the efficacy of DA as a component of English speaking tests. I will address these aims by discussing the study's research questions.

4.2.1 Is there a difference between participants' NI and DA scores, as measured by the IELTS™ scoring descriptors?

Mean participant scores for the two test formats (NI and DA), as measured by the IELTS™ descriptors, were identical (see Tables 3 and 4). Individually, too, six of seven participants' mean scores were within .2 points of each other. Based on these results, it is not possible to argue for a relation between format and improved speaking test performance, either individually or in terms of the group as a whole.

This study's small sample size (n=7) may account for the lack of notable scoring differences between formats. It is also possible that the DA and NI formats were not sufficiently distinct in order to see differences in scoring. Increasing the amount of encouragement and corrective assistance, which were the features that distinguished the two formats, may have revealed relations between format and individual and/or group scores. However, the results equally suggest the possibility that the two key DA features – encouragement and corrective assistance – do not benefit EAL speaking test-takers in terms of scoring. With regard to encouragement, in the Email Survey, six of seven participants reported that they felt comfortable by DA test 3, but only 3 of those participants attributed that comfort to familiarity with me as the examiner. Since I included encouragement in order to increase participant comfort and confidence (e.g., Feuerstein et al., 1981), as well as to increase participants' sense of familiarity with me as examiner, the email responses do not strongly support the importance of encouragement for test-takers. With regard to corrective assistance, in email responses, four of seven participants reported that the corrective interruptions made them feel nervous for the first DA test, though only one of seven participants felt the same by the last test. It is possible, then, that at least for the first DA test, corrective interruptions counteracted the positive affective and familiarity-causing benefits of encouragement. On the other hand, six of seven participants reported that, by the last test, they felt comfortable and were not nervous when I interrupted with a correction. Overall, Email Survey responses did not offer strong support for correction and encouragement, though after getting used to these interactive features, participants appeared comfortable with them as part of the DA format.

With regard to DA, previous non-EAL studies (e.g., Feuerstein et al., 1981; Campione et al., 1984) have shown positive gains with DA over independent tests, though those studies did not involve EAL groups, and targeted basic cognitive skills. EAL DA studies (e.g., Aljaafreh & Lantolf, 1994; Poehner, 2007), on the other hand, while arguing theoretically for DA's potential to facilitate language development, have not sought to demonstrate DA's superiority, in terms of participant performance, over independent tests. Instead, such studies have shown DA's greater capacity (than independent tests) to discriminate between learners with similar independent test results.

In the present study's case, the DA format's unique features – encouragement and correction – were controlled in order to increase the comparability between the DA and NI formats. Controlling these features also allowed me to assess DA's potential in a standardized format, which represented a novel application of the DA approach. However, standardizing examiner-participant interaction in this way differs from the approach of some DA proponents (e.g., Feuerstein et al., 1981; Aljaafreh & Lantolf, 1994; Poehner, 2007), who claim that individualizing interventions to suit situations and participant needs is paramount in assisting that participant to reach higher developmental levels. Future research may assess whether tailoring correction and encouragement to participants relates to greater gains in DA than NI scores. Ultimately, though, in order reliably to evaluate DA's unique features (and therefore to make systematic improvements to the approach), it is necessary to control the format, to a certain extent, in research designs. For administration and examiner training, too, a standardized version of DA facilitates making DA accessible to a wider range of teachers and examiners than is currently the case.

4.2.2 Is there a difference between participants' NI and DA scores, in terms of gains on successive tests?

This question targeted an important aspect of DA, which is the dynamic component of the approach. According to DA proponents, interaction between examiners and test-takers reveal language areas in need of development. Addressing those needs then creates a new level with new needs, and this cycle continues (e.g., Lantolf & Poehner, 2004; Rieber, 1993). Additionally, DA studies have shown the importance of testing individuals' ability to transfer new learning to subsequent tasks (Campione et al., 1984; Poehner, 2007). Beyond assessment purposes, the DA process is proposed to be facilitative of learners' genetic development (e.g., Aljaafreh & Lantolf, 1994; Rieber & Carton, 1993; Wertsch, 1985). For these reasons the present study included three DA tests, which facilitated learning in line with a genetic model, but also included three NI tests, which increased comparability. Given the limited time-frame for the study (i.e., all tests took place within a four-week period), and the fact that I corrected extensive grammatical and lexical errors as they occurred (rather than specifically targeted), it would not be possible to ascribe score gains to grammatical and/or lexical learning. At the same time, it was important to assess whether three administrations of DA, which is designed to be offered multiple times, related to

higher gains than three administrations of the NI format, which simulates a test (IELTS™) that is usually offered only once.

Overall, participants improved, in terms of their test scores, with both NI and DA test formats. The mean gain for NI tests (+.4) was the same as the mean gain for all (randomized) tests (see Tables 3 and 5). This suggests that test familiarity could just as easily account for the improvements as test format. For the DA style tests, the mean gain (+.6) was slightly higher than the mean gain on NI and all randomized tests (see Table 4). The addition of corrective support in the DA style, which included correction and encouragement, may explain the greater gains in test scores. However, because the mean scores for NI and DA tests were the same, it is equally likely that the gains were due to chance, and/or participants adjusting themselves to a format that included interruptions for corrective support. Explaining DA gains, at least partly, in terms of test familiarity concurs to a certain extent with Email Survey responses, which indicated that four of seven participants felt nervous when interrupted for correction in the first DA test, but by the third DA test only one of seven still felt nervous when corrected.

4.2.3 Is there a difference between participants' DA scores, as measured by the IELTS™ scoring descriptors, and their DA scores measured by the Regulatory Scale?

The RS framework was developed by Aljaafreh & Lantolf (1994) and provided a series of prompts that I used in the present study to assist learners to self-correct. RS scores measured the number of examiner prompts that participants required before they were able to self-correct. Prioritizing self-correction reflected a theoretically-rooted goal of learners achieving independent functioning (e.g., Wertsch, 1985). Based on this theoretical model, lower RS scores pointed to participants' greater capacity to control language forms without examiner assistance.

Overall, participants' DA scores from the IELTS™ descriptors were consistent with their RS scores for the same DA tests (see Tables 4 and 6). In other words, above average (or below average) IELTS™ descriptor scores coincided with above average (or below average) RS scores for the same test, in 71% of cases. This is despite the fact that RS scores focused only on corrective exchanges, and participants' capacities to self-correct, which were taken as indices of lexical and grammatical performance. In contrast, the IELTS™ descriptors (IELTS™, 2009) provided scores across broad criteria. Correspondences between IELTS™ descriptor and RS results additionally matched in terms of *gains*. Four out of five participants who made notable

gains on DA tests measured by the IELTS™ descriptors also made notable RS gains. Moreover, the two participants with negative gains on DA tests measured by the IELTS™ descriptors also evidenced negative gains in RS scores.

In the present study, then, both scoring systems produced commiserate ratings for participants. In terms of this study's assessment of DA's efficacy with formal speaking tests, this concordance between the two rating systems lends some validity to this study's novel use of the RS for corrective assistance and for scoring resultant corrective exchanges.

However, it is important to discuss RS gains in terms of participants' comfort with the DA format. Email Survey responses indicated that four of seven participants were nervous with corrective interruptions for DA test 1, but six of seven participants were comfortable with such interruptions by DA test 3. This suggests that familiarity with the DA format may be an important variable affecting participant performance. Indeed, for certain participants RS scores for the first DA test and possibly the second may have been affected by nervousness. This concern is supported by P7, who was the only participant to report that she was nervous when corrected for all three DA tests, and was also one of the two participants who did not improve in RS scoring. This is not true for another participant (P5), however. P5 reported that he was not nervous by DA test 3, yet his RS scores (like P7) declined from test 1 to 3. Moreover, P5's lack of RS score improvement concurred with a lack of improvement in IELTS™descriptor scores.

Based on participants' relative comfort with corrective interruptions, a more tenable explanation for RS scores' concurrence with IELTS™ descriptor scores is that once participants were comfortable with the DA format, the RS seemed to be able to distinguish between participants in terms of lexical and grammatical performance. This explanation reflects concurrences between the two rating systems, but also reflects the fact that participants needed to get used to DA's novel format. Such an explanation is in line with other EAL DA studies (e.g., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000; Poehner, 2007), which found that DA was able to provide richer assessment information than independent tests, particularly in terms of learners' grammatical understanding. In the present study, too, at least in DA test 3, results point to DA's capacity to reveal differences between participants in terms of lexical and grammatical performance. An important example is DA's capacity to distinguish between slips that often occur in connected speech and genuine errors that reflect a gap in the participant's language, or a

degree of inconsistency in using language forms correctly. The NI format and IELTS™ descriptors do not offer such detailed assessment information. Generally matching RS and IELTS™ descriptor scores suggest that using the combination of accuracy-focused RS scores with a more broadly focused scoring system can provide a more richly detailed picture of a test-taker's speaking ability than the NI test alone. Due to reliability concerns regarding which errors and how many errors to correct, however, DA may not meet strict psychometric standards for major tests such as IELTS™. However, this does not detract from DA's attractiveness to placement, in-class, and exit tests with less stringent psychometric reliability standards.

4.2.4 What is the Relation Between Participants' Culture, as Measured by Degree of I/C, and Their DA and NI Scores?

Mean SCS individualism (4.7) and collectivism (4.8) scores for the seven participants were almost identical (see Table 7). The Korean participants' lack of a clear collectivist orientation generally runs counter to earlier research (e.g., Gudykunst et al., 1996; Kim, 1994; Oyserman et al., 2002), which associated Koreans with high collectivism and/or low individualism. However, Oyserman et al.'s (2002) meta-analysis of I/C research reported studies in which contrary results were also found. Given that Korean groups have not uniformly revealed collectivist orientations, this study's SCS scores are not wholly surprising. A larger sample may have revealed overall clearer tendencies toward either individualism or collectivism.

This study's participant SCS scores may have been affected by validity problems, meaning that the SCS did not reliably tap into basic I/C differences. Somewhat low Cronbach's Alpha scores (indiv.= .63; collect.=.61), which did not reach a generally-accepted threshold of >.70, makes content validity questions a concern. Singelis (personal communication, August 20, 2009) argued, however, that given the breadth of issues that I/C covers, Cronbach's Alpha scores above >.60 should be considered acceptable. Another explanation for this study's I/C inconsistencies could be that the SCS did not tap into the particularities of Korean collectivism (e.g., Triandis & Gelfand, 1998). However, Korean groups have not distinguished themselves in idiosyncratic ways from other collectivist East Asian groups (e.g., Japanese and Chinese) in previous I/C studies (e.g., Gudykunst et al., 1996; Kim, 1994; Oyserman et at., 2002).

The evenly distributed I/C scores in the present study may also reflect the fact that some Korean individuals identify more strongly with individualism than collectivism. Gudykunst and

Lee (2003) argued that self-construal scales, such as the SCS, can reveal the ways that individuals differently mediate the cultures in which they were socialized. In other words, with a large sample, such scales are likely both to affirm underlying cultural frame (i.e., individualist or collectivist predominance) while also pointing to the emphases that individuals place on the varied factors that the scale items targeted. In terms of the present study, while the small sample cannot form the basis of claims regarding larger Korean groups, smaller samples may still be able to illuminate *individual*-level I/C differences. In the present study's case, variability in terms of participants' I/C scores allowed me to create the HI and HC groups, and then to evaluate whether the groups differed in their responses to corrective exchanges.

Correlational analysis relating I/C and speaking test scores revealed no statistically significant correlations (see Table 8). Based on this quantitative analysis, it is not possible to argue for a relation between I/C orientation and speaking test performance, nor between I/C orientation and speaking test format. Focusing on I/C measurement, possible reasons for the lack of correlations include the small sample size (n=7), relatively low internal reliability scores for the SCS, and the compactness of the scale. Previous research reported that internal reliabilities were better for larger scales (e.g., Gudykunst et al., 1996; Triandis & Gelfand, 1998), though the size of those measures made them unsuitable for my small-scale project. Gudykunst et al.'s (1996) scale in particular, because it was designed to tap into I/C communication style differences, may have revealed more detailed links between participants' individualized I/C and aspects of their speaking test performances.

Focusing on speaking tests, it is possible that the feature that I targeted – examiner interactiveness – was not salient enough to tap into participants' I/C differences. In particular it was expected, following I/C and speaking test studies (e.g., Gudykunst, 1998; He & Young, 1998; Kim, 1994; Oyserman et al., 2002), that the NI format, which involved little examiner-participant interaction, would disfavour collectivist participants. This expectation was based on collectivists' other-orientation, and, therefore, the importance of the interlocutor (in this case the speaking test examiner) for guiding communication. While other-orientation in communication appears to be a reliable discriminant between individualists and collectivists, it is possible that the way I operationalized examiner interactiveness in the DA test format (i.e., through encouragement and corrective exchanges) did not sufficiently differentiate the two formats. It

was essential to limit such interaction to allow for comparability with the NI format, in order to assess the effects of that interaction on test performance, and in order to present the DA format and RS scoring system as legitimate additions to formalized speaking test designs. However, freer examiner-participant interaction may have better differentiated the formats for purposes of correlating test and I/C scores. With the NI format, too, the familiarity I established with participants while administering three tests may have diminished the format's negative effects for collectivists.

In addition to I/C-related concerns that the DA and NI formats were not sufficiently dissimilar, it is likely that participants' speaking scores were affected by non-I/C factors, including differences in oral proficiency (despite level being controlled) and affective factors, including motivation and comfort with me as examiner. These mediating variables may have contributed to the lack of significant correlations between I/C level and speaking scores. With this possibility in mind, I also carried out qualitative analysis to see whether I/C level related to differences in interaction patterns between the examiner and participants.

4.2.5 What is the Relation between Variability in I/C scores, and Characteristics of DA Corrective Exchanges, as Realized in Test Data Recordings?

The discussion in Section 4.2.4 pointed to difficulties in quantitatively assessing relations between this study's participants' I/C orientation and their speaking test performances. However, in addition to the correlational analysis, I carried out a qualitative analysis to evaluate whether I/C differences corresponded to communication pattern differences in the speaking tests. This analysis focused largely on HI and HC group corrective exchanges, as these were the primary interactive feature of the DA format tests. Responses to an Email Survey (see Table 10) added support to the qualitative analysis' findings.

The size of the HI and HC groups (i.e., two participants each) prohibits any generalizations beyond the present study's participants. With this limitation in mind, qualitative analysis and Email Survey responses pointed to a notable difference between HC and HI groups in terms of accuracy/task-focus orientation. In their Email Survey responses, the HC participants reported that they had focused primarily on accuracy rather than task topics during DA tests, even though they (as well as the HI group) perceived my expectation to be a focus on task topics. In line with this accuracy focus, the HC group accounted for 65% of "Attempts self-correction"

instances, and 100% of "Attempts correction after minimal prompt" instances, with the latter type pointing to a high sensitivity to apparent errors during DA tests. In contrast, the HI group reported in Email Survey responses that they had focused primarily on task topics during DA tests. This declaration matches with the HI group only accounting for 35% of "Attempts self-correction," and no instances of "Attempts correction after minimal prompt." Not surprisingly, perhaps, the HI group produced 86% of "Ambiguous corrective exchange" instances, in which one or more participant responses in corrective exchanges were not clearly attempts to self-correct. Instead, such responses took the form of defending the correctness of the previous utterance, prompting the examiner for a correct form rather than attempting to self-correct, or offering an affirmative answer to a presumed clarification request. Overall, then, HC communication patterns, as well as their own email responses, evidenced a focus on accuracy over test topics. In contrast the HI group focused to a greater extent on task topics. Whether or not examples (13) and (14) are treated as the HI group interpreting my prompts as clarification requests, the 12 instances of "Ambiguous corrective exchange" in HI group DA tests show that P3 and P4 did not shift their focus from content to accuracy, which the corrective interruptions required them to do, as readily as the HC group. Furthermore, the "Attempts correction after minimal prompts" type points to the HC group's sensitivity (and the HI group's lower sensitivity) to my corrective interruptions. Results from both types support an argument that the HC group was more accuracy-oriented than the HI group.

It could be argued that my corrective interruptions had the effect of causing participants to be more accuracy-oriented. However, only the HI group increased their attempted self-corrections from the first to last DA tests. HC participants self-corrected fewer times in the second (9 instances) and third tests (11 instances) than they did in the first test (15 instances). Therefore the HC group appeared to be oriented towards accuracy regardless of my corrective interruptions. Taking instances of participant-initiated self-correction and sensitivity to my corrective interruptions as indications of accuracy orientation, the HC group certainly appeared to focus on accuracy to a greater extent than the HI group.

Results with "Initiates accuracy check," in which the HI group produced more instances than the HC group (64% vs. 36%), appear to run counter to my claim that the HC group was more accuracy-oriented than the HI group. However, while such requests are clearly accuracy-

focused, they are also to a certain extent status-challenging moves. This is because, like the "Participant takes initiative" type, accuracy checks and requests, by temporarily usurping the right to ask questions, borrow control of test management from the examiner. It is notable that P4, who evidenced the most accuracy checks/requests (7 of 14 instances), also evidenced the most instances of "Participant takes initiative" (10 of 13 instances). In this way, "Initiates accuracy check" represents a somewhat unreliable indicator of accuracy orientation. Similarly, "Participant takes initiative" does not clearly discriminate between HI and HC groups. Although the HI group evidenced far more instances of this interaction type (92% of instances), almost all of these examples of status-challenging moves came from P4.

With regard to I/C, communication style research has not yet established a concern for accuracy as a point of discrimination between individualists and collectivists. On the other hand, in a discussion of Korean learners, Lee (2001) emphasized that a pressure to produce correct speech often leads to a reluctance to speak altogether. Likewise, with regard to Koreans learning English, Han (2005) described a fear of mistakes, which she related to concerns over losing face. Based on these accounts, it seems likely that a concern for producing accurate speech guided the HC group's approach to the DA tests as a whole. Although a concern for accuracy has not been targeted by I/C research, Lee and Han's reports point to cultural sources for this behaviour.

In terms of corrective feedback, too, studies involving East Asian participants (e.g., Lyster & Mori, 2006; Sheen, 2004) support the idea of cultural valuations of correctness. In one study (Lyster & Mori, 2006), a Japanese teacher's strong emphasis on correct production helped to explain learners' high sensitivity to feedback, in terms of uptake and attempted repair. In another study, Sheen (2004) reported high percentages of uptake and repair among Korean learners (compared to Canadian ones), despite the fact that prevailing feedback types did not provide explicit correction. Apparently both the Japanese teacher and Korean learners held relatively high expectations that teacher-learner interaction would focus on accurate forms rather than the content of classroom talk. In the present study, too, my first two prompts ("Sorry?", and a repetition of the erroneous utterance) did not provide explicit correction, but the HC group overwhelmingly responded to the prompts (95% of instances) as corrective. The HI group likewise responded to my prompts as corrective in the majority of cases (79% of instances), but evidenced 86% of "Ambiguous corrective exchange" instances as well. These results add to, and

also complicate Lyster and Mori's (2006) proposal that a classroom's relative communicative or accuracy focus relates to the amount that learners respond to correction and self-correct. Based on the present study's findings, it appears that learners' own orientations either to communication (i.e., task topics) or accuracy also affect the consistency of their responses to correction. Moreover, this responsiveness may relate to culturally-rooted communication style differences.

Email Survey responses did not support an argument that evidence of accuracy orientation (or task topic orientation) related to my expectations as examiner. Uniformly, HI and HC group participants perceived that my expectation was a focus on task topics rather than accuracy. Only in the context of corrective interruptions did participants perceive my expectation to be a focus on accuracy. This suggests that the HC group's accuracy orientation was largely an attribute of the participants themselves, rather than a result of their sensitivity to my correcting and/or evaluative role. This question relates importantly to I/C, because a primary determinant of collectivists' communicative behaviour is their interpersonal context (e.g., Kim, 1994; Kim et al., 2001; Gudykunst et al., 1996; Gudykunst, 1998). In other words, collectivists tend to adjust their communicative behaviour in relation to the role expectations implied by the interlocutor and situation. Taking into consideration Korean cultural expectations of correctness (Lee, 2001; Han, 2005), which was also apparent in Sheen's (2004) corrective feedback study, it seemed likely that the HC group's focus on accuracy reflected their perceptions of examiner expectations. However, Email Survey responses indicated that, aside from corrective interruptions, my presence as examiner was not a primary cause of the HC group's accuracy orientation. Instead this focus appeared to come from the HC participants themselves.

I/C research has shown that individualists tend not to adjust communicative behaviours to the situation or interlocutor to the same extent as collectivists (e.g., Gudykunst, 1998; Kim, 1994; Oyserman et al., 2002). Instead, they value clear self-expression as a communicative goal, and consistency in self-expression as reflective of a stable, underlying individuality. To a certain extent the HI group's responses to correction corresponded to such typically individualist communicative behaviours. The HI group evidenced more instances of "Initiates corrective exchange" (64% of instances) than the HC group, as well as 92% of instances of "Participant takes initiative." HI participants also accounted for 86% of "Ambiguous corrective exchange"

instances, which involved responses other than expected attempts to self-correct. These moves represent status-challenging turns in the sense that they temporarily take control of test management from the examiner, who typically wields an exclusive right to control the sequence of events and ask questions (e.g., Weir, 2005). In the above instances, then, in line with typical individualist behaviour, the HI group did not exclusively adjust their behaviour in deference to my authority as examiner. On the other hand, such status-challenging response types were not distributed evenly between P3 and P4. With regard to "Initiates self-correction," P4 initiated 11 of 16 (69%) of total instances. Likewise, P4 accounted for 10 of 13 (77%) total instances of "Participant takes initiative." P3 evidenced more of such types than the HC group members. However, it is clear that P4 evidenced the majority of status-challenging communicative behaviours. For this reason, further research with a larger sample is necessary to determine whether individualist orientations consistently predict such behaviours.

To a greater extent than the quantitative analysis, this qualitative analysis of corrective exchanges suggested that cultural differences may relate to speaking test behaviours. On the other hand, key I/C correlates relating to communication style, namely collectivists' adjusting communication to contextual (and interlocutor) expectations, and conversely individualists valuing consistent self-expression regardless of interlocutor, did not demonstrably coincide with HI and HC responses in corrective exchanges. Relations between I/C orientation and accuracy focus are not readily interpretable in terms of previously studied I/C correlates. However, non-I/C studies have reported that Koreans typically place a high valuation on accuracy and/or not making mistakes in communication (Han, 2005; Lee, 2001). Given this apparent cultural source for an emphasis on accuracy, it is certainly possible that future I/C studies will reveal correlations between collectivism and accuracy focus. Overall, due to the small sample in question (i.e., only two participants in each group), any claims regarding I/C-related accuracy or task-topic focus are tentative. At the same time, the differences between the two groups provides a stimulus for future research involving I/C and speaking task interaction between examiners (or teachers) and second language learners.

4.3 Limitations of the Study

This section will focus on limitations in connection with the study's primary objectives: assessing (a) I/C's applicability to EAL teaching and learning studies, and (b) DA's efficacy with formal speaking tests.

4.3.1 Individualism and Collectivism Measurement

A primary limitation with regard to I/C measurement was this study's small sample (n=7). Other factors may have contributed to non-correlations between participants' I/C orientation and speaking test format scores, relating to the reliability of the SCS, and the issue of whether the test formats adequately tapped into I/C differences. However, for future research a larger number of participants will be necessary to validate relations that may emerge between I/C and speaking test scores.

Operationalizing learner culture in terms of I/C orientation was motivated by I/C's consistent results in discriminating between cultural groups, and particularly Euro-American and East Asian nationalities, across factors and using different measures (e.g., Hofstede, 1980; Kitayama et al., 2009; Oyserman et al., 2002; Singelis, 1994;Trafimow, Triandis, & Goto, 1991; Triandis & Gelfand, 1998). This study involved similar cultural groups (i.e., a Euro-Canadian examiner and Korean test-takers). A second reason for using I/C was its suitability to second language teaching and learning studies, in light of a body of research that has questioned individualist teaching practices in East Asian classrooms (e.g., Spratt, Humphreys, & Chan, 2002; Sullivan, 2008; Wen & Clement, 2003). However, other factors that Hofstede (1980) isolated as major differentiators between cultures may complicate this study's culture dimension. Future research relating culture and speaking tests needs to account for Power Distance, which describes a culture's relative acceptance of hierarchical power structures. Given the status differences implicit in examiner and test-taker interaction, this factor may also have had an effect on speaking test scores. Similarly, Masculinity, which describes the degree to which a culture clearly demarcates male and female roles, may also complicate cultural measurement in terms of speaking tests. Considering that six of seven participants in this study were female, there is equally reason to think that this factor may have mediated I/C's effects on speaking test performance.

The present study used Singelis' (1994) SCS, a 30-item Likert-type questionnaire, to measure participants' degree of I/C. This scale had previously evidenced low internal reliability scores (e.g., Oyserman et al., 2002). However, according to Oyserman et al.'s (2002) meta-analysis of I/C research, the SCS was not alone in this regard. No other widely-used compact I/C instrument had evidenced consistently high internal reliabilities. Other I/C scales with multiple dimensions and large numbers of items have evidenced improved internal reliabilities (e.g., Gudykunst et al., 1996; Triandis & Gelfand, 1998). Although I determined that using such large scales was not feasible for this small-scale project, there are legitimate concerns that the low number of items (n=30) in the SCS may not reliably tap into I/C differences. After piloting the SCS for this study, Cronbach's Alpha scores did not reach a generally acceptable threshold of >.70 for either the individualism or collectivism dimension. After removing two negatively-correlated items from the collectivist half of the instrument, reliabilities for both dimensions reached >.60, which Singelis (Personal communication, August 20, 2009) argued was acceptable given the breadth of issues covered by the scale. Additionally, Gudykunst and Lee (2003) argued that internal reliability measurement for self-construal scales did not account for an individual-culture interface in which individuals with similar underlying cultural frames emphasize different correlates of the shared culture. For these reasons I deemed that the SCS was acceptable for use with this study.

More recent approaches to I/C measurement, including priming (e.g., Oyserman & Lee, 2008) and direct measurement (e.g., Kitayama et al., 2009), offer more consistently reliable I/C measures. On the other hand, these approaches involved priming only one aspect of I/C in order to measure its effects on subsequent narrowly-focused tasks. As this study was focused on speaking tests, which apparently related to a variety of I/C correlates, including self-concept, relationality, and communication style, such a narrowly-focused I/C measure was not desirable. Despite a thorough review of previously-used scales, it was difficult to find an I/C instrument that measured general I/C, was compact and had shown consistent internal reliabilities.

Qualitative analysis found notable differences between HI and HC groups' interaction patterns in the DA tests, in terms of participants' relative focus on accuracy or task topics. I/C communication style studies have not explicitly focused on accuracy-orientation (e.g., Gudykunst, 1998). Although results may be explainable in terms of generalized individualist

valuations of self-expression, and collectivists' valuation of adjusting communication to roles and situations, it is not possible at this point to claim relations between SCS I/C scores and the HI and HC groups' communicative behaviours. Given the small sample constituting the HI and HC groups, these results cannot be held as representative of the communicative behaviours of larger I/C-differentiated populations. The results can be a source of further inquiry into I/C and EAL task performance connection. Yet, in order to move beyond I/C generalizations and make stronger claims for links between I/C and EAL task performance, it is clear that a more specific focus will be necessary, both in terms of I/C measurement and in terms of particular features of second language learning tasks.

4.3.2 Using DA with Formal Speaking Tests

The present study evaluated a novel application of DA, which retained DA's unique qualities while partly standardizing the format. Such unique qualitites include examiners intervening to assist test-takers with language that they cannot yet control independently, and in so doing eliciting rich assessment information about the test-takers (e.g., Aljaafreh & Lantolf, 1994; Lantolf & Poehner, 2004). To partly standardize the format, I employed the RS both as a framework for correcting participants and also as a means of scoring corrective exchanges. Participants' scores amounted to the number of corrective prompts they required to produce a correct form, with each successive RS prompt increasing in explicitness.

In practice, it was not possible to standardize all aspects of administering the DA format tests. Previous DA studies (e.g., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000; Poehner, 2007) employed DA in interactions that focused explicitly on assisting learners in correcting mistakes. While administering the speaking tests, however, a primary consideration was minimizing imposition from corrective interruptions. This entailed providing corrections in a manner that did not detract from the participants' test performances. As a result, it was not possible to correct all errors that emerged, as this approach may have interrupted participants' continuity while responding to test tasks, and may also have negatively affected participants' confidence. Because of these considerations, I was not able to control such variables as how many errors I corrected in each test. This raises concerns about the consistency of the DA format administrations, and by extension the reliability of resultant scores. On the other hand, in practice, it was not difficult to determine minimally disruptive times to intervene with a correction.

Moreover, the Email Survey responses suggest that, once participants became accustomed to the corrective interruptions, they did not have adverse effects in terms of nervousness. Three of seven participants reported that the corrective interruptions did not make them feel nervous for any of the three DA tests, while six of seven participants reported that, by the third test, the interruptions no longer made them feel nervous.

A second concern with the DA format administration was that I was not able to provide corrections (as I had originally intended) across the four IELTS™ descriptor criteria (IELTS™, 2009), which would have maximized comparability between the NI and DA format tests. I determined that it was not possible consistently to correct discourse-level errors and pronunciation errors while following the RS, and while prioritizing keeping corrective interruptions minimally disruptive. As a result, my corrections were limited to extensive grammatical and lexical errors as they arose. This approach raised several concerns. Firstly, it was not possible consistently to select which errors to correct, in terms of degree of difficulty and range, while simultaneously administering the DA tests. For this reason, there is a possibility that RS scores were affected by the difficulty of the errors that I corrected. To partly compensate for this variable, RS scores for each test were determined by averaging the scores from each test's corrective exchanges. This system limited the effect that range of error difficulty had on a test's RS score. Moreover, the fact that participants' positive and negative RS gains (see Table 6) were in line with IELTS™ descriptor gains (see Table 4) lends some validity to RS correcting and scoring methods that this study used.

Secondly, the DA format's relatively narrow corrective focus on grammatical and lexical errors limits its compatibility with general oral proficiency tests such as IELTS™. Using DA during test interaction meant that it was not feasible to correct pronunciation and discursive errors. To a certain extent this reduces the format's attractiveness as a complement to speaking tests that include pronunciation and discursive skills in their scoring criteria.

An additional limitation with regard to multiple DA tests in the present study is the difficulty in demonstrating relations between DA scoring gains and participant learning. For one thing, the limited duration of the study (i.e., four weeks) raises questions about the possibility of participants being able to consolidate learning from the corrective exchanges that they engaged in. Furthermore, in previous DA studies (e.g., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000;

Poehner, 2007), corrective assistance focused on limited types of errors. Such a limited focus increases the validity of such studies' suggestions that participants' improved use of targeted language forms constituted "microgenetic" development (Aljaafreh & Lantolf, 1994). It is noteworthy, however, that the above studies lasted four to six weeks, raising similar questions about the ability to claim that learning has taken place. With regard to the present study, it is not possible to make claims that the present study's extensive correction of grammatical and lexical errors as they arose related to participants' gains in DA scores. This does not discount the value of such corrrection as a means of achieving a DA goal of assisting learners to reach higher developmental levels (e.g., Lantolf & Poehner, 2004). Indeed, extensive correction of grammatical and lexical errors arguably does a better job of individualizing corrective assistance (e.g., Aljaafreh & Lantolf, 1994; Poehner, 2007) than target-focused assistance. More extensive assistance represents a more emic DA approach, since the corrected errors are not pre-determined, but emerge from the participants' own talk. However, such participant-focused correction approach comes at the cost of substantiating links between examiner intervention and learning outcomes.

Additionally, it is arguable that the RS scoring system is biased in favour of accuracy-focused test-takers. Furthermore, the RS prompts that I used ((a) "Sorry?"; (b) a repetition of the erroneous utterance; and (c) identifying the particular type of error) contain a degree of ambiguity that may have led some participants to interpret them as other than corrective (e.g., Lyster & Mori, 2006; Sheen, 2004). This particularly applies to P3 and P4 (the HI group), who accounted for 12 of 14 "Ambiguous corrective exchange" instances in those first two DA tests, which appear to relate to the participants' topic-focused orientation. The HI group's Email Survey responses were commiserate with this interpretation, as they reported that they had focused on topics rather than accuracy. These concerns raise doubts about the HI group's RS scores, and particularly their below-average RS scores for the first two DA tests. Conversely, the HC group (P2 and P5) accounted for 65% of "Attempts self-correction" instances, which is a move that is likely advantageous in terms of RS scores. This is because participants received a 0 (i.e., the best possible score) if they were able to self-correct without the examiner intervening.

On the other hand, "Attempts self-correction" included instances where self-correction was unnecessary and/or unsuccessful, so that in practice this move may not have resulted in

superior RS scores for the HC group. Indeed, the HC group, who were the only two participants to report in Email Survey responses that they focused primarily on accuracy, did not apparently benefit from this orientation in terms of RS scores. Although P2's mean RS score (1.7) was better than the group mean (1.9), the same was not the case for P5's mean score (2.3). Moreover, in two of three cases where P2 and P5's RS scores were not in line with their IELTS™ descriptor scores, this was because the RS scores were *below* average. For the HI group, too, in two of the four tests in which their RS scores were below average, their IELTS™ descriptor scores, with its broader assessment criteria, were also below average. Overall, in 15 of 21 tests, participants' RS scores were in line with their IELTS™ descriptor scores. Additionally, while the degree of ambiguity in the examiner's corrective prompts possibly confused the HI group more than the HC group, by and large both groups responded to the prompts in line with their corrective intention. For the HC group, "Responds as Correction" accounted for 42 of 44 corrective exchanges, and 45 of 57 corrective exchanges for the HI group. For this reason, there is little evidence to support an argument that the RS is biased in favour of accuracy-focused participants. Instead it appears that the RS did its job of discriminating between participants in terms of grammatical and lexical performance.

4.4 Implications and Directions for Future Research

This section discusses the implications of the present study's findings in terms of the two major research objectives: (1) evaluating the relation between learner culture (degree of I/C) and English speaking test performance, and (2) evaluating the efficacy of Dynamic Assessment (DA) for use with oral proficiency tests.

4.4.1 Individualism/Collectivism and Speaking Test Performance

One guiding question in the present study was whether learner culture, operationalized as degree of I/C, could illuminate learner differences in EAL speaking tasks. A review of the literature reveals I/C's promise as a means of understanding performance-related teacher-student and student-student cultural differences in international language classrooms. Particularly in terms of Euro-American and East Asian comparisons (also the focus of the present study), I/C has reliably differentiated between groups in terms of relationality, attribution style, values, self-concepts and communication styles (e.g., Gudykunst et al., 1996; Kim, 1994; Kim et al., 2001; Kitayama et al., 2009; Oyserman et al., 2002). In particular, communication style I/C differences (e.g., Gudykunst et al., 1996; Kim, 1994; Kim et al., 2001), have promising applicability for second language learning and teaching research, including speaking task performance and interaction between peers, and between teachers and learners.

The present study sought to evaluate I/C's applicability in the domain of speaking tests. Results offer mixed support in terms of indicating relations between I/C differences and differences in EAL task performance. The most striking findings came from qualitative analysis, which focused on corrective exchanges and revealed notable differences between the HI and HC groups. Keeping in mind validity concerns relating to the small size of these groups, a review of the HI and HC groups' corrective exchanges with the examiner found that the HC group showed a relatively high orientation towards accuracy, while the HI group oriented towards task topics to a greater extent. Participants' Email Survey responses supported these findings. Models of individualist communication styles (e.g., Gudykunst et al., 1996; Kim, 1994; Kim et al., 2001) may help to explain the HI group's task focus and (particularly for P4) their relatively high number of status-challenging moves in terms of a primary communicative goal of self-expression, over and above deference to the examiner's authority. In contrast, a collectivist model of communication style may help to explain the HC group's accuracy focus in terms of a primary

other-focus, and therefore a goal of adjusting communication to meet interlocutor expectations. In the latter case, however, the HC group's Email Survey responses attributed this accuracy focus to themselves and not my expectations as examiner. Yet the HC group's orientation towards accuracy is in line with accounts of Korean educational practices, which traditionally place a high emphasis on accuracy and/or avoiding mistakes (e.g., Han, 2005; Lee, 2001). This apparent cultural source for focusing on accuracy and/or avoiding mistakes creates a new hypothesis for correlating I/C and EAL task performance. Testing such a hypothesis represents one area where future research can build upon this study's findings.

In terms of I/C measurement, it is clear that small-scale research using I/C requires a instrument that is able more directly and reliably to tap into individual-level culture. One reason for this is the frequently low internal reliabilities of compact questionnaires such as the SCS (Oyserman et al., 2002; Oyserman & Lee, 2008). In addition, the questionnaire format "assumes that cultural frame is a form of declarative knowledge" (Oyserman et al., 2002, p.7). Unfortunately, this assumption creates a problematic causal gap in research that seeks to demonstrate relations between reported ideas and feelings and actual communicative behaviours. Kitayama et al.'s (2009) implicit I/C measurement represents a means of shrinking this gap in causality. This measurement is similar to previous I/C priming studies (e.g., Oyserman & Lee, 2008; Trafinow et al., 1991; Utz, 2004), in that it is designed to elicit either individualism or collectivism. Unlike other priming research, however, primed participants in Kitayama et al.'s (2009) study are not then compared on subsequent tasks to measure correlations. Instead the priming and task are combined into one component, based on a premise that individuals express I/C orientation through important cultural tasks. The job of researchers, then, is to design direct (implicit) tasks that are analogous to cultural tasks through which individuals express cultural identity. Highly systematic results in Kitayama et al.'s study (2009) strongly endorsed both the value of I/C as a means of differentiating cultural groups and the implicit priming approach to I/C measurement. In terms of second language teaching and learning research, it certainly seems possible to design implicit tasks that directly elicit basic I/C differences, yet exist independently as tasks or situations common to EAL teaching and learning. An example is a dilemma task (e.g., Utz, 2004) in which participants have the opportunity to express individual uniqueness or similarity to a group, but not both. Another example is a task differentiating consistent self-

expression across interlocutors with communication that adjusts to different interlocutors. In such a task participants would be asked to express themselves with older peers (i.e., they have lower status), with younger peers (i.e., they have higher status), and with an instructor (i.e., they have lower status). Such tasks tap into central I/C communication style differences, while removing many of the mediating variables that may have affected results in the present study. Implicitly relating EAL behaviours and I/C in this way strengthens claims for causality between I/C and actual behaviours. This in turn would strengthen arguments for I/C's implications in everyday language learning and teaching communication, including speaking tests. Designing and piloting a series of such tasks was beyond the scope of this study, but represent a promising line of future research.

In terms of speaking tests, learner culture represents one of many learner variables that may affect test performance (e.g., Fulcher & Davidson, 2007; McNamara, 1997; Weir, 2005). However, speaking test research has not previously operationalized learner culture to assess its relation with test performance. Keeping in mind the need for more reliable I/C measurement techniques, as well as the small sample size, the present study's quantitative analysis found no significant correlations between I/C and speaking test performance. Specifically, test format did not correlate with higher scores for individualist over collectivist participants. Although further research is needed before firm conclusions can be drawn, the NI test, which simulated the IELTS™ speaking test format, did not appear to disfavour collectivists in line with I/C-based expectations. These results represent a small contribution to the "multifaceted" approach to test validation that Weir (2005, p. 13) called for. Such a mode of test development demands that test design, examiner training and evaluation, rating design and rater training and evaluations, among other test components, must all be the scrutinized to ensure optimal test reliability.

4.4.2 Implications for Corrective Feedback

CF research involving East Asian (i.e., Japanese & Korean) and Euro-American and Euro-New Zealand interlocutors (e.g., Lyster & Mori, 2006; Sheen, 2004) has suggested possible connections between learner/teacher culture and orientations towards CF, without pursuing those potential links. Results from the present study were in line with Sheen's (2004) arguments that the conversational implication of feedback types affects resultant student repair. In this study, the examiner's prompts contained a degree of ambiguity and may have been interpreted as

confirmation checks rather than cues for self-correction. However, such misinterpretations (i.e., "Ambiguous corrective exchanges"; see Table 9) were clearly divided along I/C lines. This suggests that not only the feedback type's conversational implication, but also the learner's general orientation towards accuracy (or task topic) may affect the degree that CF is noticed and leads to self-correction. Moreover, the present study's results suggest that such learner orientations towards accuracy or task focus may be culturally based. It should be stressed that I/C's relation to relative accuracy focus is not strongly substantiated, particularly given the small HI and HC group size (n=2), and needs to be tested in future research. The same is true for relations between participant accuracy focus and the degree that certain types of CF are misunderstood. For EAL teachers, however, this study's results tentatively provide a rationale for teachers to select CF types based on their awareness of learners' relative accuracy orientation. This is because the HC group appeared to have little difficulty interpreting my prompts as corrective, while the HI group appeared to have greater difficulty in this regard. More generally, EAL teachers need to be aware that a variety of factors, including learner orientations towards accuracy, are likely to affect the success of proffered CF.

4.4.3 Dynamic Assessment in Speaking Tests

The second major research objective in the present study was to assess DA's suitability to formal second language speaking tests. This was a novel application for DA, which researchers previously had assessed with written tutorials (i.e., Aljaafreh & Lantolf, 1994; Nassaji & Swain, 2000), and story reconstruction tasks (i.e., Poehner, 2007). To allow for reliable administration and scoring, this study partly controlled DA's unique features, including corrective support and encouragement for participants. This study also employed multiple tasks over time. Offering multiple tasks reflects DA's emphasis on examiners assisting learners to reach higher developmental levels, as well as the need for examiners to assess whether improvements have taken place. However, to ensure comparability and to account for practice effects, both DA and NI format tests were given an equal number of times. Controlling the DA format differs from many DA proponents' approaches (e.g., Aljaafreh & Lantolf, 1994; Feuerstein et al., 1981; Lidz, 2000; Poehner, 2007). Such researchers argue that the goal of assisting learner development is the examiner's primary mandate, and necessitates adjusting assistance to learners and

circumstances. However, for DA to become available for wider use by teachers and examiners, and to allow for evaluating DA's unique features, a degree of standardization is necessary.

There were no notable differences between NI and DA mean scores and gains, which may lead critics of standardizing DA to argue that more individualized corrective support would have led to greater DA gains over NI tests. It is not possible to dispel this concern, though future research may compare this study's controlled DA format with an individualized format. Results did, however, generally support DA's power to discriminate between test-takers in terms of grammatical and lexical performance. In this study, once participants were accustomed to corrective interruptions, corrective support (in conjunction with RS scoring) provided relatively rich assessment information. In Sociocultural terms, corrective exchanges may be an indice of test-takers' proximity to independently controlling the targeted language forms (e.g., Poehner, 2008). One specific advantage of DA is that it has the capacity to differentiate grammatical or lexical slips from genuine gaps in participants' second language, which is a distinction that the NI format (in conjunction with the IELTS™ descriptors) cannot make.

DA's comparable mean scores (to the NI format) indicate that corrective interruptions, at least by DA test 3, did not adversely affect participants' test performances. Impressionistically, as examiner, I noticed an increased comfort with the format by DA test 3, as participants were able to shift smoothly between talk about task topics and corrective exchanges. Commiserate RS and IELTS™ descriptor gains, both positive and negative, lend some support both to using the RS for corrective exchanges, and to the reliability of RS scoring. Given these results, and the unique information that DA can provide, the format showed its potential to be an effective component of a formal speaking test. This finding differs from previous DA research (e.g., Aljaafreh & Lantolf, 1994; Poehner, 2007), which has tended to present DA's features as incompatible with independent testing. At the same time, DA studies have been limited by a narrow focus on discrete grammatical points. This study's results, however, pointed both to the compatibility of DA's interactive features with formal tests' formats, and also to the added discriminatory power that DA is able to bring to such tests.

At the same time, certain factors detract from DA's attractiveness with formal speaking tests. For one thing, results support a concern that test-taker familiarity with the DA format is necessary before it can be an effective assessment approach. Since major English speaking tests

such as IELTS™ tend to be one-time interviews, the nervousness caused by corrective interruptions from stranger examiners may limit test-takers' capacities to produce their best performances. Furthermore, while this study's proposed DA format was partly controlled, in terms of using the RS and offering encouragement, it was not possible to make these interactive elements completely uniform. In particular, the question of which (and how many) errors to correct was not controlled, in order to minimize corrective interruptions' effects on participants' confidence and continuity while answering task questions. In this study, findings did not point to this aspect of the DA format adversely affecting RS scoring reliability. On the other hand, the degree of examiner interpretation that it introduces to the test may deter major test makers, who are wary of any threats to reliability that such ambiguities create. For these reasons, the DA format is unlikely to be adopted as a component of major speaking tests such as IELTS™.

For examiners and test-makers who are not bound by strict psychometric standards of reliability, however, the DA format has value as an additional component to existing second language speaking tests. My impressions as examiner, Email Survey reports, as well as parallel gains in IELTS™ descriptor and RS gains, point to DA's potential in providing detailed grammatical and lexical assessment information, while not detracting from a test-taker's overall test performance, and therefore from the reliability of the test as a whole. The standardized format introduced in this study represents a feasible means of bringing DA's unique features to wider use for language testing. Employing the RS for corrective support and for scoring corrective exchanges allows for relatively straightforward examiner and rater training. In particular, DA lends itself to ongoing within-course assessments and exit testing in which teachers have established familiarity with learners, and learners have become familiar with the format. The importance of familiarization does not distinguish DA from other speaking tests, including IELTS™, whose idiosyncratic formats may confuse test-takers who have not encountered them previously. Nonetheless, in-course and exit tests remain an optimal fit for DA, for the additional reason that ongoing assessments can take advantage of DA's equal emphasis on assisting learners in developing language skills. Corrective exchanges can reveal language areas that require assistance, both for individuals and for classes as a whole. These areas can then be integrated into future instruction or curriculum development. The DA format, through its series of contingent prompts, as well as its emphasis on multiple tests, has the potential to

provide teachers with information on the degree to which their students have assimilated newly taught grammatical and lexical items into their second language, and are able consistently to employ those items across novel tasks. Teacher/examiners can both offer corrective assistance within DA tests, and also make note of common errors for future whole-class or individualized instruction. In this way, integrating DA into ongoing assessments responds to calls for a wider usage of testing for formative, rather than narrowly summative assessment purposes (e.g., Sternberg & Grigorenko, 2002; Weir, 2005), while not compromising the reliability of the speaking test as a whole.

CHAPTER 5: CONCLUSION

This thesis explored the relations between learner culture, operationalized as degree of individualism/collectivism (I/C), and speaking test performance. The study was motivated by a growing awareness that language and culture are deeply intermingled (e.g., Lantolf, 2006; Magnan, 2008; Savignon & Sysoyev, 2002), as well as by studies that have questioned the appropriacy of individualist English teaching practices in non-Western – and particularly East Asian – classrooms (e.g., Spratt, Humphreys & Chan, 2002; Sullivan, 2008; Wen & Clement, 2003). In light of I/C's consistent power to discriminate between Euro-American and East Asian groups (e.g., Gudykunst et al., 1996; Kitayama et al., 2009; Oyserman et al., 2002), this study enlisted the construct to assess whether learner culture related to performance differences in language tasks. This project represented an evaluation of I/C's potential to illuminate tangible ways that culture may affect second language learning and teaching practices. In particular, this study focused on speaking tests, because they are critical encounters for second language learners, and because there was reason to believe that the International English Language Testing System (IELTS™) speaking test format disfavoured collectivist test-takers. As a result, this study also evaluated a potentially more culturally equitable speaking test, Dynamic Assessment (DA) (e.g., Lantolf & Poehner, 2004), in terms of its applicability to formal speaking tests.

Participants responded to a compact I/C questionnaire, and completed six speaking tests. The tests were divided into two formats, DA and Non-Interactive (NI) (simulating the IELTS™ speaking test format), which differed in terms of the amount and type of examiner interaction with participants. Results found no difference between formats in terms of participants' mean IELTS™ descriptor scores, though DA tests accounted for a slight edge in gains (over NI tests) from test 1 to test 3. For this reason, it was not possible to claim that DA's additional features – corrective support and encouragement – notably benefited participants in test performance. However, results support DA proponents' claims (e.g., Aljaafreh & Lantolf, 1994; Poehner, 2007) that, through corrective assistance, the format is able to provide richer assessment information than non-interactive formats. Results also supported DA's capacity to differentiate between test-takers in terms of grammatical and lexical performance. As this study represents a first attempt to evaluate a standardized version of DA for speaking tests, more research is needed before generalizations can be made. Nonetheless, given Regulatory Scale (RS) scores' compatibility

with IELTS™ descriptor scores in the present study, a standardized version of DA shows promise as a component of formal English tests. The best fit for DA appears to be ongoing in-class or exit tests which will allow both teachers and students to familiarize themselves with the format.

I/C Self-construal Scale (SCS) results evenly divided the study's Korean participants in terms of individualism and collectivism. This result runs counter to Korean groups' typically collectivist orientations (e.g., Kim, 1994; Oyserman et al., 2002; Triandis & Gelfand, 1998), though contrary findings have also been reported (e.g., Oyserman et al., 2002). I/C and speaking test scores revealed no significant correlations, which contradicts concerns that the NI format would disfavour collectivist participants. On the other hand, qualitative analysis comparing High Individualism (HI) and High Collectivism (HC) groups revealed notable differences between participants in DA tests' corrective interactions. Specifically, the HC participants evidenced greater emphasis on accuracy than HI participants, while the HI group apparently focused primarily on test task topics. Participants' Email Survey responses served as confirmation for this group difference. Although the small group sizes prohibit making strong claims in this regard, the finding seems to relate to a general valuation of accuracy (and an avoidance of making mistakes) in Korean education (e.g., Han, 2005; Lee, 2002; Song, 1994). However, I/C studies have not previously assessed accuracy as a point of discrimination between individualists and collectivists, and HI and HC group differences cannot reliably be explained in terms of I/C communication style models. As such, the present study's results generate an accuracy hypothesis that future I/C research may be able to examine.

In spite of mixed results in relating I/C and speaking test performance, the I/C construct remains a highly promising framework for illuminating teacher-learner and learner-learner differences in EAL contexts. Specifically, communication style differences (e.g., Gudykunst, 1998) have high relevance to EAL interaction discussions. On the other hand, reliability concerns about compact self-construal scales mean that future research in this area needs to find a more effective I/C instrument. One measurement approach that has shown high reliabilities involves implicit I/C tasks (e.g., Kitayama et al., 2009), which tap into the ways that individuals express underlying cultural orientations. It seems feasible to develop a series of I/C tasks that target established communication style I/C differences, yet also directly relate to EAL teaching

and learning situations. Combining I/C and EAL task measures in this way has the potential to illuminate relations between individual culture and EAL learning and teaching practices.

REFERENCES

Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In J.P. Lantolf & M.E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 57-86). London: Equinox.

Aljaafreh, A., & Lantolf, J.P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal, 78*, 465-483.

Anton, M. (2009). Dynamic assessment of advanced second language learners. *Foreign Language Annals, 42*, 576-595.

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

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 1,* 1-25.

Brown, A. (2009). Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *Modern Language Journal, 93,* 46-60.

Brown, A., & Ferrara, R.A. (1985). Diagnosing zones of proximal development. In J. Wertsch (Ed.), *Culture, communication and cognition: Vygotskian perspectives* (pp. 273-305). Cambridge: Cambridge University Press.

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

Campione, J.C., Brown, A.L., Ferrara, R.A. & Bryant, N.R. (1984). The zone of proximal development: Implications for individual differences and learning. In B. Rogoff & J.V. Wertsch (Eds.), *Children's learning in the 'zone of proximal development'* (pp. 77-91). San Francisco: Jossey-Bass.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. & Barkhuizen, G. (2005). *Analyzing learner language.* Oxford: Oxford University Press.

Feuerstein, R., Miller, R., Rand, Y., & Jensen, M.R. (1981). Can evolving techniques better measure cognitive change? *Journal of Special Education, 15*, 201-219.

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

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment: An advance resource book*. London: Routledge.

Galaczi, E.D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*, 89-119.

Grice, H.P. (1975). Conversational implicature and metaphor: Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts. Volume 3* (pp. 41-58). New York: Academic Press.

Gudykunst, W.B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K. & Heyman, S. (1996). The influence of cultural individualism-collectivism self-construals, and individual values on communication styles across cultures. *Human Communication Research, 22*, 510-543.

Gudykunst, W.B. (1998). Individualistic and collectivistic perspectives on communication: an introduction. *International Journal of Intercultural Relations*, *22*, 107-34.

Gudykunst, W.B., & Lee, C.M. (2003). Assessing the validity of SCSs: A response to Levine et al. *Human Communication Research, 29*, 253-274.

Guthke J., & Beckmann, J.F. (2000). "The Learning Test Concept and its Application in Practice." In C.S. Lidz & J. Elliott (Eds.), *Dynamic Assessment: Prevailing models and applications* (pp. 17-70). New York: Elsevier.

Hall, E. (1976). *Beyond Culture*. New York: Anchor.

Han, S-A. (2005). Good teachers know where to scratch when learners feel itchy: Korean learners' views of native-speaking teachers of English. *Australian Journal of Education, 49,* 197-213. Retrieved from: http://findarticles.com/p/articles/mi_hb6475/is_2_49/ai_n29206625/

He, A.W., & Young, R. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam: J. Benjamins.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills: Sage.

Holtgraves, T., & Yang, J.-N. (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of Personality and Social Psychology, 62*, 246-256.

Hughes, R. (2002). *Teaching and researching speaking*. London: Longman.

Hui, C.H., & Yee, C. (1994). The shortened individualism-collectivism scale: Its relationship to demographic and work-related variables. *Journal of Research in Psychology, 28,* 409-424.

Hymes, D. (2001). On communicative competence. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (pp. 53-73). Malden, MA: Blackwell.

IELTS™. (2009). IELTS speaking band descriptors (public version). Retrieved from: http//www.IELTS.org/researchers.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction, 28,* 171-183.

Kim, M.-S. (1994). Cross-cultural comparisons of the perceived importance of conversational constraints. *Human Communication Research, 21*, 128-151.

Kim, M.-S, Aune, K.-S, Hunter, J.E., Kim, H.-J, & Kim, J.-S. (2001). The effect of culture and self construals on predispositions toward verbal communication. *Human Communication Research, 27,* 382-408.

Kim, K., & Suh, K. (1998). Confirmation sequences as interactional resources in Korean language proficiency interviews. In R. Young & A.W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency*. (pp. 297-332). Amsterdam: J. Benjamins.

Kitayama, S., Duffy, S., & Uchida, Y.K. (2007). Self as cultural mode of being. In S.Kitayama & D. Cohen (Eds.), *The handbook of cultural psychology*. (pp. 136 174). N.Y.: Guilford Press.

Kitayama, S., Park, H., Sevincer, A.T., Karasawa, M., & Uskul, A.K. (2009). A cultural task analysis of implicit independence: Comparing North America, Western Europe, and East Asia. *Journal of Personality and Social Psychology, 97,* 236-255.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal, 70*, 366-372.

Kuhn, M.H., & McPartland, T.S. (1954). An empirical investigation of self attitudes. *American Sociological Review, 19*, 68-76.

Kumaravadivelu, B. (2003). Problematizing cultural stereotypes in TESOL. *TESOL Quarterly, 37*, 709-719.

Lantolf, J.P., & Poehner, M.E. (2004). Dynamic assessment: Bridging the past into the future. *Journal of Applied Linguistics, 1*, 49-74.

Lantolf, J.P., & Thorne, S.L. (2006). *Sociocultural theory and the genesis of second language development.* Oxford: Oxford University Press.

Lantolf, J.P. (2006). Re(de)fining language proficiency in light of the concept of 'languaculture.' In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 72-94). London: Continuum.

Lantolf, J.P., & Poehner, M.E. (2008). Introduction to sociocultural theory and the teaching of second languages. In J.P. Lantolf & M.E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 1-33). London: Equinox

Lazaraton, A. (1992). The structural organization of a language interview: A conversation analytic perspective. *System, 20,* 373-386.

Lee, J.-A. (2001). Korean speakers. In M. Swan & B. Smith (Eds.), *Learning English*. (pp. 325-342). Cambridge: Cambridge University Press.

Lidz, D. (2000). The application of cognitive functions scale (ACFS): An example of curriculum-based dynamic assessment. In D. Lidz & J. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 407-440). N.Y.: Elsevier.

Long, M., Inagaki, S., & Ortega, L. (1998). The role of implicit negative feedback in SLA: Models and recasts in Japanese and Spanish. *Modern Language Journal*, *82*, 357-371.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition, 19,* 37-66.

Lyster, R. (1998). Recasts, repetitions and ambiguitiy in L2 classroom discourse. *Studies in Second Language Acquisition, 20,* 37-66.

Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition, 38,* 269-300.

Magnan, S.S. (2008). The unfulfilled promise of teaching for communicative competence: Insights from sociocultural theory. In J.P. Lantolf, & M.E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 349-379). London: Equinox.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

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

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

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

Nabei, T., & Swain, M. (2002). Learner awareness of recasts in classroom interaction: A case study of an adult EFL student's second language learning. *Language Awareness, 11,* 43-63.

Nassaji, H., & Swain, M. (2000). Vygotskian perspective on corrective feedback in L2: The effect of random versus negotiated help on the learning of English articles. *Language Awareness, 9,* 34-51.

Nassaji, H. (2007). Elicitation and Reformulation and their relationship with learner repair in dyadic interaction. *Language learning*, *57,* 511-548.

O'Sullivan, B., Weir, C.J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19,* 33-56.

Oyserman, D., Coon, H.M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin, 128,* 3-72.

Oyserman, D., & Lee, S.W.S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin, 134,* 311-342.

Pett, M.A. (1997). *Nonparametric statistics for health care research: Statistics for small samples and unusual distributions.* Thousand Oaks, CA: Sage.

Poehner, M.E. (2007). Beyond the test: Dynamic assessment and the transcendence of

mediated learning. *Modern Language Journal*, *91*, 323-340.

Poehner, M.E. (2008). Both sides of the conversation: the interplay between mediation and learner reciprocity in dynamic assessment. In J.P. Lantolf & M.E. Poehner (Eds.), *Sociocultrual theory and the teaching of second languages* (pp. 33-56). London: Equinox.

Rieber, R.W., & Carton, A.S. (Eds.) (1993). *The collected works of L.S. Vygotsky: Volume 5: Child Psychology*. Trans. Knox, J.E & Stevens, C.B. Plenum: N.Y.

Riggenbach, H. (1998). Evaluating learner interactional skills: Conversation at the micro level.. In A.W. He & R. Young (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 53-67) Amsterdam: J. Benjamins.

Ross, S. (1998). Divergent frame interpretations in language proficiency interview interaction. In A.W. He & R. Young (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 333-353). Amsterdam: J. Benjamins.

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest schematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Savignon, S.J., & Sysoyev, P.V. (2002). Sociocultural strategies for a dialogue of cultures. *Modern Language Journal*, *86*, 508-24.

Schegloff, E.A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361-382.

Seedhouse, P. (1997). The case of the missing ‘no.’: The relationship between pedagogy and interaction. *Language Learning, 47*, 547-583.

Sharkey, W.F., & Singelis, T.M. (1995). Embarrassability and self-construal: A theoretical integration. *Personality and Individual Differences, 19*, 919-926.

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research, 8*, 263-300.

Singelis, T.M. (1994). The measurement of independent and interdependent self construals. *Personality and Social Psychology Bulletin, 20,* 580-591.

Singelis, T.M., & Brown, W.J. (1995). Culture, self, and collectivist communication: Linking culture to individual behaviour. *Human Communication Research, 21*, 354-389.

Song, M.-J. (1994). A study on common factors affecting East Asian students' English oral

interaction. *English Teaching, 49*, 191-219.

Spratt, M., Humphreys, G. & Chan, V. (2002). Autonomy and motivation: Which comes first? *Language Teaching Research, 6*, 245-266.

Sternberg, R., & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge: Cambridge University Press.

Triandis, H.C. (1989). The self and social behaviour in differing cultural contexts. *Psychological Review, 96,* 506-520.

Triandis, H.C., & Gelfand, M.J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology, 74,* 118-128.

Triandis, H.C., & Trafimow, D. (2001). Cross-national prevalence of collectivism. In C. Sedikides & M.B. Brewer (Eds.), *Individual Self, Relational Self, Collective Self* (pp. 259-276). Philadelphia: Taylor & Francis.

Trafimow, D., Triandis, H.C., & Goto, S.G. (1991). Some tests of the distinction between the private self and the collective self. *Journal of Personality and Social Psychology, 60,* 649-655.

UCLES. (2005). *Cambridge IELTS™: Examination papers from University of Cambridge ESOL examinations: English for speakers of other languages*. Cambridge: Cambridge University Press.

UCLES. (2009). Information for Test-takers. Retrieved from http://www.IELTS™.org/test_takers.

Utz, S. (2004). Self-construal and cooperation: Is the interdependent self more cooperative than the independent self? *Self and Identity, 3,* 177-190.

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

Wen, W.P., & Clement, R. A. (2003). Chinese conceptualization of Willingness to Communicate in ESL. *Language, Culture and Curriculum, 16,* 18-38.

Wertsch, J.V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Wigglesworth, G. (2001). Influences on performance in task-based oral assessments. In

M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 186-209). Edinburgh: Pearson.

Yoshida, R. (2008). Teachers' choice and learners' preference of corrective feedback types. *Language Awareness, 17,* 78-93.

Young, R., & Halleck, G.B. (1998). Let them eat cake! Or how to avoid losing your head in cross-cultural conversations. In A.W. He, & R. Young (1998), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 355-382). Amsterdam: J. Benjamins.

Appendix A

Regulatory Scale (RS)

0. Construction of a "collaborative frame" prompted by the presence of the examiner as potential dialogic partner.

1. Examiner indicates that something may be wrong in a speaking turn ("Sorry?").

2. Examiner rejects unsuccessful attempts at recognizing the error.

3. Examiner narrows down the location of the error (e.g., examiner repeats or points to the specific speaking turn that contained the error).

4. Examiner indicates the nature of the error, but does not identify the error (e.g., "There was something wrong with the tense marking there").

5. Examiner identifies the error ("You can't use an auxiliary there").

6. Examiner rejects learner's unsuccessful attempts at correcting the error.

7. Examiner provides clues to help the learner arrive at the correct form (e.g., "It is not really past but something that is still going on").

8. Examiner provides the correct form.

9. Examiner provides some explanation for use of the correct form.

10. Examiner provides examples of the correct pattern when other forms of help fail to produce an appropriate responsive action.

Appendix B

Main Study Participant Background Information

Name: _____

Age: _____

Number of years of formal English study: _____

Other important sources of English learning (self-study, pen-pals, books, movies, etc.):
_____

Length of stay in Canada (so far): _____

Length of previous English study-abroad experience (if any): _____

Length of time spent in English-speaking countries: _____

Age of first contact with English: _____

Familiarity with the IELTS test: _____

Familiarity with other major English tests (TOEFL, TOEIC, TSE, MLAB, etc.):
_____

Years and scores of previous English tests (if any): _____

Appendix C

Self Construal Scale[10]

INSTRUCTIONS

This is a questionnaire that measures a variety of feelings and behaviours in various situations. Listed below are a number of statements. Read each one as if it referred to you. Beside each statement write the number that best matches your agreement or disagreement. Please respond to every statement. Thank you.

1=STRONGLY DISAGREE	4=DON'T AGREE OR	5=AGREE SOMEWHAT
2=DISAGREE	DISAGREE	6=AGREE
3=SOMEWHAT DISAGREE		7=STRONGLY AGREE

____1. I enjoy being unique and different from others in many respects.
____2. I can talk openly with a person who I meet for the first time, even when this person is much older than I am.
____3. Even when I strongly disagree with group members, I avoid an argument.
____4. I have respect for the authority figures with whom I interact.
____5. I do my own thing, regardless of what others think.
____6. I respect people who are modest about themselves.
____7. I feel it is important for me to act as an independent person.
____8. I will sacrifice my self interest for the benefit of the group I am in.
____9. I'd rather say "No" directly, than risk being misunderstood.
____10. Having a lively imagination is important to me.
____11. I should take into consideration my parents' advice when making education/career plans.
____12. I feel my fate is intertwined with the fate of those around me.
____13. I prefer to be direct and forthright when dealing with people I've just met.
____14. I feel good when I cooperate with others.
____15. I am comfortable with being singled out for praise or rewards.
____16. If my brother or sister fails, I feel responsible.
____17. I often have the feeling that my relationships with others are more important than my own accomplishments.
____18. Speaking up during a class (or a meeting) is not a problem for me.
____19. I would offer my seat in a bus to my professor (or my boss).
____20. I act the same way no matter who I am with.
____21. My happiness depends on the happiness of those around me.
____22. I value being in good health above everything.
____23. I will stay in a group if they need me, even when I am not happy with the group.
____24. I try to do what is best for me, regardless of how that might affect others.
____25. Being able to take care of myself is a primary concern for me.
____26. It is important to me to respect decisions made by the group.
____27. My personal identity, independent of others, is very important to me.
____28. It is important for me to maintain harmony within my group.
____29. I act the same way at home that I do at school (or work).
____30. I usually go along with what others want to do, even when I would rather do something different.

[10] Singelis (1994)

Appendix D

Sample Simulated IELTS™ Practice Speaking Test[11]

Part 1

The examiner asks the test-taker about him/herself, his/her home, work or studies and other familiar topics.

EXAMPLE

Family

- Do you have a large family or a small family?
- Can you tell me something about them?
- How much time do you manage to spend with members of your family?
- What sorts of things do you like to do together?
- Did/Do you get on well with your family? (Why?)

Part 2

Describe a teacher who has influenced you in your education. You should say: -where you met them -what subject they taught -what was special about them and explain why this person influenced you so much.

You will have to talk about the topic for 1 to 2 minutes. You have one minute to think about what you are going to say. You can make some notes to help you if you wish.

Part 3

Discussion Topics:

Development in Education

Example Questions:

How has education changed in your country in the last 10 years?

What changes do you foresee in the next 50 years?

A national education system

Example Questions:

How do expectations of today's school leavers compare with those of the previous generation?

What role do you think extracurricular activities play in education?

Different styles/methods of teaching and learning

Examples Questions:

What method of learning works best for you?

How beneficial do you think it is to group students according to their level of ability?

[11] UCLES, 2005, p. 29

Appendix E

Email Survey

1. In the speaking tests, you had to talk about different topics, and you also had to make correct sentences. When you were speaking, did you...
   (a) ...think more about making correct sentences (no mistakes)?
   OR
   (b) ... think more about the topics that you were talking about? Is your answer A or B?
2. In the speaking tests, you had to talk about different topics, and you also had to make correct sentences. Because I was an examiner, did you feel that...
   (a) I wanted you mostly to focus on making correct sentences (no mistakes)?
   (b) I wanted you mostly to focus on talking about different topics? Is your answer A or B?
3. In half the speaking tests, I interrupted you to correct your mistakes. In those 3 tests, because I was correcting you, did you feel that...
   (a) I wanted you mostly to focus on making correct sentences (no mistakes)?
   (b) I wanted you mostly to focus on talking about different topics? Is your answer A or B?
4. In half the speaking tests, I interrupted you to correct your mistakes. When I interrupted you, I always said "Sorry?". Did you...
   (a) ...ALWAYS understand that I wanted to correct you?
   (b) ...SOMETIMES understand that I wanted to correct you? Is your answer A or B?
5. In half the speaking tests, I interrupted you to correct your mistakes. In these tests, did my interruptions make you feel...
   (a) ...nervous in the first test, but by the last test you felt comfortable?
   (b) ...nervous for all 3 tests?
   (c) ...not really nervous for any of the tests? Is your answer A, B or C?
6. Did you feel more comfortable by the last test because...
   (a) You were familiar with me?
   (b) You were familiar with the style of the speaking tests?
   (c) I didn't feel comfortable by the last test. Is your answer A, B, or C?

Participant responses to email survey

	Q1	Q2	Q3	Q4	Q5	Q6
P1	B	A	A	A	A	B
P2	A	B	A	A	A	A
P3	B	B	A	A	C	C
P4	B	B	B	B	C	A
P5	A	B	A	B	A	A
P6	B	B	B	A	C	B
P7	B	A	A	B	B	B

Appendix F

Transcription Conventions[12]

(...)	A short pause
-	False start
<u>Underlined</u>	Speaker emphasis
[	Beginning of overlapping speech
Ah:::	Lengthening of the previous sound
<hhh>	Intake of breath

[12] Ellis and Barkhuizen, 2005, pp. 226-227