VALIDATING TASK-BASED ASSESSMENT OF L2 PRAGMATICS IN INTERACTION USING MIXED METHODS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

SECOND LANGUAGE STUDIES

AUGUST 2013

By

Soo Jung Youn

Dissertation Committee:

James Dean Brown, Chairperson
Graham Crookes
Thom Hudson
Gabriele Kasper
Seongah Im

for my mother, Myeongsoon Jung, in loving memory

ACKNOWLEDGEMENTS

I am very grateful to the members of my dissertation committee, Dr. James Dean Brown, Dr. Thom Hudson, Dr. Gabriele Kasper, Dr. Graham Crookes, and Dr. Seongah Im. I also thank Dr. John Norris for his critical advice at an earlier stage of my dissertation. I sincerely appreciate their time, support, encouragement, and help whenever necessary on every step of my journey. First and foremost, I owe special gratitude to my academic advisor Dr. James Dean Brown for understanding me and believing in my decisions throughout my study, which I am most grateful for. His unflagging encouragement, critical advice, and generous support will never be forgotten. My special appreciation also goes to Dr. Thom Hudson. His expertise, shrewd advice whenever needed, and thoughtful attention to the details of my dissertation were invaluable, which made my dissertation better than it was. Dr. Gabriele Kasper's guidance and help have been instrumental as well. Her inspiring work, demand for excellence, exquisite attention to details, and sage advice during my study has always motivated me to do better.

I would like to acknowledge my research participants' help and ELI director Kenny Harsch's support in the data collection. This dissertation was also possible thanks to generous funds from Small Grants for Doctoral Research in Second or Foreign Language Assessment from ETS, Language Learning Dissertation Grant, and Elizabeth Carr Holmes Scholarship Fund from the Department of Second Language Studies at UH.

My heartfelt gratitude also goes to Anna, Hyunchul, Sihyun, and Soohee for their love, support, and laughter when I needed it most. I thank them, now and always. Finally, I am deeply indebted to my parents who taught me the value of hard work and perseverance, not to mention their love and prayers.

ABSTRACT

This study investigates the validity for task-based assessment of L2 pragmatics in *interaction* in an English for Academic Purposes (EAP) setting for classroom assessment and meaningful score interpretations for stakeholders. As a validity framework, Kane's (2006) argument-based approach to validity was employed. In view of the complexity of assessing L2 pragmatics in *interaction* and score interpretations from observed L2 pragmatic performances, score interpretations were built upon interpretive arguments comprising of four inferences, which guided the types of data analyzed and research steps taken in this study. Using the sequential mixed methods approach (Green, 2007; Tashakkori & Teddlie, 2003), qualitative and quantitative evidence was collected to provide valid support for the inferences and to strengthen the validity argument.

Based on a large-scale needs analysis on EAP L2 pragmatic learning needs, two open role-play tasks that are meaningful and relevant to stakeholders in an EAP context were developed. Unlike the closed role-play task format, examinees were allowed to negotiate and naturally interact with interlocutors. In order to provide meaningful score interpretations to the stakeholders and to assist raters in making accurate evaluations of examinees' interaction-involved L2 pragmatic performance, conversation analysis (CA)-informed analytical rating criteria with detailed descriptions were developed based on the detailed analyses of examinees' L2 pragmatic performances in the role-play tasks. One hundred two adult ESL examinees completed the role-play tasks and monologic tasks. Four rater groups, consisting of 12 raters in total, scored each examinee's pragmatic performance. A many-facet Rasch measurement using FACETS (Linacre, 2006) indicated that the role-play tasks displayed different levels of difficulty, which stably

differentiated between the varying degrees of the 102 examinees' pragmatic abilities. The raters showed internal consistency despite their different degrees of severity. Stable fit statistics and distinct difficulties were reported within each of the interaction-sensitive rating criteria, indicating that they contribute to measuring L2 pragmatic competence. In particular, the two rating categories for interactional competence were distinct in their difficulty levels, which supports the CA findings.

Based on the qualitative and quantitative analyses, all of the validity evidence was woven into the validity argument for task-based assessment of L2 pragmatics in interaction, focusing on the four types of inference, *domain description*, *evaluation*, *generalization*, and *extrapolation*. The current study exemplifies how the construct of L2 pragmatics in interaction can be operationalized based on the qualitative scrutiny of the target performance that is also supported by the quantitative findings, which contributes to the advancement of measuring L2 pragmatics. Lastly, this study provides additional grounds for the recent development of the validity argument approach in the field of language assessment at large.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

STATEMENT OF RESEARCH GAPS

Introduction

This dissertation examines the validity argument for task-based assessment of L2 pragmatics in *interaction* in the context of English for academic purposes (EAP) using mixed methods. In order to locate my dissertation in the field of L2 pragmatic assessment, a discussion of research gaps in L2 pragmatic assessment research is relevant based on a synthesis of previous research in L2 pragmatic assessment. Thus, this chapter starts with an overview of previous L2 pragmatic assessment research followed by a discussion on three research gaps in the theory, methodology, and intended uses of previous L2 pragmatic test instruments.

Overview of Previous L2 Pragmatic Assessment Research

L2 pragmatics assessment is a relatively late arrival in language testing research compared to other components of language competence. Yet, a growing body of research on assessing L2 pragmatics has been available over the last decade. Various theories of pragmatics have influenced the operationalizing of L2 pragmatics as an assessment construct, particularly Searle's speech act theory (1969, 1975), politeness theory (Brown & Levinson, 1987), and the distinction between sociopragmatics and pragmalinguistics (Leech, 1983). Aside from a few early studies on assessing L2 pragmatics (Farhady, 1980; Shimazu, 1989), one of the most systematic and extensive of the earlier studies on assessing L2 pragmatics was conducted by Hudson, Detmer, and Brown (1992, 1995). In their framework, careful consideration of three sociological variables from politeness

theory (Brown & Levinson, 1987), social distance, relative power between interlocutors, and absolute ranking of imposition in the particular culture, played an important role in operationalizing L2 pragmatics. To assess the three speech acts, request, refusal, and apology, six prototype measures were developed. The measures included multiple-choice discourse completion tests (MCDCTs), open-ended written DCT (OPDCT), oral DCT, role-play, self-assessment for the DCT, and self-assessment for the role-play. These six instruments were designed to assess diverse aspects of pragmatic competence. For example, the OPDCTs and oral DCTs, measured learners' offline written and spoken pragmatic production by first having learners either read or listen to certain situations and then either write or speak about they would do. On the other hand, the role-plays measured learners' online pragmatic performance by requiring them to have a conversation with an interlocutor on various situations. To score learners' pragmatic performances, six analytical rating criteria, which reflect the two intersecting domains of pragmatics, sociopragmatics and pragmalinguistics (Leech, 1983) were developed as well: (a) ability to use the correct speech act, (b) formulaic expression, (c) amount of speech in a given situation, (d) formality level, (e) directness level, and (f) overall politeness level.

Since then, the Hudson *et al.* framework has been greatly influential in subsequent L2 pragmatics assessment. The test instruments have been translated into other foreign languages, including Japanese (e.g., Yamashita, 1996) and Korean (e.g., Ahn, 2005; Youn, 2008), and have been used in assessing Japanese ESL students' pragmatics competence (e.g., Yoshitake, 1997). The framework was consistently reported as having a reasonably high reliability and validity measure for all of the test types except the

multiple-choice DCT. However, according to Brown (2000, 2001) and Hudson (2001a, 2001b), who compared results from the previous studies, variation in the method effects among the different test types was reported depending on the group of examinees. The Hudson *et al.* framework was also adapted as the research instruments to assess learners' L2 pragmatic abilities (e.g., Taguchi, 2007a; Takimoto, 2009). In addition to examining the reliability and validity of the Hudson *et al.* instruments, methodological and pedagogical concerns regarding the test instruments were investigated as well. For example, Brown and Ahn (2011) investigated the relative effects of various facets, including raters, functions (Power, Distance, Imposition), and item types using G-theory. Findings suggest raters and item types were relatively important sources of error. Interestingly, role-play was found to have the most consistent variance.

Other studies developed their own instruments using diverse methodological and analytical foci to assess pragmatic competence (e.g., Grabowski, 2009; Liu, 2007; Roever, 2005, 2006, 2007; Tada, 2005; Timpe, 2012; Walters, 2007; Youn, 2010). Roever (2005, 2006) investigated the construct validity of pragmalinguistics using routines, implicatures, and speech acts as web-based assessment instruments. Various test types were employed including the MCDCT format for routines and implicatures and written DCT with a rejoinder for speech acts. A strong L2 proficiency effect was found on both implicatures and speech acts, although routines were more related to the amount of L2 exposure than L2 proficiency. Liu (2007) also engaged in research on the assessment of offline pragmalinguistics, specifically focusing on the development of MCDCT items in a Chinese EFL context to further investigate mixed results reported on MCDCT test items in previous studies. While Roever (2005, 2006) and Tada (2005)

found reliable results using MCDCTs, the studies that employed MCDCT from Hudson *et al.* framework reported low reliability possibly due to the challenge of developing plausible distracters in MCDCTs. Considering the critical importance of developing sound distracters in MCDCTs, Liu (2007) carefully developed MCDCT items using various analyses, such as pilot study results from English native speakers and ESL learners, statistical analyses, and verbal protocol data from examinees. Liu's research findings show how MCDCT test items can reliably and practically measure Chinese EFL learners' pragmatic competence.

Despite the dominance of quantitative research, some studies have employed qualitative analytic perspectives to investigate the construct of L2 pragmatics (e.g., Grabowski, 2009; Walters, 2007, 2009). Grabowski (2009) employed a mixed methods approach to examine the construct validity of pragmatics and grammar. In addition to statistical analyses, Grabowski used discourse analysis to examine the pragmatic meanings displayed in responses from native speakers and nonnative speakers to further support statistical analyses. Other qualitative studies were done by Walters (2007, 2009). Employing a hermeneutic approach to validity, Walters argues for the questionable validity of DCT-elicited test results, pointing out the limitations of DCT instruments. For example, Walters points out that DCT instruments do not consider actual language behavior, such as the sequential organization of talk-in-interaction. As an alternative, Walters employed conversation analysis-informed test methods (CAIT) to serve as a more valid way of assessing pragmatic competence. In his study, pragmatic competence was operationalized as three sub-skills, assessment, compliment, and pre-sequence, and the study used a closed role-play as the test instrument in which three oral pragmatics

prompts were embedded. The study hypothesized that CA-trained raters can better assess targeted pragmatic competence than raters only familiar with speech act-theoretic taxonomies of utterances, as the CA-trained raters would be able to notice certain interactional features relevant to pragmatic competence. The study cautiously argues that the pilot CAIT is practical, in terms of test administration and rating efficiency, and provides empirically valid rating results, although further investigations are indeed necessary.

Research Gaps in L2 Pragmatic Assessment

As reviewed above, some studies focus on developing valid and reliable L2 pragmatic assessment instruments and others address various methodological issues involved in assessing L2 pragmatics. These studies have greatly contributed to advancing our understandings of L2 pragmatics as an assessment construct. However, some research gaps do exist. From a theoretical perspective, previous studies primarily defined and operationalized L2 pragmatic competence in individual-cognitive terms from a rationalist speech act perspective. This approach inherently put less emphasis on the discursive aspects of L2 pragmatics, namely how pragmatic meanings can be achieved in the turn-by-turn sequences of *interaction* (Kasper, 2006). Such under-emphasis of L2 pragmatics in *interaction* along with the lack of understanding of the co-constructed nature of interaction lead to the under-representation of the construct and validity threat for interaction-involved assessment of L2 pragmatic performance specifically (Roever, 2011) and interaction-involved language proficiency generally (e.g., McNamara, 1997; McNamara & Roever, 2006, Roever, 2011; Young & He, 1998). Thus, reorientation in

the theoretical conceptualization of L2 pragmatics is crucial to ensure the validity of L2 pragmatic performance assessment.

Methodologically, DCTs and closed role-play task, as they are predominantly used, are limited in their ability to provide valid evidence for measuring L2 pragmatic performance. DCTs are a common research instrument in research assessing L2 pragmatics due to their practicality in administration. Written and oral DCTs, in which examinees either read or listen to certain situations and are asked to write or speak about what they would do, measure learners' offline spoken and written pragmatic production rather than performance. Although these instruments have been reported as reliable and practical, the validity of DCTs has long been questioned. As shown in Golato (2003), compliment expressions elicited from DCTs were different from naturally occurring conversations. Therefore, such emphasis on the practicality of test instruments at the expense of authenticity is a potential threat to the validity of L2 pragmatic assessment. When it comes to measuring L2 pragmatic performance in interaction, a closed role-play task where speakers are asked to follow predetermined interactional outcomes (e.g., accept request rather than decline) in role-play scenarios has been widely employed. However, the validity of the closed role-play format has been also questioned (Kasper & Rose, 2002). The fixed role-play scenarios might not reflect authentic pragmatic performance, as speakers usually do not know interactional outcomes beforehand. Furthermore, the systematic development of reliable and valid interaction-sensitive rating criteria for online pragmatic performance has not been emphasized, which is another threat to the validity of performance assessment (Linn, Baker, & Dunbar, 1991; Messick,

1994). A systematic and valid approach to developing test instruments for assessing L2 pragmatics *in interaction* is necessary.

Lastly, the intended uses and test impact of previous L2 pragmatic assessment research have been limited. Previous studies primarily focused on developing valid and reliable test instruments to assess learners' general pragmatic competence, rather than their assessment practices for specific learning contexts, resulting in the narrow interpretations of test scores. In particular, classroom assessment uses of L2 pragmatic assessment in relation to language program curriculum or learners' pragmatic learning needs have not been emphasized in previous research. As a result, sound L2 pragmatic assessments are available for mainly norm-referenced test purposes, and such test instruments are not intended for specific L2 educational settings in assessing learners' pragmatic learning progress. More attention on utilizing L2 pragmatic assessment practices for specific educational contexts is needed

These research gaps can jeopardize the validity of L2 pragmatic assessment. Responding to such research gaps in assessing L2 pragmatics, the present study investigates a validity argument of task-based assessment of L2 pragmatics in *interaction* in the context of EAP. In view of the complexity of assessing L2 pragmatics in *interaction* and score interpretations from observed performances, score interpretations were based on the building blocks of interpretive arguments composed of various inferences and assumptions using mixed methods (Greene, 2007; Tashakkori & Teddlie, 2003).

Overview of the Dissertation

Chapter 2 discusses two theoretical frameworks used by the current study. First, in order to conceptualize and operationalize the construct of L2 pragmatics in *interaction*, the discursive approach to pragmatics (Kasper, 2006) is discussed. Interactional competence in language testing research is reviewed, as it is relevant to conceptualizing L2 pragmatics in *interaction*. Second, as the validity framework for the current study, Kane's (1992, 2006) argument-based approach is considered in detail. Chapter 3 introduces the current study's methodological frameworks, looking first at the mixed methods framework used in this study. The rationales for employing mixed methods to strengthen the validity argument in the current study along with definitions and the current status of mixed methods at large will be explicated. Second, a task-based language assessment (TBLA) framework and an approach to performance data-driven rating rubric development are reviewed. The review focuses on how TBLA's characteristics and strengths can be useful to operationalize L2 pragmatic competence. Additionally, as the development of the rating criteria to measure interaction-involved L2 pragmatic performance in an EAP context is crucial in this study, the benefits of the data-driven rating criteria approach in measuring context-specific language performance are also discussed. Chapter 4 then details the research questions and data collection methods. Chapter 5 provides the qualitative and quantitative results, and Chapter 6 discusses the research questions. In particular, the research questions address assumptions and warrants made in the structure of the interpretive validity argument and each of them are discussed in the form of the validity argument. Lastly, in Chapter 7, the dissertation ends by looking at the study's limitations, implications, and future research.

CHAPTER 2

THEORETICAL FRAMEWORK

This chapter discusses the two theoretical frameworks, one for the construct of L2 pragmatics in interaction and another for the validity framework. The chapter first starts with a discursive approach to speech act pragmatics (Kasper, 2006) along with conversation analysis (CA)'s contributions to L2 pragmatic research. Considering that the target construct being investigated in this study involves interaction in L2 pragmatic performance, interactional competence is also discussed, focusing on its theoretical backgrounds and how it has been measured in language testing. The second part of the chapter starts with a brief overview of conceptions of validity in the field, followed by Kane's validity argument focusing on its strengths and characteristics.

Discursive Approach to L2 Pragmatics and CA

L2 pragmatics has been conceptualized under ontologically distinct theoretical frameworks over the last three decades (Kasper, 2009). Prevalently, L2 pragmatics has long been influenced by theories and concepts from rationalist pragmatics, particularly Searle's speech act theory (1969, 1975) and politeness theory (Brown & Levinson, 1987). A theoretical premise of these approaches is that speakers are seen as individual rational actors who choose their means to meet the actor's intended goals following encoding/decoding communication models. Here, such key concepts as *intentionality* put an emphasis on the listener's recognition and inference of the speaker's intention (Grice, 1957). Here, the following two notions are critical to understand rationalist pragmatics. Firstly, the notion of 'recognition of intention' is critical to conceptualize communication

or comprehension where a speaker produces an utterance that causes the listener to formulate the particular psychological states the speaker intends. Additionally, the notion of 'rationality' is critical in conventional speech act research, which is "defined as a goal-directed means-ends relationship" (Kasper, 2006, p. 285). It is necessary to be rational to have an effective means-ends reasoning which enables a rational actor to choose means for intended goals.

Following the conventional speech act research paradigm, research orientation in early interlanguage pragmatics has been grounded with a model of cross-cultural pragmatics from a speech act model (Searle, 1969), and therefore it was essentially comparative rather than acquisitional (Bardovi-Harlig, 1999; Kasper & Schmidt, 1996). When it comes to theorizing L2 pragmatic learning, an individual-cognitive perspective has been dominant. Inherently, primary interlanguage pragmatic research interests have focused on how individual learners develop their ability to comprehend and produce pragmatic meaning and learn diverse pragmatic objects in L2 (Kasper & Blum-Kulka, 1993). Such learning theories in SLA as the two dimensional models (Bialystok, 1993) and noticing hypothesis (Schmidt, 2001) have been employed to examine L2 pragmatic development from a cognitive processing perspective. Along with these theoretical premises, an extensive range of interlanguage pragmatics research issues has been investigated (Kasper, 2009; Kasper & Rose, 2002) including pragmatic transfer (Takahashi, 1996), learning context and environment (Olshtain & Blum-Kulka, 1985), pragmatic comprehension and production (Taguchi, 2007a, 2007b), effects of instruction (Alcón Soler & Martínez-Flor, 2005; Bardovi-Harlig, 2001; Jeon & Kaya, 2006;

Takahashi, 2001), and a relationship between grammar and pragmatics (Bardovi-Harlig, 1999; Bardovi-Harlig & Dörneyi, 1998; Niezgoda & Röver, 2001; Schauer, 2006).

However, there have been critiques and expressions of discontent with these theoretical notions, data collection methods, and data analysis within the long-standing rationalist speech act research. Especially, disagreement to applying the rationalist pragmatic theory to interaction-involved pragmatics have been discussed and diverse alternative theoretical approaches to extending speech act theory have been proposed in order to account for interaction empirically. As surveyed in D'hondt (2009), such attempts are rooted in diverse intellectual and disciplinary foundations including Goffman's explanation of the interaction order (1983), conversation analysis (CA) (Sacks, Schegloff, & Jefferson, 1974), Garfinkel's ethnomethodology (1967), and interactional sociolinguistics (Gumperz, 1982). Among these, CA has been considered as a highly productive alternative to rationalist pragmatics especially due to CA's emphasis on the centrality of empirical evidence from an emic perspective and its sound conceptual and methodological framework on talk-in-interaction (e.g., Arundale, 1999, 2005, 2008; Bilmes, 1993; Clark, 1996; Sanders, 1987; Schiffrin, 1987).

With its roots in ethnomethodology (Garfinkel, 1967) under the field of sociology, CA aims at finding interactional organization of how interactants produce and understand in talk-in-interaction by closely examining sequential organizations of turns in not only ordinary conversation but also institutional talk (Sacks, Schegloff, & Jefferson, 1974). One of the conceptual and theoretical rationales of such analytical endeavor is the importance of conceptualizing conversation in its own right (Drew, 2005). Namely, conversation is regarded as the primordial site of social action as people *do* things largely

through conversation. Equally importantly, conversation is viewed as practical action whose meaning is jointly constructed and established through speakers' use of language rather than considering language as mere representation of reality.

Central to understanding CA, one of the fundamental organizations of talk-in-interaction is the organization of *turn-taking*. As the extensive CA literature shows, the turn-taking organization for conversation is extremely effective and robust. Other empirically uncovered types of interactional organization include *adjacency pair* and *preference organization* (Hutchby & Wooffitt, 1998; Schegloff, 2007; Seedhouse, 2004). First type is adjacency pairs, which is paired utterances by different speakers that become conditionally relevant, such as *greeting-greeting*, *question-answer*, *offer-accept/decline*, or *grant a request-thank*. It should be noted that the adjacency pair concept is not about claiming that second parts should be provided for first parts. But, it is about *normative* accountability of actions, which is the key to understanding the ethnomethodological basis of CA (Seedhouse, 2004). Based on this *normative* frame of reference of actions, an absence of the second part can be interactionally accounted for.

A second different, but related, concept is preference organization. Generally speaking, some actions implemented in sequence organization are positive or *preferred* such as acceptance, granting, and agreements, while some are negative or *dispreferred*, such as rejections, declinings, and disagreements. It should also be noted that the concept of preference and dispreference is not about liking or wanting, but related to issues of affiliation, disaffilitation, accountability, and sactionability conducive to social solidarity, which are also derived from ethnomethodogical principles (Seedhouse, 2004, p. 23). In terms of turn-taking structures of these actions, preferred actions invoke overlap between

turns or are exchanged without any delay. However, for dispreferred actions, such as disagreement, often next turn is delayed with interactionally appropriate pauses. Breaking such norms of turn-taking is noticeable or interactionally accounted for. However, there are exceptions. For example, agreement can be dispreferred and disagreement can be preferred after a self-deprecation (e.g., I'm so dumb I don't even know it!) (Pomerantz, 1984, pp. 83–90).

Following CA's theoretical and empirical undertakings, key notions in rationalist pragmatics research, such as intentionality and action, have been challenged and reconceptualized. One of the active discussions centers on the notion of intentionality (e.g., Arundale, 1999, 2008; Duranti, 2006; Edwards, 2006, 2008; Grice, 1957; Haugh, 2008; Heritage, 1990/1; Levinson, 2006). Arundale (1999, 2008) argues the Gricean notion of intention recognition conveys a monologic view of talk and is descriptively inadequate as it is based on "the unexamined assumption that the summative pairing of the cognitive processing activities of two separate individuals provides *sufficient* accounts for what is empirically an interactive phenomenon" (Arundale, 2008, pp. 240–241). Arundale further argues for a lack of attention on "how the speaker and recipient come to know whether the recipient's inference and attribution is consistent with the speaker's intention, and hence whether or to what degree they have engaged in communication" (Arundale, 2008, p. 252). Essentially, Arundale advocates 'interactional achievement' models of communication that emphasize how meaning in communication is achieved in and through interaction, which coincides with one of the key tenets of CA.

When it comes to the seminal notion of 'action' in pragmatics, CA contributes the fundamental insight that actions are jointly accomplished in *sequences* by participants

rather than by outcomes of an individual rational actor's mind. As such, interpreting pragmatic meanings within interaction should be based on a recipient's response to what the co-participant produced in a prior turn in its sequential organization, rather than linguistic conventions or what transferred between individual minds. Recognizing the need for critical examinations of L2 speech act research's theoretical premises and analytical practices, Kasper (2006) argues for a discursive approach to speech act research from a CA perspective. Kasper empirically shows how sequential analysis of participants' turns-at-talk enables us to examine how the participants establish pragmatic meaning and accomplish actions jointly in organized sequences, without assuming any actions intended by participants as analysts. Furthermore, as evidenced in Kasper's analyses, such rationalist notions as 'illocutionary ambiguity' is not a concern anymore in discursive pragmatics from participants' point of view, as any ambiguity that might rise during interaction can be clarified by participants themselves by displaying understanding of the previous utterance in its sequential organization. Following the emic perspective of CA, Kapser notes analysts do not need to involve analyzing and interpreting participants' motivations, intentions, and other mental states unless participants themselves orient such mental events which becomes relevant with the topic of their talk.

Such perspective on the discursive approach to L2 pragmatics in interaction has been taken up in numerous empirical studies which examined various areas of L2 pragmatics study in conjunction with diverse learning theories, such as L2 Japanese speakers' use and learning of the Japanese particle *ne* and modal expressions (Ishida, 2006), L2 Korean speaker's development pattern of Korean discourse connective *–nuntey*

(Kim, 2009), and CA investigation of pragmatic learning opportunities in L2 Japanese classroom (Tateyama & Kasper, 2008)

Conceptualizing Interactional Competence

The discussion of interactional competence has emerged from diverse research traditions in social sciences. One of the vigorous discussions centers around characterizing interactional competence as one of the language competences. The discussion on language competence traces back to Chomsky's (1965) distinction between competence and performance which invoked critiques from scholars and led to the development of various models on language competence, which have greatly influenced L2 teaching and assessment (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972). The term 'interaction' has appeared in these models and has been interpreted as competence of speakers' communicative abilities in interactional contexts.

Communicative competence has been mainly characterized as individual cognitive knowledge systems among applied linguists (Bachman, 2007; Doughty & Long, 2003). However, in an effort to conceptualize the construct of 'interaction' both theoretically and empirically, the predominant focus on individual-based cognitive charge of interaction bas been criticized and a call for a balanced view toward interaction has been raised to encompass its co-constructed nature (e.g., McNamara, 1997, 2001; Young, 2000). Kramsch (1986) first used the term 'interactional competence' in contrast with the notion of language proficiency as the target of L2 learning, and problematized previous models on competence and called for contemporary and transformed understandings of the competence. Arguing for a reconceptualization of SLA, Hall (1995) coined a notion

of ‘interactive practices’ to describe talk in interaction. Opposing the idea of an individual who is creating utterances regardless of social constraints, she described interactional practices as co-constructed by participants in relation with sociocultural significance employing diverse linguistic and pragmatic resources to the practice. He and Young (1998) developed their own theoretical notions of the term that differs from communicative competence in several accounts and motivated subsequent research on L2 intearactional competence not only in SLA research but also language testing research. Taken together, He and Young characterize interactional competence as joint construction of interactive practices by all participants, which is specific to that practice. Here, they emphasize that the interactional practices acquired by an individual are not general but local and practice-specific interactional competence.

Central to the concept of interactional competence proposed by several applied linguists, co-construction and intersubjectivity have been critical among others. With its roots in a wide range of disciplinary perspectives, including CA, linguistic anthropology, ethnomethodology, and Soviet psychology, the notion of co-construction has been fundamental to understand human social interaction, mind, and child language development (Jacoby & Ochs, 1995). Jacoby and Ochs refer to co-construction as “the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality” (1995, p. 171). The concept covers a wide range of interactional processes including collaboration, cooperation, agreement, or disagreement, whether it is affiliative or supportive or not. In this concept, conversation is conceptualized as a primordial site for constructing social meaning and action. Regarding implications of the co-constructed view of interaction, they note

"language, discourse, and their effects cannot be considered deterministically preordained by alleged "inherent" properties of linguistic structures, by assumed constructs of individual competence and so-called shared knowledge, or by assigning participants to membership categories presumed to be relevant to the occasion" (p. 177). Also, taking the co-constructed view of interaction means that participants, as active agents rather than passive robots, respond to the rich flow of unique interactional moments on-line which entails complex human cognition and communication behavior.

The notion of intersubjectivity is also closely tied with interactional competence (Heritage, 1984; Seedhouse, 2004; Young, 2008). During talk in interaction, intersubjectivity, or publicly displayed mutual understanding or interpersonal alignment during interaction on what is going on at a moment of interaction, is accomplished and maintained. Here, CA's distinctive concept, the adjacency pair, is indispensable to understand intersubjectivity, which is "the basic building-blocks of intersubjectivity" as Heritage (1984, p. 256) put it. It is a paired utterance that consists of the first part of the pair (e.g., request) and the second part (accept). The adjacency pair has been used not only as an action template within sequential organization but also a fundamental unit for interpretation. In the adjacency pair, any first action plays a role of creating a normative expectation for a next action and becomes a basis for interpretation, which is termed as "next-turn proof procedure" in CA literature (Sacks, Schegloff, & Jefferson, 1974, p. 729). Also, the process of analyzing the adjacency pair in interactional organization is fundamental to an emic perspective to knowledge construction that focuses on explicating meaning from a participant-based perspective in and through interaction. In

this way, the process of achieving intersubjectivity can be examined from participants' points of view during interaction.

Measuring Interactional Competence

Along with emerging interests in interaction-involved speaking tests, language testers also have started paying attention to conceptualizing and operationalizing the construct of interactional competence. Despite some pessimistic views (e.g., Bachman, 2007), the discussion on addressing interactional competence in language testing has been initiated and sustained through empirical research articles, book-length publications, and special issues in refereed journals using diverse analytical attempts (e.g., Brown, 1995, 2003, 2005; Chalhoub-Deville, 2003; Chalhoub-Deville & Deville, 2005; Johnson, 2000, 2001; Kasper & Ross, 2003, 2007; Lazaraton, 2002; May, 2009, 2011a, 2011b; McNamara, 1997; McNamara, Hill, & May, 2002; Okada, 2010; Ross, 2007; Swain, 2001; Taylor & Wigglesworth, 2009; van Lier, 1989; Weir, 2005; Young, 1995, 2000, 2002; Young & He, 1998). The important premise behind such increasing interest is to figure out ways in which interactional competence is defined, ensuring the validity of performance-based assessment, especially interaction-involved assessment, such as group speaking tests, paired speaking assessment, oral proficiency interview (OPI). One of the pressing concerns for language testers when it comes to employing paired speaking tests includes the impact of one examinee on another in terms of scores given by a rater. Following up such concern, Galazi (2004, 2008) and May (2009, 2011a) investigated how raters perceive the co-constructed natures of interactional competence and how it affects on awarding scores. The following features positively affected the raters, such as

high degrees of equality in terms of topic initiations and mutuality achieved conversations. However, asymmetric patterns, such as a low level of equality between examinees and dominant/passive roles during conversation, negatively influenced the raters. Additionally, in-depth qualitative analyses on investigating interactional competence in the context of language assessment have been available. Among various methods, CA has been prevalently employed in such recent studies as Okada (2010) that investigated the validity of employing the role-play format in eliciting authentic interactional competence or Gan (2010) on examinees' interactional features in the group oral assessment context. These recent empirical studies that employed qualitative analyses on understanding interactional competence provide promising new possibilities and important implications for developing interaction-involved assessment tasks and rating criteria.

This section discussed the theoretical framework of conceptualizing L2 pragmatic performance in interaction. Following the Kasper's argument for a discursive approach to L2 pragmatics, CA's theoretical and empirical tenets that can contribute to resolving limitations of rationalist views toward speech act research were discussed. Additionally, CA has been considered as one of the most promising analytical programs to explicate how participants jointly construct pragmatic meaning and action during interaction. Inherently relating to the construct of L2 pragmatics in interaction, the discussion also focused on ways in which interactional competence has been theorized in social sciences at large and empirically addressed in the field of language testing in particular. The comprehensive understandings and subsequent operationalization of interactional competence is critical in ensuring the validity of performance assessment, which still

needs further attentions and research. Let us then move to the discussion of validity to fully understand its roles in assessment.

Validity in Assessment

This section surveys conceptions of the validity in educational measurement at large and language testing specifically. Followed by a short overview of how validity has been addressed in the field, Kane's (2006) validity argument, which is used as the validity framework in this study, is discussed in detail.

Within educational measurement, over a long period, validity theory has evolved over the course of sustained and seminal discussions among theorists and researchers, and its definition has evolved around such discussions. Early validity practices, from the 1950s through the 1970s, mainly concerned particular types of validity. According to the 1974 APA/AERA/NCME standards, validity was defined as "the appropriateness of inferences from test scores or other forms of assessment" (p. 25) and it was characterized with three interrelated aspects, *criterion*, *construct*, and *content* validity. However, it was criticized that content and criterion validity provide relevant but insufficient evidence as primarily supporting roles for the construct validity, and thus they are not logically distinct or have equal importance (Loevinger, 1957). Since then, the 1985 AERA/APA/NCME Standards for Educational and Psychological Testing changed the definition of validity as three validities with a single unified view of validity putting construct validity as central, supported by content and correlation analyses as methods to investigate construct validity. As noted in the definition, validity centers around a theoretical construct within a nomological network that could be used to generate clear

hypotheses about test performance. In 1989, Messick proposed a unified model of validity which considers not only empirical aspects of construct validation but also values and social consequences of test interpretation and use. Messick characterized validity as:

> An integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores other modes of assessment (1989, p. 13)

In Messick's view, validity is not an inherent property of a test but the degree to which we are justifying in making inferences about a construct based on test scores. Here, the most fundamental to understanding validity is what is being validated, which is not tests themselves but *interpretations* and *uses* of tests. Previously, validity was considered a characteristic of a test and focused on the extent to which a test measures what it is supposed measure. However, since Messick's conception of validity, validity is considered an argument of the extent to which test *interpretations* and *uses* can be justified based on integration of evidence.

The advances made in educational measurement also affected validity research in language testing (Chapelle, 1999). In particular, Messick's validity framework has been greatly influential for language testing research (e.g., Bachman & Palmer, 1996; McNamara, 2006). Shedding a critical light on the legacy of Messick's framework, McNamara (2006) emphasizes that Messick's framework provides the most comprehensive conceptualization of the validation process to date. Messick's views toward values and social consequences of test use allowed the discussion of ethics, impact, and washback in language testing.

At the same time, Chapelle's (1999) discussion on challenges in validity of language assessment helps us to evaluate the current status of validity research in the field. Chapelle discusses the following three challenges. Firstly, more work on different theoretical approaches to construct definition in test development and validation is needed, which is a still contentious issue. Relating to the first issue, the second challenge concerns ways in which appropriate types and numbers of justifications are identified and then integrated to make a validity argument. However, delimiting types and ways to interpret validity evidence is not a simple matter, especially considering context-specific uses for language tests. On this challenge, Chapelle (2009) notes "a validity conclusion is an argument-based, context-specific judgment, rather than a proof-based, categorical result" (p. 264). Lastly, the issue of applying and adapting validity theories, which are originally derived from the field of educational measurement, to language testing practice is still under-researched in practice despite the presence of basic tenets in the literature for decades. These concerns are still relevant to date. Although the importance of making a validity argument focusing on the adequacy of test score interpretations and uses in test development has been early recognized (Cronbach & Meehl, 1955; Cronbach, 1988; Messick, 1989), in practice validity has been often perceived to be a quality of tests themselves based on restricted ranges of evidence (e.g., correlations with criterion measures or content analysis). Furthermore, explicit articulation of the internal logic of validity argument itself is not present, thereby lacking an explicit methodology for generating hypotheses and formulating an intended test interpretation. Toward these challenges, researchers (e.g., Chapelle, Enright, & Jamieson, 2008, 2010) who have worked on building a validity argument for a new TOEFL demonstrate how Kane's

(2006) validity framework offers explicit guidance and conceptual framework of evaluating the adequacy of test score interpretations and uses.

Kane's (2006) Argument-based Approach to Validity

In Kane's term, the validity argument is an interpretive argument in which backing and evidence have been provided to support assumptions. In simplest terms (Kane, 2006), the argument-based framework involves two steps. Firstly, specifying the proposed interpretations and uses of the scores in some detail in the form of an *interpretive argument*. The interpretive argument involves building blocks of inferences and assumptions which are identified to support score interpretations and uses. Secondly, evaluating the overall plausibility of the proposed interpretations and uses in the form of *validity argument*, which essentially evaluates the interpretive argument focusing on its coherence, reasonability, and plausibility. Kane (2006) further underscores validating an interpretation and use of test is "to evaluate the rationale, or argument, for the claims being made and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (p. 17).

Accurate understandings of the meaning of inference in developing a validity argument are also very crucial (Mislevy, Steinberg, and Almond, 2003). In order to fulfill this purpose, Toulmin's (2003) schema for the structure of arguments has been utilized, which comprises various components, including *grounds*, *claim*, *warrant*, *backing*, and *rebuttal*. The following example of speaking performance from the work from Chapelle, Enright, and Jamieson (2008, 2010) well demonstrates how the components for the argument structure can be woven together. The observation of frequent

mispronunciations and hesitations in ESL students' speaking performance during an individual class presentation is an example of *grounds,* which become the basis for making a *claim*, a proposition we wish to support with the data. The *claim* that one can make based on the ground in this example (i.e., mispronunciation, hesitations) is relatively low student's academic speaking abilities. This process is to make an inference which proceeds by means of a *warrant*, which is a generalization used to justify the inference from the data to a claim. In other words, the evaluation of inference is supported by a *warrant*. Warrants themselves need *backing*, in the form of theories, research, data, or experience. From the example here, a warrant in this case is 'Performance on the speaking task reflects relevant language abilities needed in an academic context' and this warrant is further supported by backing, such as a generally held assumption that hesitations and mispronunciations are usual characteristics found from students with low-level speaking ability who struggle in an academic context. However, at the same time, the initial argument can be weakened by *rebuttal* data. In the example, the observation of hesitations and mispronunciations are usually signs of low levels of English speaking abilities based on the ESL teachers' experiences, but at the same time they can be results of highly technical vocabularies from an unfamiliar topic in the presentation or unexpected interruptions from the audience. This rebuttal weakens the inferential link between the ground and the claim. Such a rebuttal can include particular cases that deserve further investigations or delimit the extent to which the argument is accepted. As illustrated in this simple example of speaking performance in the form of interpretive argument, the claims, grounds, warrants, backing, and rebuttals are dependent each other and involve chains of reasoning with various data, which are also

used in this study. Lastly, it should be noted that an interpretive argument specifies the interpretation to be drawn from the grounds to a claim via an inference, rather than relying on the observation itself to evaluate the student's speaking abilities.

In order to link grounds with conclusions, multiple types of inferences in a chain of inferences or bridges, which is a metaphor referring to inferences, are required in the interpretive argument (Kane, 1992; Kane, Crooks, & Cohen, 1999). Three types of inferential bridges were identified: *evaluation*, *generalization*, and *extrapolation* in performance assessment in particular. One of the basic tenets is that each inference in the interpretive argument rests on *assumptions* that require *support*. Based on Kane *et al.*'s (1999) original discussion, Chapelle, Enright, and Jamieson (2008) detailed how these inferences can be supported based on the evidence specific to language testing and expanded to encompass a complex theoretical construct being measured in TOEFL and its comprehensive score uses. *Evaluation* refers to giving a certain score on the observation of test performance. Important assumptions behind the evaluation inference include the extent of appropriateness behind implementing scoring criteria and test administrations. Such assumptions can be investigated by empirical research on raters' performances and soundness of scoring rubrics. *Generalization* refers to whether the observed score can be expected similarly across parallel tasks and test forms. Among various methods for generalization, reliability estimates are used to support the generalization inference. Lastly, *extrapolation* is invoked when examinees' expected score is interpreted as possible performance in the target domain. One of the assumptions underlying extrapolation is that task characteristics are conducive to real tasks in the target domain and this assumption can be supported by a logical analysis of the

correspondence between tasks and the target domain. Beyond the three inferences, different types of inferences can be added to include a detailed definition of the target domain, a theoretical construct, and score use, namely *domain definition*, *explanation,* and *decision-making* respectively.

Kane's approach to a validity argument is distinct from previous approaches to validity in many ways. Recent discussions well illustrate important advancements made based on Kane's framework for L2 language assessment (e.g., Chapelle, 2012; Chapelle, Enright, & Jamieson, 2008, 2010). Kane's approach greatly emphasizes the need to state interpretations explicitly in detail, called an interpretive argument, and to evaluate the overall plausibility of the proposed interpretations and uses by formulating a validity argument (Kane, 2006). Unlike previous approaches to validity, it provides systematic disciplines, structures, and internal logic of the validity argument. Furthermore, its less strict treatment of theoretical construct definition allows more explicit focus on score interpretations and uses and can address the inherently challenging issue of dealing with different approaches to construct definitions in the field of applied linguistics.

Following the new possibilities that the Kane's framework provides for validity in language assessment, the current study will investigate the four inferences, *domain description*, *evaluation*, *generalization*, and *extrapolation*, using the Toulmin's (2003) schema for the structure of arguments in order to operationalize the complex nature of interaction-involved L2 pragmatic competence. The next section then discusses the methodological framework used in this study.

CHAPTER 3

METHODOLOGICAL FRAMEWORK

This chapter details methodological frameworks in this study. Mixed method is first discussed focusing on its characteristics and potentials to strengthen a validity argument. Secondly, a task-based language assessment framework and an approach to data-driven rating rubric development, which were used in developing pragmatic assessment tasks and rating criteria in this study are discussed.

Mixed Methods as an Independent Research Paradigm

Out of discontent with philosophical and methodological debates, which have been characterized as the 'paradigm wars' between quantitative and qualitative methodology (Gage, 1989) in social science, and recognizing the importance of the dynamic interplay between theory and practice, mixed methods started emerging in the field. Regarding its inception and position, Greene (2007, p. 32) notes "mixed methods approaches to social inquiry is a natural and logical development" and Johnson and Onwuegbuzie (2004, p. 14) refer to as "the natural complements to traditional qualitative and quantitative research". Some researchers argue that the practice of mixed methods has existed for a long time in addition to the two main research paradigms, quantitative and qualitative methods (Tashakkori, 2009). Some social scientists in applied fields, such as education, nursing, and evaluation, have routinely sought to employ various methods mainly due to practical demands from each field. Greene (2007) argues that "the practice of mixing methods in empirical studies is not a new phenomenon, especially in the highly applied domains of social science that are dedicated to understanding and improving

human practices" (p. 32). To date, the development of mixed methods theory has been generative and productive.

Contemporary mixed methods research has been strengthened and has evolved into a distinctive methodology (Greene, 2008). In an attempt to position and advance mixed methods research as an independent research paradigm, constructive discussions among the researchers in the field have been made. One of them is to examine and establish a definition and common terminologies in the field, especially considering diverse names given to refer to the mixed methods research over the years, such as blended research, integrative research, and multimethod research among others. Regarding an issue of definition of mixed methods, some researchers would argue a truly mixed approach methodology "(a) would incorporate multiple approaches in all stages of the study (i.e., problem identification, data collection, data analysis, and final inference) and (b) would include a transformation of the data and their analysis through another approach" (Tashakkori & Teddlie, 2003, p. xi). Johnson, Onwuegbuzie, and Turner (2007) undertook research to examine how the field is being defined and conceptualized by asking 19 current scholars in the field. According to thematic analysis of 19 definitions, five major themes were found: what is mixed within research paradigms, mixing stage, breadth of mixed research, why mixing is carried out, and orientation of the mixed methods research. Based on a synthesis of 19 definitions, Johnson *et al*. generated the following definition:

> Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis,

inference techniques) for the broad purposes of breath and depth of understanding and corroboration (p. 123).

Regarding the interpretation and use of methods in the mixed methods design, Johnson *et al.* keep a position of a broad interpretation allowing various issues and strategies around data collections, research methods, and philosophical issues, which is also supported by other researchers in the field (e.g., Greene, 2006).

Even among scholars in the field of mixed methods, as noted in diverse themes in the definition above, diverse views and approaches have been undertaken toward mixed methods. Creswell and Tashakkori (2007) examined how researchers adopted mixed methods and identified four different perspectives, although they are not mutually exclusive. The first is a *method* perspective in which researchers view mixed methods as strategies that are developed for data collection, analysis, and interpretations for both quantitative and qualitative methods. The second is a *methodology* perspective. In this perspective, researchers conceptualize mixed methods as a distinctive methodology beyond simply the methods. The third is a *paradigm* perspective in which researchers discuss philosophical foundations for mixed methods. The forth is the *practice* perspective. Researchers who hold this perspective consider mixed methods as means or a procedure to complete a piece of research regardless of the research design.

Despite the varying views toward mixed methods, an effort of establishing common terminologies has been made. Teddlie and Tashakkori (2006) discuss that comprehensive mixed methods research framework can be categorized into four families of designs: concurrent, sequential, conversion, and fully integrated. Firstly, *concurrent* mixed designs refer to a design in which at least two independent strands are used, for

example qualitative and quantitative. Based on an independent but simultaneous procedure of data collection and analysis from each strand, results are synthesized to make meta-inferences at the end of the study. Here, the findings from each of two strands are integrated simultaneously by comparing and contrasting to formulate a more comprehensive understanding of research issues. Secondly, *sequential* designs utilize qualitative and quantitative strands chronologically. For example, research findings and conclusions made from the first strand guide the construction of research questions, data collection and analysis for the next strand. Then, the results from both strands are used to make a final inference. Thirdly, *conversion* designs refer to a multistrand concurrent design where qualitative and quantitative strands occur in all stages of a study. In this design, data from either of the strands is collected and analyzed accordingly and then converted for further analyses using the other strand. For example, after collecting and analyzing qualitative interview data, it is transformed into quantifiable data. Lastly, similarly with conversion designs, *fully integrated* design also employs qualitative and quantitative techniques throughout all stages of conducting research. A critical difference is how it is done. For all stages, including conceptualization, methodological, analytical, and inferential stages, two methodological approaches inform each other either to formulate research questions or to influence formulation of meta-inferences after crossover analyses in an interactive, iterative, and reciprocal manner. While the distinction among the various mixed methods designs can be very helpful in guiding and designing a study, difficulties of exclusively following one of these four mixed methods designs might arise because actual studies in practice can be quite complex to follow only one design (Teddlie & Tashakkori, 2006; Greene, 2006). Nonetheless, such typologies

within the research tradition can guide a research process and help establish common terminologies for researchers.

Potential Challenges and Controversies

In conceptualizing and utilizing mixed methods design, various paradigmatic assumptions are still in debate. Such debate has emerged around the critical issue of commensurability within mixing paradigms and methods and whether such philosophical differences are reconcilable (Teddlie & Tashakkori, 2003). Especially, paradigm purists who argue each paradigm should hold its integrity without being influenced by other paradigms express a discontent toward mixed methods. Central to their arguments are assumptions that different traditional paradigms are fundamentally incommensurable about human nature and worldview since each paradigm represents a coherent whole, which must be respected and preserved (Guba & Lincoln, 1989). For example, as Lincoln and Guba (2000) asserted, certain critical theory, constructivist, and participatory paradigms are not commensurable with positivist or post-positivist traditions as they share fundamentally different assumptions and stances. Yet, regarding a question of whether paradigms are commensurable and mixed, Guba and Lincoln (2005) argue for a cautious 'yes' stating, "if the models (paradigms) share axiomatic elements that are similar, or that resonate strongly between them" (p. 174).

Regarding this concern, some mixed methods researchers take the pragmatic stance (Johnson *et al.*, 2007; Tashakkori & Teddlie, 2003) in favor of taking an action by collecting useful and practical evidence to seek answers to research questions rather than focusing on incompatibilities. Others keep a dialectic stance (Greene, Caracelli, &

Graham, 1989) by recognizing irreconcilable different assumptions behind paradigmatic stances. Taken together, considering each paradigm's distinct ontology, epistemology, and methodology, much cautious and deep understanding of each paradigm will be crucial to enhance mixed method practices. Nevertheless, Greene (2008) offers a positive view on mixed methods that can play a deep and potentially catalytic role in dealing with complex social inquiries by embracing multiple paradigmatic traditions. On this positive note, the mixed methods approach has been increasingly employed as an alternative to the two main research paradigms, quantitative and qualitative, to address complex social inquiries in diverse areas of social science, including education, evaluation research, health, nursing, and sociology (Teddlie & Tashakkori, 2003). Such effort was also made in language assessment research, which is discussed in the next section.

Mixed Methods for Validity

Within L2 language testing research, although sometimes it is not explicitly framed as the mixed methods research, various studies on investigating validity in diverse L2 educational contexts using mixed methods have been available (e.g., Grabowski, 2007; Jang, 2005; Jang, McDougall, Pollon, Herbert, & Russell, 2008; Kim, 2009; Lee, 2005; Lee & Greene, 2007; Norris, 2008; Walter, 2007). Jang (2005) undertook research on the validity of formative reading assessment to make synthesized evaluative judgments about the effects of reading skills diagnosis using mixed methods. The validity argument evaluated in Jang's study was that skills diagnosis in the context of Next Generation Test of English as a Foreign Language (NG TOEFL) can provide pedagogically useful diagnostic feedback for teachers and students. By seeking multiple

perspectives and multiple sources of evidence, the validity argument was strengthened. Lee and Greene (2007) investigated relationships between ESL students' placement test results and three measures of academic performance, GPA, faculty evaluations, and self-assessments, using qualitative and quantitative data collection methods. Ultimately, they tried to examine complex interplay when it comes to the predictive validity of an ESL placement test. The study well demonstrated how mixed methods analyses either support or complement findings from statistical analyses. Jang *et al.* (2008) employed a concurrent mixed methods approach in which thematic analysis of qualitative data and factor analysis of quantitative data were analyzed independently and then researchers synthesized the results to make meta-inferences. In the context of US college foreign language education, Norris (2008) undertook a three-year case study to evaluate assessment practices of an innovative German language program using comprehensive mixed methods in various stages of his study. Arguing for validity evaluation in educational assessment, Norris emphasized considering real constraints of a particular educational context for validity evaluation processes and outcomes. Each stage of the study was composed of multiple evaluation activities, various types of evidence, quantitative and qualitative data analyses, and various decisions made jointly on the basis of negotiated discussions among diverse stakeholders. The study showed how the implementation of validity evaluation process led to a variety of constructive outcomes and meeting intended assessment purposes. These empirical studies well illustrate how mixed methods are used to find either convergent and divergent findings, which ultimately strengthen validity arguments within various educational assessment contexts.

In addition to the effort to strengthen the validity argument using mixed methods, systematic methods for developing assessment tasks and appropriate rating criteria are also crucial to ensure the validity of performance-based language assessment. The next section discusses a task-based language assessment framework and an approach to data-driven rating criteria development, which were both used in this study to develop meaningful L2 pragmatic assessment tasks and interaction-sensitive rating criteria to measure context-specific L2 pragmatic performance.

Task-based Language Assessment

With an increasing interest of assessing learners' communicative and authentic language use, task-based performance language assessment (TBLA) (e.g., Norris, 2002; Norris, Brown, Hudson, & Yoshioka, 1998; Brown, Hudson, Norris, & Bonk, 2002) has been introduced. While traditional language assessment is concerned with the construct to be measured first, TBLA primarily focuses on whether examinees can engage in meaningful language communication and a fundamental question of "why and how task-based assessments are being used in contexts?" (Norris, 2002). The principal concern of this assessment framework is to measure learners' performance on authentic tasks that are close to real-life tasks rather than exclusively concerning with theoretical constructs to be measured.

One of the key elements in TBLA is addressing needs of diverse stakeholders, which is one of the unique contributions of TBLA compared with other assessment frameworks. Such emphasis on the needs plays an important role of not only developing assessment and pedagogical tasks (e.g., Gysen & Van Avermaet, 2005) but also

strengthening language programs (e.g., Long & Norris, 2000). When it comes to assessment uses, both formative and summative assessment is emphasized in TBLA depending on motivations of implementation of TBLA, namely intended uses of assessment. Furthermore, Norris (2009) argues TBLA is closely related to strengthening task-based language classrooms and task-based language programs by providing rich information needed to support language learning and foster learners' abilities to do things with language.

Despite TBLA's unique strengths, they are not immune from constraints and criticism. In the performance assessment literatures in general, issues of ensuring reliability and validity have been discussed mainly due to fewer tasks which also causes an inferential processes and generalization of learners' performance from fewer tasks (Brindley, 1994; Kane, Crooks, & Cohen, 1999; Moss, 1992; Shepard, 1991). Criticism toward TBLA more specifically, Bachman (2002) criticized TBLA has "problems with supporting such predictions" (p. 453) in relation to task selection, generalizability from performance tasks, and extrapolation from test tasks to target domains. While Bachman's critiques can be constructive in several accounts, they were mainly grounded from psychometric theories rather than considerations of complexity and diversity of particular language assessment settings, which is one of the key tenets of TBLA (Norris, 2002). Bachman's critiques on the conceptual framework of TBLA coincide with a tension between diverse assessment paradigms, such as those between classroom assessment and psychometric concepts (Shepard, 1991; Teasdale & Leung, 2000). Whether such tension is reconcilable or not might depend on how either construct-based or task-based assessment is used in diverse assessment settings. Considering these frameworks draw on

quite distinct theoretical and intellectual concepts, criticizing each assessment framework to become what it is not supposed to be might not be the most constructive way.

Yet, an increasing line of empirical studies on TBLA has showed its constructive and educational values in diverse L2 learning settings (e.g., Byrnes, 2002; Gysen & Van Avermaet, 2005; Norris, 2008; Youn, 2010). Norris *et al*. (1998) and Brown *et al*. (2002) embarked on a pioneering university-level second and foreign language TBLA study by developing carefully-designed prototype task-based performance tests along with task-specific and holistic rating scales. Depending on intended uses of TBLA, some studies such as Gysen and Van Avermaet (2005) developed specific target tasks and task types that can extrapolate task performance to make high-stakes decisions with summative uses in a Dutch as a second language setting. Other studies such as Byrnes (2002) and Norris (2008) employed TBLA to reform an entire collegiate foreign language curriculum by implementing tasks into the curriculum and each lesson with both pedagogical and assessment purposes. Also, such attempt has extended to a large-scale standardized assessment including TOEFL and Canadian Language Benchmarks. Accordingly, the unique characteristics of TBLA, such as an emphasis of stakeholders' needs, a focus on real-life tasks, and formative assessment, can provide constructive insights for future L2 pragmatics assessment research.

Performance Data-driven Rating Rubric Approach

The role of a rating rubric is critical in ensuring the validity of performance assessment design (Messick, 1994). Unlike a fixed-response item format, performance-based language assessment inevitably involves raters and rating criteria. As nature of

language uses in various language tasks become complex, as seen in the diverse language needs in the TBLA research above, great needs for different ways of developing rating criteria has been voiced. Two main approaches to rating rubric design have been available: the measurement-driven approach and the performance data-driven approach (Fulcher, Davidson, & Kemp, 2011). The measurement-driven approach has been the oldest and the most popular one in the field that focuses on developing the rating rubric based on a measurement model. Under this model, holistic and analytic rating rubrics have been widely employed and a debate between the rubrics has been a long-standing issue in the field of performance-based language assessment.

On the other hand, the performance data-driven approach, as an alternative to the measurement-driven approach, has been proposed which prioritizes observations of language performance (Fulcher, 1987, 1996b; Fulcher, Davidson, & Kemp, 2011; Turner, 2000; Turner & Upshur, 2002; Upshur & Turner, 1995, 1999). This approach considers performance data as important and primary source for empirical rating rubric construction process by describing them in detail or by establishing differences between levels ultimately to make a direct relationship with scale descriptors and performance data (Fulcher *et al.*, 2011). Based on the collection of language performance data elicited from target tasks, conversation analysis or discourse analysis can be conducted to identify key performance features, which become empirical basis of the resulting scale. While this approach involves an extremely time-consuming process and raters find it complex to employ, it enables measurement of a complex dimension of language competence, such as interactional competence.

Arguing for the importance of rich descriptions of performance data in rating rubrics for a better validity claim, Fulcher *et al.* (2011) proposed an innovative scoring instrument, a *Performance Decision Tree* (PDT), which is particularly useful for measuring context-specific language performance. Fulcher *et al.* illustrated an example of the development of the PDT for measuring the performance in service encounter interaction based on qualitative analysis of native speaker service encounter interaction. The PDT provides a clear picture of required competencies and skills for successful service encounter interaction, including discourse competence, discourse management, and pragmatic competence.

The performance data-driven approach takes the complexity of language discourse in particular language-use contexts into account by investigating language performance data in its own right. Therefore, it can be particularly useful to develop rating criteria that tap into interactional competence and pragmatic performance in which particular discourse and interactional features are considered important, such as L2 pragmatic performance in an EAP setting. Furthermore, qualitative analyses of language performance will allow capturing various language features empirically, that were not prioritized in the measurement-based approach to the rating criteria design. However, as the performance data-driven rating criteria include specific descriptions of target performance, it cannot be used to rate performances across other contexts other than the one for which they were developed. This will limit the practicality and generalizability of the rating criteria. Additionally, empirical questions still remain, including how raters perceive rich descriptions of each rating category and employ them to make reliable scoring decisions. Although the raters' perceptions are not directly addressed in this

study, quantitative findings on the raters' performances in this study will at least provide an answer for the inquiry that whether raters can consistently use the data-driven rating criteria with detailed descriptions.

CHAPTER 4

METHOD

Study Purpose and Research Questions

The present study investigated the validity of task-based L2 assessment (Long & Norris, 2000; Norris, 2009) with a particular focus on L2 pragmatics in *interaction* in an English for academic purposes (EAP) classroom setting. Although quantitative methods were employed to analyze the quality of the performance ratings in assessment tasks, conversation analysis (CA) was also employed to develop data-driven rating criteria to measure the interaction-involved performances in the open role-play tasks due to the social and co-constructive nature of interactional competence. Additionally, as the validity framework for this study, Kane's (2006) argument-based approach to validity was used to evaluate interpretive arguments on the proposed score interpretations and uses. The following research questions guided the study, each addressing a different type of inference (*Target Domain Description*, *Evaluation*, *Generalization*, and *Extrapolation*) sought to investigate in this study, except for research question 5, which directly addresses the mixed methods utilized in this study:

1. What kinds of role-play tasks can be developed to assess EAP L2 pragmatics in interaction? To what extent do the role-play tasks reflect stakeholders' pragmatic learning needs and measure L2 pragmatics in interaction in a valid way?
2. What kinds of interactional features can be included in analytical rating criteria to assess examinees' L2 pragmatic performances in interaction? To

what extent do conversation analysis findings of interaction data enhance the validity of the rating criteria?

3. How reliable is the rating process between and within raters' performances on rating the open role-play tasks using the data-driven interaction-sensitive rating criteria?
4. To what extent are performances on role-play tasks attributed to a construct of L2 pragmatics and L2 proficiency?
5. To what extent are findings from mixed methods trustworthy and how do they help to strengthen the validity argument?

Table 1 provides an overview of how the research questions specifically address the different types of inferences, warrants, and assumptions along with data sought to support them, as components of the validity argument. These will be further discussed as components that constitute the validity argument in Chapter 6 based on qualitative and quantitative results. The last column also describes how the data will be analyzed.

Table 1
Overview of Validity Argument and Research Questions

Inference	**Warrants supporting the inference**	**Assumptions underlying warrants**	**Research Question**	**Data (sought to support assumption)**	**Analysis**
Construct Definition	The construct of L2 pragmatics in interaction is theoretically defined.	The theory-defined construct of L2 pragmatics in interaction is measurable and operationalizable via real-life role-play tasks.	N/A	Theoretical Framework, Previous literature on the construct, Research gaps	Research Synthesis
Target Domain Description	Observations of performances in ORPTs (open role-play tasks) reveal L2 pragmatic knowledge, skills, and abilities in academic discourse.	ORPTs require important L2 pragmatic skills and are **representatives** of L2 pragmatics related to the academic domain.	1. What kinds of open role-play tasks can be developed to assess EAP L2 pragmatics in interaction? To what extent do the role-play tasks reflect stakeholders' pragmatic learning needs and measure L2 pragmatics in interaction in a valid way?	Needs analysis from Youn (2010)	Qualitative Interviews with Stakeholders, Descriptive Statistics, Factor Analysis (*already done in previous study*)
Evaluation (Task Design & Scoring) *(emphasized greatly in view of complexity of measuring the target construct)*	Observations of performances in ORPTs are evaluated to provide observed scores reflective of targeted L2 pragmatic abilities.	***ORPTs are designed*** to balance **authenticity** and **standardization**. ***Rating criteria*** is ***appropriate*** for providing **evidence** of targeted pragmatic abilities and intended uses. ***Raters*** are adequately trained with scoring the observed performances using the interaction-sensitive rating criteria.	2. What kinds of interactional features can be included in analytical rating criteria to assess examinees' performances of L2 pragmatics in interaction? To what extent do findings from conversation analysis of interaction data enhance the validity of the rating criteria?	Open role-play task design using role-play cards Interactional features of examinees' performance for developing data-driven rating criteria Rater training materials, post-rating interviews	***Sequential Mixed Design (Qualitative Phase)*** Conversation Analysis
Generalization	Observed scores are estimates of expected scores across raters (and parallel versions of tasks).	ORPTs provide stable estimates of test takers' pragmatic performances. Raters employed the rating criteria consistently across examinees.	3. How reliable is the rating process between and within raters' performances on rating the role-play tasks using the data-driven rating criteria?	Characteristics of open role-play tasks 12 raters' scoring results of 102 examinees	***Sequential Mixed Design (Quantitative Phase)*** Descriptive Statistics, FACETS
Extrapolation	Expected scores are attributed to a larger construct of L2 pragmatics and L2 proficiency.	ORPT performance relates to other measures of L2 pragmatics. ORPT performance relates to other criteria of language proficiency.	4. To what extent are performances on open role-play tasks attributed to a construct of L2 pragmatics and L2 proficiency?	Examinees' performance on pragmatic monologic tasks and TOEFL monologic tasks	Correlation btwn ORPT and monologic prag task /TOEFL spk tasks
Utilization (not directly addressed)	ORPTs and rating rubric are useful as instructional materials for EAP L2 pragmatics.	The meaning of test scores is clearly interpretable by stakeholders. ORPTs will have a positive influence on how L2 pragmatics in interaction is taught and assessed.	Not directly researched, but all decisions were made in relation to the intended uses	***Future research***	***Future research***
Mixed Method RQ: 5. To what extent are findings from mixed methods trustworthy and how do they help to strengthen the validity argument?					

Participants

Examinees. A total of 102 adult learners of English using English for academic purposes in a university setting voluntarily participated as examinees. At the time of data collection, the examinees were international students enrolled in either a 4-year university or a community college in the United States. Table 2 provides examinees' demographic information. The examinees came from 11 different first-language backgrounds: Korean (37%), Japanese (18%), Chinese (16%), Indonesian (10%), Vietnamese (10%), Farsi (4%), Spanish (2%), Thai (1%), Hindi (1%), Yoruba (1%), and Singhalese (1%). Of the 102 examinees, 70% were female and 30% were males. Almost an equal number of graduate and undergraduate students participated in this study. The length of the examinees' residence in English-speaking countries (mostly in the US) varied from one month to 22 years, with a mean of 27 months and a median of 10 months. To measure examinees' L2 language proficiency, the researcher requested examinees' standardized test scores (TOEFL, IELTS, and TOEIC). Comparison charts based on linking TOEFL iBT scores and IELTS ("Compare TOEFL® Scores," 2012) were used to compare the different scores. Examinees' converted TOEFL iBT scores ranged from 65 to 111. This proficiency range is not as wide as potential TOEFL examinees, but the language levels of the examinees were considered appropriate because they were able to at least function at an English-medium university, making examinees the target population for this study. Additionally, because of the variations between the proficiency tests, two independent monologic speaking tasks adapted from the current TOEFL iBT speaking task format, which all examinees completed, served as an additional criterion measure of examinees' L2 speaking proficiency.

Table 2
Examinees' Demographics

L1	*N*	Percentage
Korean	38	37%
Japanese	18	18%
Chinese	16	16%
Indonesian	10	10%
Vietnamese	10	10%
Farsi	4	4%
Spanish	2	2%
Thai	1	1%
Hindi	1	1%
Yoruba	1	1%
Singhalese	1	1%
Gender		
Female	71	70%
Male	31	30%
Academic Status		
Graduate	55	54%
Undergraduate	47	46%

Interlocutors for open role-play tasks. Two female and two male English native speakers who are familiar with academic discourse, as demonstrated by their experience in advising and teaching in EAP contexts, participated as professors for the role-play task. Three of these interlocutors were PhD candidates who had experience with content teaching at the university level. The other interlocutor was a MA degree-holding university ESL instructor who had also engaged in student advising. The four interlocutors received training to standardize the conversation between examinees. For example, the interlocutors were asked to accept the students' request or prefer a certain option when two choices are given. This was needed to minimize the effect of having four different interlocutors. This decision is also justified as the participants being evaluated were examinees rather the professor interlocutors. Although the examinees

were randomly assigned to each interlocutor, the examinees with a low-intermediate level English proficiency had to complete the role-play tasks with the same interlocutor due to logistic reasons.

Raters. A total of 12 individuals, consisting of three males and nine females, participated as raters who scored the examinees' audio-recorded performances of two role-play tasks and three monologic tasks. Seven raters were either English native speakers or bilingual English speakers. Five raters were non-native English speakers with advanced levels of academic English. Seven raters held MA degrees in Second Language Studies (SLS) and five raters were pursuing their MA degrees in SLS at the time of the ratings. All raters had language teaching experience, although duration and contexts varied.

Test Instruments

Interaction-involved tasks: Open role-play tasks. Two open role-play tasks were developed based on results from a large-scale needs analysis in a university-level EAP program (Youn, 2010). The needs analysis was intended to investigate students' L2 pragmatic learning needs (see Appendix A). Qualitative interviews with diverse stakeholders enabled the researcher to identify 20 different situations that students might encounter with a range of interlocutors in an EAP setting, such as *appropriately requesting a recommendation letter from a professor*, *politely refusing a professor's request*, and *disagreeing with classmates during class discussion*.

Unlike the closed role-play task format, examinees were not asked to follow predetermined interaction outcomes during the open role-play tasks, but to negotiate and

interact naturally for authentic interaction. Role-play cards (see Appendix B) for each role-play situation were employed to balance both authenticity and standardization. A different role-play card for each role-play situation was given to the interlocutors so that each interlocutor would not know what another interlocutor would say in the role-play situation. For example, in the role-play with a professor, an examinee who requests a recommendation letter from a professor with one week due does not expect to hear that the professor has a time conflict, which is a conference trip. In the role-play with a classmate, each examinee, who plays an classmate, does not know another interlocutor's class schedule or a preference of meeting mode. At the same time, the role-play cards were designed to ensure similar contents across examinees. Table 3 summarizes characteristics of the open role-play tasks.

Table 3
Summary of Open Role-play Tasks

Characteristics	**Role-play with professor**	**Role-play with classmate**
General Description	• Talk with a professor during an office visit • Three sub-situations: (1) requesting a recommendation letter; (2) requesting time to discuss a project; (3) responding to a professor's request	• Talk with a classmate about a group project meeting after class when the third member is absent • Two sub-situations: (1) decide on a good time to meet; (2) decide on whether to meet face-to-face or through online chat
Constructs being measured	• Pragmatic performance in interaction • Make requests and handle a professor's request	• Pragmatic performance in interaction • Express opinions for agreement and disagreement
Duration and Procedures	• Each examinee met with a professor interlocutor (English native speaker) • Enough time was given for an examinee to understand situations • Audio-recorded • No time limit, took about 15min for each examinee to complete	• Two examinees were randomly paired as classmates (their proficiency levels were controlled) • Enough time was given for an examinee to understand the situations • Audio-recorded • No time limit, took about 15min for each examinee to complete

Authenticity	• An examinee does not know professor's travel schedule when making a letter request • An examinee does not know that a professor will make a course rescheduling request	• Both examinees received different weekly schedules to negotiate a time naturally • Both examinees expressed opinions between two meeting options (face-to-face vs. online) based on own preferences
Standardization	• Each examinee delivered similar contents using the role-play card • Professor interlocutor was trained to follow certain scenarios (e.g., accept student's requests)	• Each examinee delivered similar contents using the role-play card • Only two meeting options were given
Role in the Validity Argument	• Provide evidence for the *evaluation* inference	• Provide evidence for the *evaluation* inference

Monologic tasks: Individual general speaking tasks and pragmatic task. In addition to the open role-play tasks, which are specifically designed to measure L2 pragmatic performance, examinees also completed three monologic tasks, comprising of two monologic speaking tasks and one pragmatic task (see Appendix C). Examinees' performances on the monologic tasks were used as additional criterion measures to examine how performance on the role-play tasks relates to the construct of L2 pragmatics and L2 proficiency. Table 4 summarizes the monologic tasks' descriptions, the constructs being measured, and their roles in building the validity argument.

Table 4
Summary of Monologic Tasks

Characteristics	**Two Monologic Speaking Tasks**	**Monologic Pragmatic Task**
Description	• Two tasks: (1) Describe a preference; (2) Describe a choice on two given options	• Giving oral comments on an e-mail written by a classmate
Constructs being measured	• Monologic speaking ability	• Monologic pragmatic performance • Ability to give constructive comments appropriately • Knowledge on writing an e-

		mail to a professor
Duration and Procedures	• Each examinee completed individually • Completed either before or after completing open role-play task with an examinee • 15 second preparation time • 45 second response time for each task	• Each examinee completed individually • Completed after two monologic speaking tasks • An examinee was asked to a researcher as if a research is a classmate • No time limit, took about 10 minutes
A role in the validity argument	• Provide evidence for the *extrapolation* inference	• Provide evidence for the *extrapolation* inference

Rating Criteria

Open role-play tasks. Following the approach to performance data-driven rating rubric development (e.g., Fulcher, Davidson, & Kemp, 2011), analytical rating criteria with five rating categories (*Contents Delivery*, *Language Use*, *Sensitivity to Situation*, *Engaging with Interaction*, *Turn Organization*) on a 3-point scale were developed. Employing a conversation analytic perspective, example performances in the role-play tasks across different levels were analyzed turn-by-turn to identify their task-specific characteristics and sequential organization (see Appendices D and E). Additionally, existing literature on EAP and L2 pragmatics were employed as evidence to enhance the validity argument. Considering the distinct contents and language performance involved in each role-play task, separate criteria for each task were developed. Detailed qualitative analyses and how such results informed the development of rating criteria are provided in Chapter 5.

Monologic tasks. The rating criteria for the two monologic speaking tasks, which were similar to the current TOEFL iBT monologic speaking task format, were adopted

from the TOEFL speaking rubric (see Appendix F). It consists of three rating categories, (a) *Delivery*, (b) *Language Use*, and (c) *Topic Development*, along with descriptions for each scale and category. The original TOEFL speaking rubric has a 4-point scale, but a 3-point scale was used due to a smaller range of examinee ability included in this study compared to the TOEFL examinee population.

The rating criteria for the monologic pragmatic task were adopted from Youn (2010) (see Appendix G). Four rating categories, (a) *Tone of giving comments/suggestions*, (b) *Delivery*, (c) *Language Use*, and (d) *Knowledge of writing an e-mail to professor*, on a 3-point scale were included in the rating criteria. Detailed descriptions and examples for each category were also included to assist the raters to make consistent and accurate decisions.

Procedures

Piloting tasks. The assessment tasks were piloted with help of four examinees with varying language proficiency levels, which resulted in the revision of the role-play task procedure. During the pilot, several sub-situations within each role-play became a source of confusion and delays during performance. For example, a role-play with a professor includes three sub-situations (making two different requests and responding to a professor's request). Originally, examinees were asked to perform all sub-situations for each role-play simultaneously in one recording. However, as different sub-situations involved different role-play cards, examinees had difficulty following several role-play cards smoothly. As this can become a source of construct-irrelevant variance, each sub-situation was performed separately with a short break between sub-situations. In other

words, each examinee was asked to complete the first sub-situation within each role-play task and then given a short break to read and understand the next role-play card before moving on to the next sub-situations.

Test administration. To complete all tasks, each examinee was required to meet twice with the researcher. The first meeting involved each examinee completing all monologic tasks and a role-play with another examinee, who played a classmate. The second meeting involved an examinee engaging in a role-play with the professor interlocutor. It took each examinee approximately 1 hour 10 minutes to complete all tasks including time for instructions. Each meeting was held with an individual examinee. Instructions for the tasks were provided in detail so that examinees clearly understood the tasks' procedures. All responses made during the tasks were audio-recorded.

In order to minimize the potential effect of an examinee's performance to another examinee's performance due to a difference in their proficiency levels especially for the role-play task with a classmate, examinees completed the role-play with another examinee who either had the same or different level of language proficiency. Based on the standardized proficiency test, examinees were assigned a proficiency level of high, mid, or low, based on their score. However, a proficiency score alone is not an adequate indicator of ability, as it does not represent one's current language proficiency, particularly if the test score was obtained more than two years ago. Thus, examinees' length of time living in the USA and their ESL program attendance were also considered in determining their language proficiency. Each examinee was randomly paired with another examinee out of the six possible pair combinations (High-High, Mid-Mid, Low-Low, High-Mid, Mid-Low, High-Low).

Rater training. Each rater participated in three separate training sessions administered by the researcher. Instead of introducing all tasks at one meeting, each training session focused on one task type and examined the complex characteristics of the task and its rating criteria. Each training session took about 50 min to 1 hour. The training and scoring order of tasks were counter-balanced, meaning some raters started scoring role-play task performance while others started scoring examinees' performances in monologic tasks. Each training session consisted of three sequential steps: (a) *familiarization* of explaining purposes of this study, assessment tasks, and various features included in the criteria; (b) *norming* to establish shared understandings of differentiating examinees' levels and pragmatic ability according to descriptions in the rating criteria; (c) *practice* of example scoring and discussion of tricky cases (see Appendix H for more details). The following aspects were also highlighted during the training: (a) establishing an understanding of what each level indicates (e.g., '1' being not a true beginner), (b) distinguishing and accurately understanding each criteria, (c) understanding interactional features (e.g., turn-takings, pauses) using CA transcripts, and (e) emphasizing rating consistency. Various training materials (see Appendix H) were prepared for the training session including a rating manual and CA transcripts to exemplify interactional features such as turn-taking and meaningful pauses. Additionally, example performances across different levels were introduced to establish a shared understandings of what each level in the criteria indicated. In particular, in order to train the raters on how to tap into interactional features, CA transcripts of several example performances were also provided. In general, raters were asked to listen to each audio file

at least twice to make consistent evaluations and the researcher was available for any consultations.

Rating. After the training sessions, each rater was given a packet that included an audio CD of examinees' responses, excel sheets of the assessment tasks, rating criteria, and training manuals. Each rater completed their ratings alone. The raters were asked to score all examinees' performances on one task first and then move on to their performances on the next task, rather than score each examinee's performance on all tasks and then move on to the next examinee. It took each rater approximately 25 hours or 35 hours to complete all ratings.

Post-rating interviews. After the rating, a post-rating interview was conducted with each rater, which took about 30 minutes to one hour. Questions on the following areas were asked: the quality of the training sessions, raters' understanding and use of the rating criteria, difficulties during rating, and suggestions for changes. The interview results were used as validity evidence in ensuring the raters' understandings of the rating criteria and quality of the rater training.

Rating Design

A fully-crossed rating design in which all of the raters evaluated all examinees on all tasks is ideal, but it is very time-consuming, expensive, and tiring, especially given the number of examinees (N =102), complex language performance, and different task types in this study. FACETS, which is the main quantitative analytic method in this study, does not require that every examinee be rated by every rater on every item. Instead, a necessary condition is to create a network of observations that connects examinees,

raters, and items so that all measures and calibrations can be placed on one common scale (Linacre, 2012; Linacre & Wright, 2002). As such, the current study employed a mixed rating design following a terminology used in Schumacker (1999, p. 325). Schumacker discusses three rating designs (crossed, nested, and mixed design) for many-facet Rasch analysis. Unlike the nested design in which each rater only rates students' performances on one task rather than on all tasks, the mixed design allows connectivity while creating a common vertical ruler in FACETS by having some raters' ratings overlapp with other raters' ratings.

Illustrating more on the mixed design, twelve raters were randomly divided into four rater groups and each rater group scored different groups of examinees rather than all of the examinees and their performances in all of the tasks. Table 5 illustrates each rater group's rating assignment. For example, Rater Group 1, composed of rater ID 1, 2, and 3, scored 67 examinees' (examinee ID 1 to ID 66) performance on two role-play tasks and 30 examinees' (examinee ID 37 to ID 66) performance on three monologic tasks. Rater Group 4, composed of rater ID 10, 11, and 12, scored 30 examinees' (examinee ID 37 to ID 66) performance on the two role-play tasks and 67 examinees' (examinee ID 37 to ID 102) performance on the monologic tasks. Dividing up the examinees between rater group allowed each examinee's performance to be rated by three different raters. In order to link subsets of the data from the different rater groups, all 12 raters scored at least 30 examinees' performances of varied proficiency levels and on all tasks, the scores of which also served as anchored data. In a Rasch measurement, measures need to be directly comparable in one frame of reference (Lincare, 2012). A gray area (performance of examinees ID 37 to 66) in Table 5 enables connectivity

between the subsets of the data. Table 6 further illustrates the structure of the rating design in the study. Each check mark (√) refers to the score given by each rater.

Table 5
Rater Groups' Rating Assignment

Examinee ID	*Two role-play tasks*	*Three monologic tasks*
1 – 36 (*N* = 36)	Rater Group 1 (Rater ID 1-3)	Rater Group 3 (Rater ID 7-9)
37 – 66 (*N* = 30)	Rater Groups 1, 2, 3, 4	Rater Groups 1, 2, 3, 4
67 – 102 (*N* = 36)	Rater Group 2 (Rater ID 4-6)	Rater Group 4 (Rater ID 10-12)

Table 6
Illustration of Rating Design

		Examinee (ID1-ID36)							Examinee (ID37-ID66)							Examinee (ID67-ID102)						
Rater Group	Task	1	2	3	..	34	35	36	37	38	39	..	64	65	66	67	68	69	..	100	101	102
1	Role-play	√	√	√	√	√	√	√	√	√	√	√	√	√	√							
(R1-3)	Monologic								√	√	√	√	√	√	√							
2	Role-play								√	√	√	√	√	√	√	√	√	√	√	√	√	√
(R4-6)	Monologic								√	√	√	√	√	√	√							
3	Role-play								√	√	√	√	√	√	√							
(R7-9)	Monologic	√	√	√	√	√	√	√	√	√	√	√	√	√	√							
4	Role-play								√	√	√	√	√	√	√							
(R10-12)	Monologic								√	√	√	√	√	√	√	√	√	√	√	√	√	√

Data Analysis

Qualitative analysis. This study employs CA in order to investigate the interactional features of examinees' L2 pragmatic performance, which informed the development of the performance data-driven interaction-sensitive rating criteria used in this study. CA's primary purpose is to explicate organizational structures of talk-in-interaction, focusing on how participants produce social actions and make sense of others' actions. A substantial body of empirical CA studies within the last four decades has shown how CA is the most productive and useful methodology in analyzing spoken

interaction. The performance of the examinees during the role-play tasks was carefully transcribed according to the notation system developed by Gail Jefferson (Atkinson & Heritage, 1984) for turn-by-turn sequential analyses (see Appendix I).

Quantitative analysis. A many-facet Rasch measurement (MFRM) (Linacre, 1989) was conducted using the computer program FACETS, version 3.0 (Linacre, 2006) to examine examinees' abilities, task difficulty, rater severity, and quality of rating criteria. MFRM is an extension of the one-parameter Rasch model. Maintaining the same mathematical properties of the Rasch model for dichotomous data, MFRM handles rating scale and partial credit observations. The model for MFRM (Linacre, 2006, p. 6) is:

$$\text{Log}\ (P_{nijk}/P_{nij(k-1)}) = B_n - D_{gi} - C_j - F_{gk}$$

where P_{nijk} is the probability of observing rating category *k* for person *n* encountering item *i* scored by rater *j* and $P_{nij(k-1)}$ is the probability of observing rating category k-1. B_n is person *n*'s ability and D_{gi} is item *i*'s difficulty in relation to rater *j*'s serverity. C_j represents the severity of rater *j*, who awards the ratings to person *n* on item *i*. F_{gk} is the difficulty of being observed in category *k* relative category *k-1*, for an item in group *g*. MFRM has been fruitfully applied in various L2 performance-based assessment studies, as it is particularly useful for understanding rater variations and statistical modeling of raters' performance well beyond what classical testing theory is capable of (Hambleton, Swaminathan, & Rogers, 1991; McNamara, 1996). More detailed information on FACETS analyses is provided in the results section.

Figure 1 summarizes the major research activities and temporal steps taken in this study.

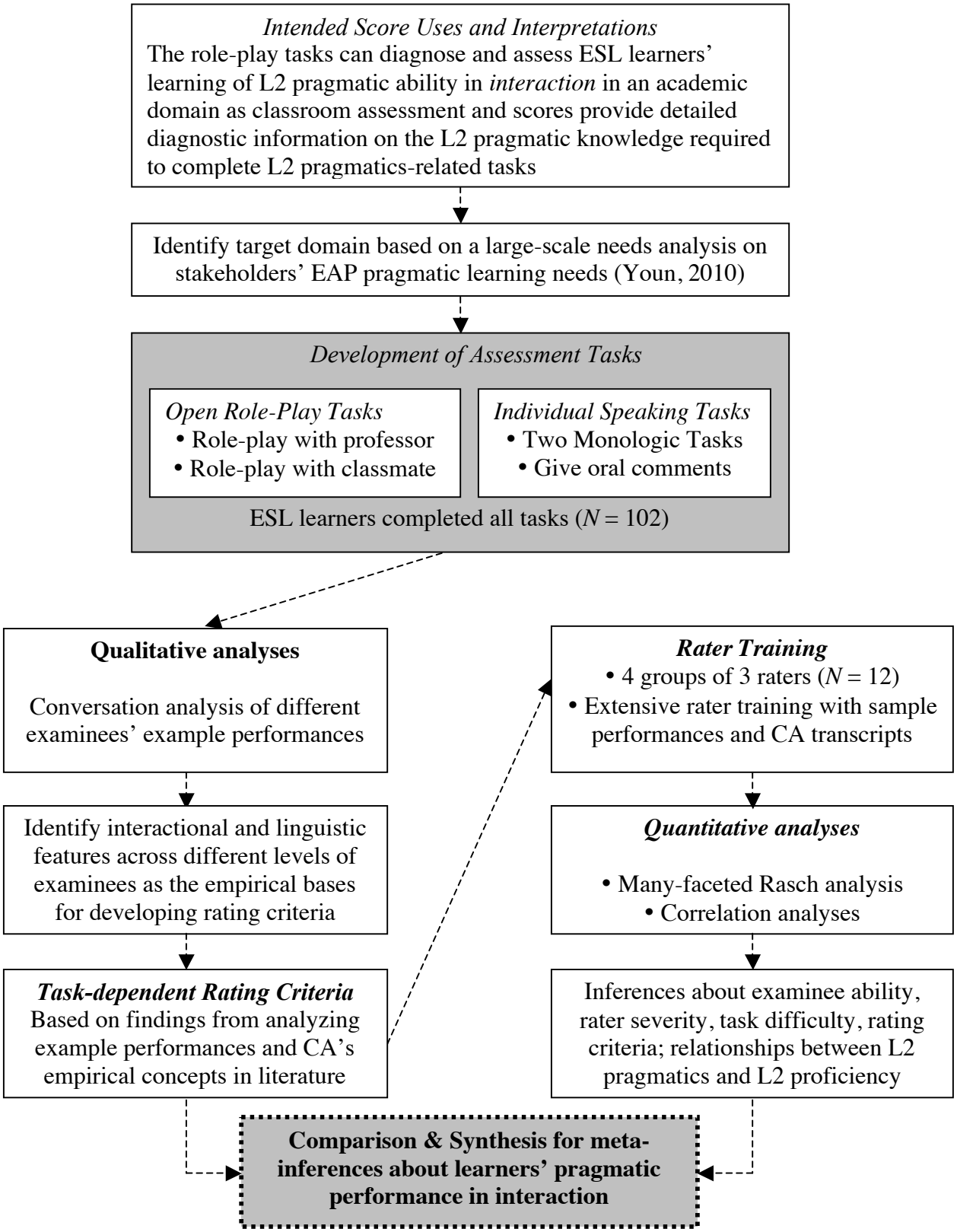


Figure 1. Flowchart of research activities

CHAPTER 5

RESULTS

This chapter first presents qualitative analyses of examinees' performances using a CA framework. Second, it discusses how this study's CA findings inform the development of rating criteria. Lastly, quantitative results from a many-facet Rasch measurement using FACETS analysis focusing on 102 examinees' abilities, 12 raters' performances, task difficulty, and rating criteria's quality are presented.

Conversation Analysis of Examinees' Example Performances

Role-play performances by examinees with various levels of language proficiency were randomly selected for CA. At the time of analyzing examinees' performances, L2 pragmatic performance level was not decided by raters, as seen in Figure 1, so examinees' L2 proficiency levels determined by TOEFL iBT and IELTS were used to select example performances across advanced, intermediate, and low-intermediate levels. This section starts with analyses of the performance in the role-play task with a professor, followed by analyses of the performance in the role-play task with a classmate. Each section presents three different levels of examinees' performances.

Role-play with Professor

The role-play task with a professor consists of three sub-situations: (a) requesting a recommendation letter with an upcoming due date, (b) requesting time to discuss a course project, and (c) responding to a professor's request to reschedule a class project presentation. For each sub-situation, a different role-play card was prepared for the professor interlocutor and examinee in order to balance authenticity and standardization

(see Appendix B) as presented in Chapter 4. All of the excerpts below involve different examinees and the four interlocutors who played the professor role.

Advanced level. Excerpt 1 presents part of an advanced level examinee's (ID 26) performance in the role-play with a professor. The examinee is playing the part of a student meeting with a professor to make a recommendation letter request. "Jessie" was the name used for all examinees for the role-play task with a professor. As presented in Chapter 4, the examinee does not know the professor's schedule constraints in advance.

Excerpt 1: Advanced Level Examinee
Make a recommendation letter to professor
P: Professor (Interlocutor 1), J: Jessie (Examinee ID26)

```
P: what can I do for you today
J: uhm, I actually have uh:: little of a big favor
   for you?=
P: =°uh huh°
J: uhm: I'm (.) applying for this department
   scholarship. and I need a letter of recommendation.
   and I was wondering if you are uhm able to write one
   for me.
   (0.7)
P: sure, I'd be happy to write the letter for you,
   uh when do you need it by?
J: uhm: it's actually the letter is due in one week
P: oh that's quite soon=
J: =Uh hum=
P: =U[hm: one problem with that would be that I'm going
J:   [°I'm sorry°
P: on a conferenc[e. I'm tryn' to leave tomorrow=
J:               [okay
J: =okay
P: and I will be gone for a week=
J: =uh huh
P: so I don't know if I will have time to do it=
J: =okay
P: would it be possible to submit the letter later?
J: Uhm::: (2.0) actually uh let me check with my department
   about the deadline for receipt? uhm: But you can also
   uhm beside the hardcopy you can also submit the letter
   by electronic copy. So, if- you might if- if you can
    >you might be able to do it< or while you are away.
   but I know you will be busy. but, uhm:
   (1.5)
J:  >what what what< do you still want me check with my
   department about °the writing°?
P: ah Yes, please do chec[k with the department.
J:                       [okay okay
```

In line 1, the professor initiates the conversation with *what can I do for you today,* displaying the professor's orientation toward an institutional setting, namely an office hour visit where a student visits a professor to accomplish institutionally-related actions. In lines 2 to 3, Jessie projects a pre-request sequence (Sacks, 1992) (*I actually have a little of a big favor*), which commonly occurs in the previous turn of an actual request, and Jessie downgrades her request using *a little of*. Here, Jessie's turn displays an understanding of the professor's previous turn as a grant to bring up her agenda. The professor's response, *uh huh,* in line 4 serves as a preferred go-ahead response, advancing the sequence's trajectory. In lines 5 to 8, Jessie projects a request sequence using diverse grammatical structures, including a present progressive (*I'm applying*), past progressive (*was wondering*), and complex structure using a dependent clause (*if you are able to*). The professor clearly understands Jessie's utterance as a request and grants it in line 10, seeking further information about the letter request (*deadline*) in line 11. In response to the professor's question, Jessie delivers a response about the short letter due date with signs of mitigations, such as an elongated *uhm* and an adverb (*actually)* in line 12. By doing so, Jessie treats the short notice as dispreferred, transforming a routine request to a problematic request. Here, the professor treats Jessie's response as unexpected information, indexed by the change-of-state token *oh* (Heritage, 1984) in line 13. Orienting to the imposition of the request, Jessie projects *I'm sorry* with a meek voice in line 16. The professor then brings up the conference travel schedule, which might delay the letter submission, in lines 15 to 20. In lines 22 to 24, the professor asks if the letter can be submitted later. Jessie treats the professor's request as a dispreferred action by deploying an elongated *uhm*, a 2.0 second pause, and the adverb *actually* in line 25, soon

issuing a response (*checking with the department*) and an alternative option (*an electronic letter submission option*) in lines 25 to 30. As this exchange illustrates, Jessie displays her sensitivity to the contingencies of the request and the professor's responses, which is achieved by utilizing diverse linguistic structures, institutional knowledge, and speech production devices such as prosody and voice quality. After Jessie's long turn and with no immediate response from the professor, Jessie launches another turn by asking the same question in a yes-or-no form in lines 32 to 33. The professor's response in line 34 is marked with the token *ah* and *yes* in an emphatic tone. Being different from the frequently used the state-of-token *oh* in line 13, the oh-prefaced response to Jessie's question in line 34 is a practice used when the question evidently requests information that is already available by inference (Heritage, 1988). The pause in line 31 shows that the professor treats Jessie's long turn as pragmatically problematic.

Summarizing Jessie's (examinee ID26) performance, firstly, the utterances were delivered accurately and smoothly within turns. Additionally, diverse linguistic resources were utilized to deliver pragmatic meanings. The examinee engaged in the conversation with the professor smoothly, as evidenced by her acknowledgement tokens and relevance in Jessie's turns in relation to what the professor said. More importantly, the examinee showed great sensitivity toward turn-taking, such as completing adjacency pairs (i.e., *question & answer*) and using meaningful pauses as a mitigation device for dispreferred action to respond to the professor's inquiry on whether the letter can be submitted after the deadline.

The various characteristics discussed above are also shown in another advanced-level examinee's (ID48) performance in Excerpt 2. However, examinee ID 48 formulates

the contingency before the professor grants a request. Unlike examinee ID26 in Excerpt 1, examinee ID48 completed the role-play tasks with another interlocutor (ID2).

Excerpt 2: Advanced Level Examinee
Make a recommendation letter to professor
P: Professor (Interlocutor 2), J: Jessie (Examinee ID48)

```
12 J: uhm I (.) was wondering if you could help me=
13 P: =uh huh
14 J: I need a recommendation letter↑ (.) to apply for a
15    scholarship=
16 P: =uh huh
17 J: ah (0.5) and: >I know it is kind of late cause due<
18    in one week so:=
19 P: =due [in one wee[k
20 J:      [I'm sorry [yeah
21 P: [one week
21 J: [>wondered you could help me<
22 P: okay Uh:m well (1.0) I have a conference that I'm going
23    to that I'm leaving out for tonight and I won't be back
24    until the end of the next week=
25 J: =°oh wow°
26 P: would it be possible for me to turn in the: a:
27    recommendation a little bit late
28 J: uhm (1.0) well (.)  if- if you have to go to a conference
29    and you won't have time then there's nothing that
30    we can do about it, Yes I can check with the department
31    office
32 P: yes please do
33 J: yeah >and and< let you know as soon as possible
34 P: okay uhm (.) also would it be possible to submit it
35    electronically
36    (0.5)
36 J: uhm Yes I think (.) that should be okay but you know
37    to make sure I will ask the department=
38 P: =uh huh=
39 J: =for the information and then I will e-mail you and or
40    I'll give you a call
41 P: yeah let me know >as soon as you can<
```

In line 12, Jessie initiates a request sequence using bi-clausal structure (*I was wondering if*) with a modal verb (*could*) in a general statement without specific information. Followed by the professor's acknowledgement (*uh huh*) in line 13, Jessie explicitly states what is needed along with a brief account in lines 14 and 15. Jessie continues the turn by providing the professor with information on the recommendation letter's deadline, which is in one week, in lines 17 and 18. Notably, this information is delivered with a hedge (*ah*), delay (0.5 second pause), and assessment of the due date

appropriateness (*late*), displaying Jessie's orientation toward treating the letter's one week due date as dispreferred and problematic. Additionally, Jessie formulates the contingency by mentioning the short letter due before the professor grants the request. In line 19, the professor repeats the content words (*due in one week*), overlapped with Jessie's explicit apology turn in line 20. Accompanied with hedge (*uhm, well*) and a 1 second pause in line 22, the professor explains the conference schedule. This information is new to Jessie as evidenced by the change-of-state token *oh* in line 25. Then, the professor asks whether the letter can be submitted late in lines 26 and 27. Marked with hedging (*uhm, well*) and delay (1.0 second pause), Jessie's response in line 28 provides her own assessment on the situation and an answer to the professor's earlier question. The professor provides the receipt (*okay*) and asks whether an electronic submission is possible. Jessie issues a personal assessment of the feasibility of this new information using a complex structure (*I think that should be okay*) and then displays her orientation toward the institutional setting by mentioning permission from the department in lines 36 to 40. The professor marks Jessie's response with the information action receipt *yeah* and then responds to Jessie's remarks with the expression *as soon as you can* in line 41, orienting to the professor's authority to give permission to a student.

The following features characterize the performances of two examinees with advanced-level L2 proficiency. Concerning the performances' linguistic features, diverse grammatical structures were utilized, such as bi-clausal structures (*I was wondering if*), conditional phrase, and various modal verbs (*might, can, could*), rather than relying on one expression to make the letter request and express politeness. Additionally, the examinees not only revealed situational knowledge specific to an academic setting (e.g.,

letter due date) but also utilized this knowledge to smoothly engage in a conversation with the professor (e.g., suggesting the electronic letter submission option to meet the deadline). When it comes to the interactional features of examinee ID26 and ID48's performances, the examinees smoothly engaged in their conversations, displaying their understandings of the professor's utterances in their sequences' subsequent turns and through their lack of noticeable delays between and within turns. At the same time, the examinees also made meaningful turns, such as completing adjacency pairs and displaying sensitivity to dispreferred actions via employing delays and hedges. Let us then move to performances of intermediate level examinees in the same role-play tasks in order to investigate whether similar characteristics are also present.

Intermediate level. Excerpt 3 presents the performance of an examinee (ID72) with intermediate level L2 proficiency engaging in the letter request role-play task with a different interlocutor (Interlocutor ID3). Again, Jessie, an examinee, visits a professor's office to make a recommendation request.

```
Excerpt 3: Intermediate Level Examinee
Make a recommendation letter to professor
P: Professor (Interlocutor 3), J: Jessie (Examinee ID72)
1  P: hello:?
2  J: hello do you have time?
3  P: Yes, Jessie please come in sit down (.) what can I do
4     for you?
5  J: ah: I- (.) I plan to apply (.) the international student
6     scholarship and I need a letter (.) recommendation letter.
7     so could you write me >could you write< it for me?
8  P: Yes I'd be happy to=
9  J: =°uh huh°
10    (2.5)
11 P: can I ask uh when the letter is due?
12 J: ah: it's due in one week
13 P: °hmm°
14    (2.0)
15 P: ah: xx the problem with that next week I'm going to a
16     COnferenc[e
17 J:           [°conference°
18 P: and: I leave for the conference tonight=
19 J: =hm:
20 P: I will try to write the lette[r while I'm awa[y
```

```
21 J:                                [°hm            [°uh huh°
22 P: hm but let me ask you can the letter be submitted
23    after the deadline? Is that possible?
24    (0.7)
25 J: I don't think so
26 P: hmm:
27    (2.0)
28 P: okay well uh::: in that case (.) let me I will TRY and
29    write it (.) while I'm away, but maybe could you CHEck to
30    see if there's an[y flexibility [with the deadline
31 J:                  [okay          [uh huh
32 J: okay I will check (.) the department office, okay
```

After greeting the professor, Jessie projects *do you have time,* a pre-sequence that establishes recipient's availability for talk (Sacks, 1992) in line 2. In lines 3 to 4, the professor launches a "go-ahead" type response, which is a preferred action that promotes the progression of the sequence (Schegloff, 2007). Understanding the professor's turn as a grant to bring up an agenda, Jessie makes a recommendation letter request in lines 5 to 7, accompanied by a brief account of the request (*plan to use the letter to apply for a scholarship*). Note Jessie's language use in making the request: although Jessie uses the modal verb *could* in line 7, her request is made using the mono-clausal phrase *I need* in line 6, despite the relatively high degree of imposition (i.e., recommendation letter request to professor). In line 8, the professor grants Jessie's request with no pauses, which is a common feature of preferred actions, such as granting a request (Pomerantz, 1984; Pomerantz & Heritage, 2013). After the professor's granting turn in line 8, Jessie provides the minimal acknowledgement token *uh huh* in line 9, displaying her understanding of the previous turn. What is noticeable during this sequence is that after the first pair part (i.e., granting a request), the second pair part (i.e., thanking) of an adjacency pair (*granting a request–thanking*) is missing. A 2.5 second pause follows in line 10, which indicates that the professor treats the absence of the second pair part as accountable.

After the pause, since no further response is given by Jessie, the professor launches into the different topic of the letter deadline in line 11. Marked with the state-of-change token *ah*, Jessie displays her understanding of the professor's question by elaborating the letter's due date in line 12, which also shows no recognition of the short letter due date imposition. Initiated by *hmm* and a 2.0 second pause in lines 13 and 14, the professor explains his time constraints in lines 15 to 20. Jessie expresses her understandings of the professor's schedule by repeating the content word *conference* (line 17) and projecting acknowledgement tokens (lines 19 and 21). The professor offers a resolution asking Jessie whether the letter can be submitted after the deadline and treats this as a request using the modal verb *can* and a mitigation device (*is this possible*) in lines 22 and 23. Interestingly, in response to the professor's request, Jessie says *I don't think so* (line 25) after a 0.7 second pause. The way that Jessie treats the professor's resolution clearly indicates her absence of pragmatic competence and institutional knowledge. Because rejection a request is a dispreferred action, mitigation or delay with meaningful pauses is normally expected. However, Jessie does not show any signs of treating the rejection as a dispreferred action, except for a short 0.7 second pause in line 24. After a long pause in line 27, which indicates that the professor treats Jessie's response as accountable, the professor continues the conversation by providing a solution to the situation.

Summarizing the examinee's (ID72) performance in Excerpt 3, diverse language use was not observed in her request, in contrast to the advanced-level examinees' performances. Although the examinee is engaged in the interaction with the professor, as evidenced by the use of acknowledgement tokens and the relevance of the examinees'

turns, the examinee did not complete the second pair of an adjacency pair (i.e. *thank to granting a request*), which is quite noticeable and which the professor treats as interactionally accountable. Additionally, a lack of meaningful pauses in handling the dispreferred action was observed.

Let us examine how an intermediate level examinee performs on another situation in the role-play with a professor. Excerpt 4 presents part of the performance of examinee ID9, who is participating in the role-play situation that requires an examinee to respond to a professor's request to reschedule a class presentation. As explained in Chapter 4, the content of the professor's request was not previously given to the examine in order to elicit how the examinee handles the situation. Additionally, the role-play card does not explicitly direct an examinee to refuse the professor's request, and instead, instructs an examinee to *respond to professor's question* so that examinees can choose the situation's interactional outcome (see Appendix B).

Excerpt 4: Intermediate Level Examinee
Respond to professor's request
P: Professor (Interlocutor 1), J: Jessie (examinee ID9)

```
1  P: you have your uh presentation scheduled for uh (.)
2     two weeks from now is that right?=
3  J: =that's right
4     (1.0)
5  P: I was wondering if you'd be able to do your presentation
6     a week early to do it next wee[k uhm
7  J:                               [oh >can I ask why?<
8  P: because the student who's supposed to present
9     next wee[k is sick she won't be here at all so would you
10 J:         [°uh huh°
11    be willing to go earl[y?
12 J:                      [that's sad but I cannot becau:se
13    I have uh exam next wee:k so (.) I cannot tsh (1.5) it-
14    next time I'm too busy to do presentation
15 P: okay that's fine I will ask somebody else then
16 J: okay °sorry°
17 P: okay
```

In lines 1 to 2, the professor asks Jessie to confirm the date of her class presentation. Jessie responds by confirming the date of the presentation in line 3. After a

1 second pause, the professor makes a scheduling request (lines 5 and 6) using the complex structure *I was wondering if*. By employing these various resources, the professor orients to the contingency of the request. Overlapping with the professor's turn, Jessie initiates a question with the change-of-state-token *oh* (line 7), which she produces more quickly than any of her other utterances in the sequence. Responding to the Jessie's request for further explanations of the situation, the professor provides an account of the earlier request (*health condition of an original presenter*) in lines 8 to 9. The professor then reformulates the request more directly using a mono-clausal in lines 9 to 11. As a response to the professor's second attempt to issue her request, Jessie refuses the request.

Thus far, it is important to note the way in which Jessie's refusal was accomplished in this sequential organization. In lines 12 to 14, the refusal sequence was made with an overlap, an assessment (*that's sad*), and a mono-clausal phrase (*I cannot*), followed by a brief account (*midterm schedule, being too busy*). Refusal is a dispreferred action involving issues of disaffiliation, noticeability, accountability and sanctionability (Boyle, 2000). As well documented in Pomerantz (1984), dispreferred actions, such as rejections or refusals, are normally delivered with hesitation, delay, or accounts at the start of the response. However, here, the examinee's way of delivering refusal is almost the opposite of what is normally expected in social action. Additionally, regarding the linguistic features involved in making a refusal, the examinee's performance lacks diverse linguistic expressions (e.g., complex structures, conditionals). Neither the examinee's tone nor content (i.e., minimal account) in the refusal delivery shows her sensitivity toward the situation. Concerning the interactional features, the examinee was able to engage in the interaction. However, the existence of abrupt turn-initiations (e.g.,

line 7) and the way that the refusal was performed without a hesitation or delay indicates that the examinee lacks sensitivity toward making proper turns in interaction.

Low-intermediate level. Moving on to the performances of examinees with low-intermediate language proficiency, the following excerpts depict part of their role-play performances and the distinct features therein. Compared to the higher-level examinees, low-intermediate language level examinees' linguistic expressions and sequential competence were quite limited, especially in delivering pragmatic meanings, and also displayed a lack of situational knowledge. However, different resources and strategies were utilized in delivering their pragmatic meanings. Excerpt 5 illustrates part of the performance from examinee ID90 with the letter request task and Interlocutor ID4.

Excerpt 5: Low-intermediate Level Examinee
Make a recommendation letter to professor
P: Professor (Interlocutor 4), J: Jessie (examinee ID90)

```
1  P: Hi Jessie come on in
2  J: Hi
3     (1.0)
4  J: so: (0.5) I nee:d (.) uh recommendation letter for
5     (.) scholarship=
6  P: =oh↑ okay=
7  J: =so:=
8  P: =what kind of scholarship?
9     (1.0)
10 J: department scholarship
11 P: okay you are applying for department scholarship?=
12 J: =uh huh >how can I do<?
13 P: uh huh (.) oh↑ are you asking me [for a letter
14 J:                                  [>yeah yeah< I have
15    a question so:
16 P: sure (.) yeah I can write a letter for you for
17    [the scholarship=
18 J: [uh huh
19 J: =okay
20 P: uh: when is it due?
```

After greeting the professor and an ensuing 1 second pause (lines 1 to 3), Jessie launches into a statement of what is needed without pre-sequence, which is highly non-normative. Her turn in lines 4 to 5 includes a prosodic emphasis on the pronoun *I* along

with several delays within the turn. Marked with the state-of-change token *oh*, the professor produces the token *okay* in line 6 granting the request, and asks a clarification question. Consequently, Jessie explains the nature of the scholarship in line 10 after a 1 second pause, and the professor reconfirms this information by recycling what Jessie had told her in line 11. After issuing the acknowledgement token *uh huh* in line 12, Jessie says *how can I do*. However, rather than elaborate her request further, Jessie then asks the professor to tell her how to apply for the scholarship. Understanding Jessie's question as a request, the professor issues the state-of-change token *oh* and projects a clarification question in line 13. Jessie's next turn in line 14 confirms the mutual understandings established between them. Although Jessie attempts to achieve the interactional action (request), her interactional competence and linguistic resources in making a request with high imposition are very limited. Additionally, Jessie's *how can I do* turn displays the examinee's orientation toward seeking help from the professor. By doing so, the examinee changes the situation from making a request to seeking the professor's help. After Jessie's indirect request is granted, Jessie projects the generic acknowledgement tokens *uh huh* and *okay* in lines 18 to19, instead of providing the normally expected second part of the adjacency pair, which in this case should be *thanking*. After this sequence, the professor launches into a topic of the letter due date, which is a critical condition for granting the letter request.

Summarizing the performance in Excerpt 5, examinee ID90 was able to carry on the task conversation with the professor. However, the examinee's utterances were minimal, lacking diverse linguistic resources when making her request. The examinee primarily employed a statement of her needs without providing enough information. In

terms of the interactional features of the conversation, although the examinee completed adjacency pairs, the examinee's second pair part was not necessarily appropriate and her sequential competence is quite limited compared to those of the previous examinees.

Examinee ID88, another low-intermediate language level examinee also demonstrates her inability to complete adjacency pairs and achieve mutual understanding in Excerpt 6, which presents part of ID88's performance in asking if a professor has the time to discuss a course project. For this situation, the examinee does not expect that the professor will provide two options for a meeting time.

Excerpt 6: Low-intermediate Level Examinee
Request more time to discuss a course project
P: Professor (Interlocutor 4), J: Jessie (examinee ID88)

```
1  J: professor uh: (1.0) uh: do you hh do you have a time?
2     (1.0)
3  P: uh huh do I have time now?
4  J: uh Yes uh I have (.) a (.) few questions about (.) my
5     course (.) project
6  P: ah↑ I see uhm I see >you wanna talk about< your course
7     project, I actually I don't have time right now,
8     to discuss it uh: (.) so maybe we can meet another
9     time? like tomorrow after class? (1.0) or even next
10    week at this time?
11 J: °okay°
12 P: uh huh↑ when would you like to meet?
13 J: uh: (0.7) uh: I can't (.) decide uh: (.) proper topic
14 P: Ah↑ that's what you want to talk about? yeah you
15    can't decide a proper topic for your class project=
16 J: =°yes°=
17 P: =yeah, uh but when do you want to meet? we could
18    meet tomorrow after class? or next week (.) at this
19    time
20 J: tomorrow is (.) good for me
```

In line 1, Jessie projects a request pre-sequence (Sacks, 1992), *do you have a time*. But, the indefinite article obscures the interactional import of the turn, as marked from a 1 second pause and the professor's confirmation question in line 3. But, Jessie does not understand the professor's action as confirmation request as she does not wait for the professor's response to her confirmation (yes) in line 4. Jessie immediately projects a pre-

request (*I have a few questions*) in lines 4 and 5, which the professor understands as a request action to discuss Jessie's course project. By providing an account (time constraint) and a proposal (a time option) in lines 6 to 10, the professor treats her action as accountable. Jessie only acknowledges the professor's response in line 11 with the token *okay* in a quieter voice without providing an answer to the professor's disjunctive proposal. Without an immediate answer to the time option, the professor soon issues another question *when would you like to meet*, which is specifically designed to seek an answer for the time option question she produced in line 12. However, in response to the professor's question, Jessie elaborates why she wants to meet with the professor (*can't decide a proper topic*) in line 13, which might be a delayed account of the Jessie's time request at the beginning. But, Jessie's turn in line 13 is not conditionally related to the professor's question and clearly displays her misunderstanding of the professor's previous utterances (lines 7 to 10, line 12). Marked with the state-of-change-token *ah*, the professor displays her understanding and repeats Jessie's delayed account in lines 14 and 15. By doing so, the professor shows her revised understanding, but with uncertainty by seeking confirmation of that understanding, followed by Jessie's confirmation token *yes* in line 16. The professor then launches the same question in order to seek an answer, repeating the options for meeting time in lines 17 to 19. Finally, Jessie provides an answer to the professor's question in line 20. To summarize examinee ID88's performance, the utterances delivered were minimal and a bit choppy, based on frequent delays and pauses within the examinee's turns. Concerning the examinee's interactional competence, the examinee at times did not provide a relevant answer to the professor,

displaying her lack of understanding in the conversation. The examinee repeatedly did not understand the professor's actions.

Excerpt 7 demonstrates part of another examinee (ID97)'s performance on the same task situation in excerpt 6. This time, the low-intermediate level examinee ID97 accepts the professor request. Again, the examinee's role-play card does not say to refuse, but allows the examinee to choose interactional outcomes.

Excerpt 7: Low-intermediate Level Examinee
Respond to professor's request
P: Professor (Interlocutor 4), J: Jessie (examineeID97)

```
1  P: I was wondering if you could do your presentation
2     a week earlier, because the student who was going
3     to present next week has gotten really sick, so they
4     can't do the presentation, could you do it next week?
5     (1.3)
6  J: uh:: (0.5) I:'m so: sorry uh I will have a midterm (.)
7     exam next friday in history class=
8  P: =I: see=
9  J: =uh
10    (2.0)
11 J: but↑ (1.3) hmm:
12    (4.0)
13 J: I will (.) try=
14 P: =uh huh=
15 J: =uh
16    (2.0)
17 J: becau:se until no[w(.) uh (1.0) I have studied=
18 P:                  [uh huh
19 P: =Oh↑ okay
20 J: uh huh (0.3) little bit
21 P: yeah
22 J: so:=
23 P: =so you think you [prepared
24 J:                   [it-(.) it doesn't (.) big matter
25 P: uh huh okay it's not [a big deal?=
26 J:                      [I'll try
27 J: =yes
28 P: okay
```

The professor engages in a long turn in lines 1 to 4, making a presentation reschedule request to Jessie. After a 1.3 second pause (line 5), Jessie explicitly projects an apology sequence along with an account of her upcoming exam schedule in lines 6 to 7. In Jessie's utterance *I'm so sorry*, the adverb *so* is elongated, creating an apologetic

tone. Followed by the professor's acknowledgement in line 8, Jessie attempts to initiate a turn after a 2-second pause in line 10. After producing the topic-transition marker *but* in line 11 and following long pauses in lines 11 and 12, Jessie launches the statement that *I will try*, providing promises to make an effort to comply. Compared to earlier sequences in this excerpt, more overlapping turns start to appear indicating their actions become more preferred. After another pause in line 16, Jessie provides an account of her action in line 17, followed with the professor's change-of-stake token *oh* in line 19. Soon, Jessie gives an assessment of her readiness as *little bit* in line 20, which is upgraded by the professor in line 23 as being prepared. Jessie then gives another assessment *it doesn't big matter* in line 24 by blending two formulaic expressions *it's not a big deal* and *it doesn't matter*. Despite the incorrect use, this shows Jessie's sequential competence, as formulaic expressions are often used in sequence-closing position (Drew & Holt, 1998). Line 24 is followed by the professor's turn (*it's not a big deal*) in line 25, which repairs Jessie's earlier turn. Overlapping with the professor's previous turn in line 25, Jessie projects another turn *I will try* in line 24.

Based on frequent delays within and between turns, it is evident that examinee ID97 lacked clear delivery. The minimal and incorrect grammatical expressions (e.g., *it doesn't big matter* in line 22) also indicate the examinee's low language proficiency. Interestingly, despite the lack of linguistic features that are normally present in other examinees with higher language proficiency, examinee ID97 still achieves the task's interactional goal without being impolite or inappropriate. As a way to achieve the interactional goal of the task, examinee ID97 oriented to being a good student with A good work ethic by accepting the professor's request and by assessing acceptance of the

request as not a big deal. Additionally, examinee ID97 handled the dispreferred action using a pause and hedge, as demonstrated by examinee ID97's initial apology. This aspect should not be taken for granted. As seen in Excerpt 4, intermediate-level examinee ID9, whose language proficiency is higher than examinee ID97, did not necessarily display sensitivity toward handling the dispreferred action.

Role-play with Classmate

The role-play task with a classmate consists of two sub-situations: (a) deciding a good time to meet with classmates, and (b) deciding whether to meet face-to-face and or by online chat. In order to facilitate authentic interaction, each examinee received a different weekly schedule and was allowed to suggest a meeting mode according to their personal preferences (see Appendix B). As explained in Chapter 4, the examinees were either paired up with another examinee who either had the same or different level of language proficiency. The following three excerpts present two performances. This first performance is between an advanced and an intermediate level examinee and the other is performance between two low-intermediate level examinees.

Interaction between an advanced and an intermediate level examinee. Excerpt 8 presents part of the performance of an advanced level examinee (ID67) and an intermediate level examinee (ID68) in the role-play task of deciding when the examinees should meet to discuss their group class project. Additionally, the task instruction also explains that the third group member, Tom, is absent. Different weekly schedules were given to each examinee (see Appendix B), and each examinee did not know the other examinee's schedule. Jessie and Phoenix were the names given for examinees ID67 and

ID68, respectively. Phoenix is supposed to start the conversation by proposing available time slots to Jessie.

Excerpt 8: Advanced and Intermediate Level Examinees Decide a time for a class project with a classmate
J: Jessie (Advanced, examinee ID67)
P: Phoenix (Intermediate, examinee ID68)

```
3  P: uh: (1.7) you remember↑ we have uh: (1.0) presentation on=
4  J: =yes of course I remember
5  P: okay uh: I think when we have first time to meet together
6     I think we can with the (1.0) uh: I have schedules (.) but=
7  J: =okay tell me when you are free (.) uhm:
8  P: I'm free on Tuesday how about yo[u
9  J:                                 [on Tuesday=
10 P: =yeah
11 J: what time on Tuesday
12 P: I think available anytime on Tuesday
13 J: okay, I'm I have to work until five pm, but maybe
14    we can meet later↑ when I get back from work?
15    (0.7)
16 P: °oh°
17    (1.0)
18 P: hm:: (1.0) how about the:
19    (1.0)
20 J: does- it does work for you? right?= ((laugh))
21 P: =yeah I think it's [xx
22 J:                    [oh why not?
23 P: I think how about the (1.0) uh: Saturday because it's
24    weekend
25    (0.7)
26 J: yeah sure Saturday is fi[ne
27 P:                         [saturday
28 J: I'm free all Saturday
29 P: but, I have time until five pm
30 J: I'm free all Saturday so I can work=
31 P: =oh are you you're free on °saturday°=
32 J: =yeah=
33 P: =we can okay we can meet uh from uh morning until
34 J: okay we can meet early in the morning?=
35 P: =uh huh=
36 J: =we can[xxxxx
37 P:        [but the: how about Tom?
38    (1.0)
39 J: uh:: do you have his number or his e-mail?
```

In lines 3 to 4, Phoenix brings up their group presentation topic, which Jessie's response overlaps, before Phoenix completes his utterance. After establishing shared understanding, Phoenix suggests scheduling a meeting in lines 5 to 6. However, Phoenix's utterance consists of chunks of words that lack coherence. Nonetheless, in line

7, Jessie shows an understanding of what Phoenix suggests. In lines 8 to 14, Jessie and Phoenix exchange their time availabilities on Tuesday, a day they both seem to be available. First, they agree on the day and move on to the specific time.

But in line 13, Jessie expresses that she may have a conflict that day, which is delivered with an account (*work schedule*), mitigation (*maybe*), modal verb (*can*), and an alternative (*meeting after work*). Phoenix' response to Jessie's suggestion is considerably delayed (lines 15 to 18), accompanied by the change-of-state-token *oh*, pauses, mitigation, and an utterance used for a suggestion (*how about*), all of which commonly appear for disagreement. After no immediate uptake from Phoenix in line 19, Jessie understands the delays in Phoenix's turn as signs of disagreement for meeting late Tuesday, and asks a confirmation question with laughter (lines 20 and 22). Ignoring the Phoenix' dispreference display in lines 18 to 19, Jessie shows expected agreement as evidenced in lines 20 and 22. Phoenix finally suggests a different day, Saturday, along with an account (*weekend*) in lines 23 and 24. Phoenix's suggestion is a bit abrupt and he does not respond to Jessie's previous turn nor provide an account of his disagreement, showing his limited sequential competence. After a short pause in line 25, Jessie accepts the new time and their exchanges flow from lines 28 to 35, with no delay between turns. But, Phoenix' lack of sequential competence is also shown in lines 28 to 31, such as Phoenix' delayed confirmation of Jessie's availability on Saturday in line 31 even after Jessie repeats it twice in lines 28 and 30. Eventually, they agree with meeting early on Saturday (e.g., line 30, 34). While in the middle of deciding what time to meet on Saturday, Phoenix suddenly cuts off Jessie's utterance in line 37 and brings up the

group's third member, Tom, who is absent during the conversation. After a short pause, Jessie responds to the question by referring to Tom's contact information.

To summarize the two examinees' performances, concerning the ways in which disagreements or agreements were performed, examinee ID67 (Jessie) provided an account (*work on Tuesday*), utilized linguistic resources, and gave clear content delivery. However, examinee ID68 (Phoenix) neither provided an account for the disagreement nor employed diverse linguistic resources. Also, the opinions delivered by examinee ID68 were not coherent or clear (e.g., lines 4 to 6). Examinee ID68 showed sensitivity at times toward handling the disagreement action by employing a pause and hedge, but his sequential competence was limited compared to examinee ID67.

Let us then examine how the two examinees above perform on the second sub-situation. In Excerpt 9, examinees ID67 and 68 are to decide on a meeting mode for their upcoming group meeting. Separate role-play cards were given to each of the examinees and the group's third member, Tom, is absent. Instead of having Phoenix starting the conversation (as seen in Excerpt 8), now Jessie starts the role-play conversation, proposing the two meeting options (face-to-face vs. online chat) and suggesting what she personally prefers. Here, Phoenix does not know the two meeting options until he hears from Jessie. Both examinees, though, know that the conversation is about deciding how to meet to discuss an upcoming group project.

Excerpt 9: Advanced and Intermediate Level Examinees
Decide a meeting mode for the meeting
J: Jessie (Advanced, examinee ID67)
P: Phoenix (Intermediate, examinee ID68)

```
3  J: so: I was thinking about meeting for the: project↑
4     to discuss the project? But: as both of us are:
5     >well not Tom< we have like really busy schedules we
6     could maybe do some research online and exchange some
7     e-mails or chatting? And: when we have whole information
8     meet face-to-face to uh: (1.4) finish whole the project
```

```
9      and see how we are going to present it? what do you think?
10 P: I think it's better uh: (2.5) we can meet together right?
11 J: yeah we can on saturday
12 P: yeah we can meet on saturday but (1.0) because our
13    conversation we can do that by e-mail it's better
14 J: so you don't want to meet face-to-face (0.8) you seem
15    to you don't need to=
16 P: =we meet face-to-face on- on the:
17 J: Yes
18 P: but we want to make sure it does because we cannot
19    meet Tom, right?=
20 J: =uh huh
21 P: and (0.8) and the best way I- we send we send uh: him
22    any e-mail or take him or xxx and let's xxxx he's xxxx
23    (1.0)
24 J: so what do you I don't (.) really understand
25    what do you prefer .hh do you prefer to meet
26    face-to-face on saturday? or do you prefe:r
27    to e-[mail
28 P:      [saturday you mean face-to-face
29 J: Okay and before that we e-ma[il (.) research=
30 P:                             [yeah
31 P: =yeah
```

In lines 3 to 9, Jessie makes an explicit and specific proposal on the meeting mode, namely *exchanging information via e-mails or chatting first and then meeting face-to-face*. Jessie utilizes various linguistic resources to express her opinions, including the use of a past progressive (*was thinking*), modal verb (*could*), and adverb (*maybe*). At the end of her turn in line 9, Jessie asks a question (*what do you think*) specifically designed to seek an answer from Phoenix, orienting to the need for the group to find a jointly agreed decision for their project. Responding to the Jessie's question, Phoenix' response starts with *I think it's better*, producing a cohesive tie to the Jessie's opinion solicitation question format, followed by a 2.5 second pause in line 10. Without completing the utterance, Phoenix launches a different question about whether they can meet or not. Jessie responds by suggesting that they can meet on Saturday. Phoenix repeats what Jessie mentions in line 12 and elaborates his opinion using the conjunction *but* between the *face-to-face* and *e-mail* options in lines 12 to 13, mentioning the adjective *better* after the *e-mail* option. Based on the conjunction and adjective, Phoenix

seems to prefer the e-mail exchange option. In lines 14 and 15, Jessie also shows her understanding of Phoenix' previous statement accordingly, asking whether Phoenix does not prefer the *face-to-face* option. However, in line 16, Phoenix brings up the *face-to-face* option, and Jessie confirms this option with an emphasized *yes* in line 17. So far, Phoenix does not clearly express a preference, which causes confusion for Jessie. In lines 18 and 19, Phoenix mentions Tom, who is absent during the conversation, and further elaborates his opinion in lines 21 and 22. However, Phoenix' utterance is neither clear nor coherent. It seems that Phoenix is concerned with the fact that Tom is not present and that his absence makes it difficult to come to a decision. After a short pause in line 23, Jessie finally explicitly expresses the lack of clarity in Phoenix' preference, and asks a questions designed to seek an answer from him in lines 24 to 27. Phoenix mentions the face-to-face option, which Jessie treats as Phoenix' preference. Jessie also adds the e-mail exchange option in line 29. This option is also seemingly agreed to by Phoenix, as he issues the acknowledgement tokens *yeah* in line 30.

Summarizing the two examinees' performances, confusion and troubles expressed by the participants themselves emerged mainly due to examinee ID68's (Phoenix) unclear language use and content delivery. In the course of the task, examinee 67 (Jessie, advanced level) also oriented to the confusion and attempted to resolve it. Contrastingly, examinee 67 expressed her opinions and expressions clearly, as her turns were accompanied by not only accounts (e.g., line 5, 13) but also diverse linguistic resources.

Interaction between low-intermediate level examinees. Let us then examine how low-intermediate level examinees perform in the role-play task with a classmate. Excerpt 10 demonstrates part of the performances of two low-intermediate level

examinees in the role-play of deciding the meeting time. Examinee ID79 played Jessie, and examinee ID80 played Phoenix.

Excerpt 10 Low-intermediate Level Examinees
Decide a time for a class project with a classmate
J: Jessie (low-intermediate, Examinee ID79)
P: Phoenix (low-intermediate, Examinee ID80)

```
1  P: hi Jessie
2  J: hi
3  P: hmm h- how about our article presentation?
4     (1.0)
5  J: hmm: so I'm also xx next saturday I'm free how about you?
6  P: next thursday?
7  J: yeah
8  P: I- I have um classes from 10p- 10a[m to 3 pm (.)=
9  J:                                   [hm
10 J: =hm
11    (1.5)
12 P: so, (0.9) hm I can I can do it after 3pm
13    (1.3)
14 J: ah↑ yeah ((laugh)) yeah I- I can do it after 3pm,
15    so (1.5) we we have to tell Tom↑ (1.0) to meet after (0.8)
16    3pm on next saturday=
17 P: =yes
18    (4.3)
19 P: ok[ay, let's do=
20 J:   [okay
21 J: =mm
```

After a greeting adjacency pair (lines 1 and 2), Phoenix initiates a question in line 3, formulated with the broad opinion-asking expression *how about*. After a 1 second pause, Jessie proposes an availability on Saturday in line 5, displaying her understanding of Phoenix' broad question as a proposal to decide a good time to meet. Jessie also ends the utterance with a question *how about you* in line 5, which is designed to seek Phoenix's answer. In response to Jessie's question, Phoenix asks a confirmation question by recycling the time reference word (*Thursday*) in line 6. However, they do not share the time information correctly, since it was *Saturday*, not *Thursday*, that Jessie originally proposed (line 5). Nonetheless, Jessie confirms it in line 7. These sequences indicate a lack of establishing shared correct understandings between the two examinees, although they complete an adjacency pair (i.e., question & answer). In line 8, Phoenix further

elaborates her class schedules, accompanied by Jessie's acknowledgement tokens in lines 9 and 10. A 1.5 second pause in line 11 indicates Phoenix's orientation that it is Jessie's turn to respond to Phoenix's schedules. However, as Jessie does not provide an immediate response, Phoenix orients to the pause by further elaborating the time availability in line 12. In this second attempt, Phoenix now makes her availability more specific, compared to the previous turn in line 8, indicating her orientation to achieving the task's interactional goal, which is to find a good meeting time. Without an immediate uptake in line 13, Jessie endorses the time in line 14 with the change-of-state token *ah* and laughter. Jessie also mentions informing Tom, the group's third member who is absent, about their decisions on the time. Several pauses in Jessie's turn in lines 14 to 16 are present. After a long pause in line 18, they close the conversation with *okay*, initiated by Phoenix, and it overlaps with Jessie's *okay*, which is a common close sequence (Schegloff, 2007, p. 121).

To summarize the two low-intermediate level examinees' performances, their utterances were minimal without the use of various linguistic expressions. Regarding their interactional competence, the two examinees did not achieve the shared understanding of the correct information (about the available date of a week). This suggests that they do not fully engage in the role-play conversation. The frequent pauses between and within turns also indicate that the two examinees' lack of fluency in their language use.

Summary of the CA Findings on Examinees' Pragmatic Performances

The following features were identified across different levels of examinee based on the CA findings, as summarized in Table 7. The noticeable features included language use, interactional features, and varying degrees of situational knowledge. The language use in L2 pragmatics is theorized as pragmalinguistics, along with sociopragmatics (sensitivity to situations) (Leech, 1983). It was evident that the advanced level examinees not only used complex structures but also utilized *various* linguistic resources in delivering their pragmatic meanings. On the other hand, intermediate and low-intermediate level examinees mostly used mono-clausal structures without utilizing the various structures. For example, the single modal verb *can* was present in their language use, rather than other types of modal verb, such as *could*, *might*, and *would*. The degree of situational knowledge in an EAP setting also differed, although variations existed. Regarding the interactional features, the examinees sometimes did not complete adjacency pairs or their turns displayed the irrelevance in their second parts of the adjacency pair. Additionally, the normally expected features, such as delays or hedges, for dispreferred actions (e.g., refusals, disagreements) during the role-play interaction, were not present.

The excerpts selected for the analysis were from the examinees with different levels, all professor interlocutors participated in the study, and different role-play task situations. Additionally, the contents of the role-play tasks were quite consistent across the examinees due to the use of role-play cards, resulting rather predictable language use and structured sequential organizations. Thus, although not all examinees' performances

were analyzed, it is argued that the characteristics found from the analyses of the excerpts selected are trustworthy, representing the characteristics of the rest of the examinees' performances. Nonetheless, it is still questionable whether raters discern the characteristics consistently in their ratings, which is examined in an upcoming quantitative analysis section.

Table 7
Summary of Examinees' Performances

Characteristics	Advanced	Intermediate	Low-intermediate
Delivery	• smooth turn initiation • clear, concise, fluent within turns	• abrupt turn initiation • less fluent than advanced level examinees	• choppy, limited • noticeable pauses within a turn
Language Use	• utilize diverse linguistic resources (e.g., complex structures, conditionals, past progressive, various modal verbs (could, might, would))	• inaccurate at times • modal verbs (can, could), but lack of diverse expressions	• direct expressions (must, need to) • sometimes use modal verbs (can) • no complex structures
Sensitivity to Situations	• provided accounts for a request, refusal, and disagreements • had institutional knowledge of an EAP context • variations exist	• sometimes express situational sensitivity in tone • variations exist	• sometimes express sensitivity in tone • variations exist
Interactional Features	• completed adjacency pairs • aware of the use of meaningful pauses and mitigation between turns • formulated the contingency in making request and understood interlocutor's actions	• engaged in the conversation • inconsistently complete adjacency pairs and use meaningful pauses	• lack of establishing shared understandings • engaged in the conversation with further assistance from the interlocutor • lack of sensitivity to meaningful pauses

Integrating CA Findings into the Development of Rating criteria

The following aspects were considered in integrating the CA findings into the development of the interaction-sensitive analytical rating criteria for the two open role-play tasks. First of all, examples and descriptions included in the rating criteria were

carefully selected. Most importantly, they need to be easily understood by raters to ensure consistency and reliability, which can be achieved by ensuring clarity of each rating category. They also need to be objectively observable across the examinees. For example, an absence of certain characteristics in examinees' language uses can be a clear indication of a lack of L2 pragmatic competence. A mono-clausal request expression *I need a recommendation letter*, unless it is accompanied by other expressions, sounds less polite than a complex grammatical structure with modal verbs *I was wondering if you could write a letter*. Additionally, the interactional features that are included in the rating criteria were carefully selected focusing on concrete and objectively recognizable evidence. For example, an absence of the second pair part of adjacency pairs (e.g., question-answer) can be objectively recognizable, as it is easily noticeable during interaction.

Secondly, detailed descriptions for the ideal performance in the rating criteria, namely perfect scores for all criteria, were cautiously determined. The concept of ideal pragmatic performance is subject to a criticism as one can argue about whose performance is ideal considering great variations that exist even among native speakers. Toward this issue, the intended uses of the assessment tasks and rating criteria were primarily considered. It can be argued that the ideal performance comprising of various resources (e.g., linguistic and interactional features) can be introduced to students for classroom assessment purposes.

Thirdly, each rating criterion needs to reflect distinct features of L2 pragmatic performance in interaction, which can differentiate between varying examinees' L2 pragmatic abilities. Equally importantly, they need to be succinct enough to ease scoring

and easily understood by raters. In order to fulfill theses purposes, for example, two distinct aspects, *Engaging with Interaction* and *Turn Organization*, were included to measure interactional competence in the rating criteria. As shown in the CA findings of this study, engaging in the conversation does not necessarily entail examinees' abilities to take turns appropriately or understandings of the meaningful pauses for dispreferred actions.

Lastly, two different task-dependent rating criteria were developed for the two open role-play tasks, in consideration of not only the intended uses but also different degrees of expectation and contents embedded in each task. Although the same five rating categories were used for both criteria, descriptions included in each criteria were slightly different. Different characteristics specific to each role-play situation were considered so that raters can score target performances consistently. For example, the descriptions for the *Sensitivity to Situations* category for each task were different, as requests and refusals were the main speech acts in the role-play task with a professor while agreements and disagreements were expected in the role-play task with a classmate. Based on the considerations discussed above, the five rating categories on a 3-point scale, (a) *Content Delivery*, (b) *Language Use*, (c) *Sensitivity to Situations*, (d) *Engaging with Interaction*, and (e) *Turn Organization*, were included in the finalized version of two different rating criteria for the role-play tasks (see Appendices D and E).

Many-facet Rasch Measurement Analysis using FACETS

This section reports quantitative results from a FACETS analyses. Three separate FACETS analyses, first for the open role-play tasks only, second for the monologic tasks only, and lastly for both open role-play and monologic tasks combined based on averaged composite scores, were conducted for the following reasons. As different rating criteria with a different number of rating categories were employed for different types of tasks, composite scores, rather than a score from each rating category, need to be used to include all tasks in a single FACETS analysis. However, the composite scores (i.e., averaged scores from all rating categories) do not provide detailed information on each rating category for the open role-play tasks. However, this information constitutes important validity evidence. Additionally, as one of the primary purposes of conducting FACETS analysis in this study is to analyze the characteristics of examinees' performances in the open role-play tasks separately from those in the monologic tasks, different FACETS analyses are more appropriate. All FACETS analyses employed the Partial Credit Model (Masters, 1982), which assumes that all raters used different rating criteria when assigning scores. Results in all subsequent sub-sections are presented and discussed in the order of FACETS summary and measurement reports for each facet (i.e., examinee ability, rater severity, task difficulty, and rating criteria difficult). Before reporting the results from FACETS analysis, the descriptive statistics are presented.

Descriptive Statistics

This section reports the descriptive statistics for 12 raters across all task types (role-play and monologic tasks) and sub-situations within each task type. Table 8 details

the raw score average and standard deviation of all tasks across all raters. As maximum and minimum scores for each rater were mostly three and one respectively (or zero for the monologic tasks), they were not included in the table to minimize redundancy. Each raw score average and standard deviation was based on the composite scores from the rating categories in the rating criteria. Although the scores varied across the raters, a few noticeable patterns emerged. Out of the five role-play sub-situations, the first task of *making a recommendation letter request* received the lowest average score with the highest standard deviation from most of the raters, except for Raters 1 and 2. Out of the 12 raters, Rater 12 gave the lowest average score (1.95) for this difficult task (*making a recommendation letter request*), indicating that Rater 12 might be one of the severe raters. However, it cannot be concluded that Rater 12 was the most severe rater considering Rater 12's relatively lenient score on other tasks, such as *2. requesting more time from professor* under role-play with a professor. The most lenient rater is not clearly identified. Additionally, the monologic task with the lowest average score (i.e. the most difficult task) is not clearly identified, as the scores vary across the raters. In fact, based on the raw scores, the identification of the most difficult or the most lenient rater is not that simple. In particular, a score awarded to an examinee in performance-based language assessment is a result of the combination of various factors, such as task difficulty and rater severity. Thus, the raw scores alone do not provide accurate estimates of examinees' abilities, which is one of the limitations of classical test theory and why a many-facet Rasch measurement is needed.

Table 8
Descriptive Statistics

Type	Description	R1		R2		R3		R4		R5		R6	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
RP with professor	1. Request a letter	2.03	.58	2.14	.48	2.07	.52	2.27	.62	2.38	.55	2.12	.57
	2. Request time	2.06	.50	2.35	.46	2.20	.45	2.52	.48	2.54	.32	2.26	.48
	3. Refusal	2.02	.44	2.39	.44	2.18	.46	2.49	.39	2.50	.38	2.35	.37
RP with classmate	4. Decide a time	1.98	.49	2.07	.46	2.41	.47	2.39	.59	2.55	.32	2.33	.44
	5. Decide a meeting mode	1.99	.52	2.10	.44	2.41	.46	2.39	.59	2.55	.31	2.42	.50
Monologic task	6. Preference	1.87	.60	2.21	.49	2.12	.56	2.24	.60	2.09	.71	2.09	.64
	7. Choice	2.00	.53	2.29	.63	2.24	.61	2.21	.81	1.82	.80	2.19	.71
	8. Give comments	1.88	.64	1.88	.61	2.14	.63	1.98	.84	1.66	.72	1.91	.55

Table 8 (continued)
Descriptive Statistics

Description	R7		R8		R9		R10		R11		R12		All	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
1. Request a letter	2.40	.44	1.98	.50	2.25	.50	2.05	.52	2.13	.54	1.95	.61	2.15	.54
2. Request time	2.67	.30	2.35	.32	2.49	.41	2.41	.43	2.59	.35	2.35	.56	2.40	.42
3. Refusal	2.59	.30	2.35	.33	2.45	.49	2.30	.48	2.65	.36	2.18	.41	2.37	.40
4. Decide a time	2.65	.23	2.26	.34	2.29	.49	2.29	.42	2.68	.33	2.09	.55	2.33	.43
5. Decide a meeting mode	2.69	.24	2.46	.36	2.33	.52	2.39	.41	2.65	.32	2.15	.59	2.38	.44
6. Preference	2.04	.73	1.96	.64	2.14	.72	1.76	.67	2.46	.61	1.98	.63	2.08	.63
7. Choice	2.26	.71	1.89	.68	2.29	.72	1.89	.68	2.51	.66	2.11	.67	2.14	.68
8. Give comments	2.23	.50	1.97	.65	2.08	.68	1.93	.66	2.29	.72	2.26	.74	2.02	.66

Two Open Role-Play Tasks

FACETS Summary. All facets can be mapped into a single set of relationships, which is one of the most useful characteristics using the same scale called *logits*. By

doing so, we can investigate, for example, whether a test is too easy or too difficult for a particular group. Figure 2 presents a FACETS summary, also called *a variable map*, that places all facets on the same logit scale. The logit scale basically transforms the probability of a particular response in a more approachable fashion (McNamara, 1996, p. 165). As the logit scale is a true interval scale, it represents a true difference between differing examinee abilities, difficulties, or rater severities. By convention, the average difficulty of items is set at zero logits. An examinee ability expressed as 0 logits has a 50 percent chance of getting right an item of average difficulty.

Four facets (*examinee*, *task*, *rater*, and *rating criteria*) were included in the first FACETS analysis. Each column presents information of elements within each facet. From the left-hand side of the columns, the first column shows the measurement scales in logits. Each facet is positioned on the logit scale. The second column displays the distribution of examinees' varying abilities. Each asterisk represents each examinee. Ability measures are ordered with more proficient examinees appearing at the top of the column, and less proficient examinees appearing at the bottom. In Figure 2, more examinees are shown above the zero logit, indicating that more examinees were above intermediate ability levels.

The third column compares 12 raters in terms of their degrees of severity and leniency. More severe raters appear higher in the column and less severe raters appear lower. Noticeable gaps between elements indicate the raters varied in their degrees of severity for the rating. The fourth column shows five tasks, three (request letter, request time, refuse) from role-play with a professor and two (time, mode) from role-play with a classmate. More difficult tasks appear higher in the column while less difficult tasks

appear lower. The most difficult task, which is located at the top, was *Request a recommendation letter to professor*, while other tasks were clearly less difficult. Five rating categories' difficulty levels are displayed in the fifth column. Gaps between the rating categories indicate each category varied considerably in terms of difficulty. The most demanding category, located at the top, was *Language Use* while the least difficulty category, located at the bottom, was *Engaging with Interaction*. As a partial credit model, which assumes each item has its own scale structure (Masters, 1982; Wright & Masters, 1982), was used, the sixth to the last columns respectively present how scores for five tasks were utilized. Score assignments for the most difficult task (S.1 Request a recommendation letter to professor), for instance, the lowest score 1 was more widely used compared to the other tasks.

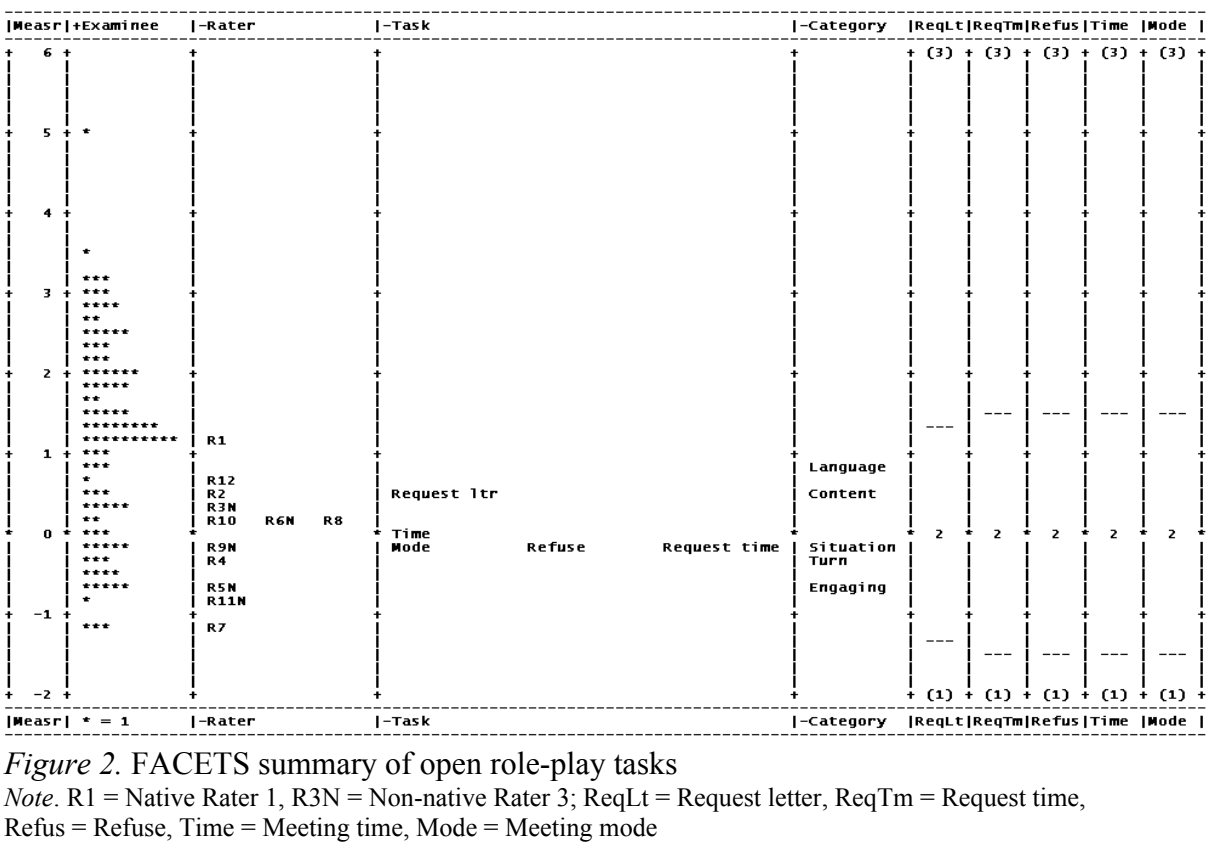


Figure 2. FACETS summary of open role-play tasks
Note. R1 = Native Rater 1, R3N = Non-native Rater 3; ReqLt = Request letter, ReqTm = Request time, Refus = Refuse, Time = Meeting time, Mode = Meeting mode

Examinee measurement report. FACETS analysis provides measurement reports for each facet along with detailed information on elements within each facet in terms of model fit and reliability. As shown in the FACETS summary, examinees were widely spread out with logit values ranging from –1.24 and 5.08 (M=1.20, SD=1.24), with a total spread of 6.32 logits. The separation value was 6.01 with its reliability index of .97, indicating that the open role-play tasks reliably separated examinees' varying ability. The reliability for the *examinee* facet in FACETS is equivalent to Cronbach Alpha (the ratio of True variance to observed variance). In terms of fit statistics, FACETS provides two fit statistics, infit and outfit, which require slightly different interpretations. As outfit statistics are more sensitive to outliers and unusual infit values cause more concern, Rasch researchers usually focus on infit statistics. Generally accepted infit values range from 0.5 to 1.5 (Lunz, Wright, & Linacre, 1990), while a more conservative fit range is 0.8 and 1.2 (McNamara, 1996). Another widely used method is to use infit mean squares adjusted by their mean (M) and the standard deviation (SD) values, which is calculated as the mean plus (or minus) twice the standard deviation (i.e., $M \pm (2 \times SD)$). An infit value over the upper-control limit is considered *misfit* indicating too much unpredictability while an infit value below the lower-control limit is considered *overfit* displaying lack of variation (Eckes, 2011, p. 58). Based on these criteria, three examinees were slightly over the upper-control limit (i.e., overfit) and four were below the lower-limit (i.e. misfit), based on $M \pm (2 \times SD)$ (i.e., 1.31~0.67=0.99±(2×0.16)). However, the degrees of misfit were all negligible. Applying the less restrictive 0.5 to 1.5 infit value range, all examinees were within the acceptable range, indicating that no examinee

behaved erratically. The full result of examinee measurement report is presented in Appendix J.

Rater measurement report. Table 9 presents logit values, errors, and fit statistics for 12 raters. The logit values and fit statistics for the *rater* facet indicate overall severity and self-consistency respectively for each rater. The higher logit values, the more severe the raters are, and vice versa for the lower logit values. The most severe rater was Rater 1 with the highest logit of 1.15 and the most lenient rater was Rater 7 with the lowest logit of –1.10, which is also clearly presented in the second column of Figure 2. However, the meaningfulness of the clear differences in severity logits need to be supported by three additional statistics: (a) the separation index, (b) the reliability of the rater separation index, and (c) the fixed (all same) chi-square. The separation index of 10.68 with its reliability, 0.99, indicates the raters reliably showed different levels of severity. In other words, if all raters' levels of severity were all equal, the reliability of rater separation would be close to 0.00. This can be an ideal situation, but the different degrees of rater severity in L2 pragmatic performance are not really surprising. Additionally, it should be noted that a reliability reported in FACETS is not inter-rater reliability, but it indicates how measures are different. This finding is confirmed by the significant chi-square statistic (χ^2=1415.9, df=11, $p < .00$), rejecting the null hypothesis that each rater is equal in terms of severity.

The last column in Table 9 presents fit statistics. The infit values for all raters were within the acceptable range of 0.77 and 1.25, based on $M \pm (2 \times SD)$ (i.e., $1.01 \pm (2 \times 0.12)$), indicating that all raters showed internal consistency in their ratings. The stable fit indices for the raters indicate that the differences in rater severity are not the

result of unpredictable rater behavior. Note that the 12 raters consisted of the eight native speakers and four non-native speakers, who were indicated with an *N* sign in Table 9. Regardless of the raters' L1s, they all showed internal consistency. Regarding the severity, no non-native raters were quite severe, as the three most severe raters (Raters 1, 12, and 2) were all native raters. Although it is a preliminary observation, the non-native raters showed either average severity or leniency, as also seen in the second column in Figure 2 above.

Table 9
Measurement Report for Raters (Open Role-play Tasks)

Raters	Severity Logit	Model Error	Infit Mean Square
1	1.15	0.04	1.02
2	0.52	0.04	0.78
3[N]	0.36	0.05	1.04
4	-0.27	0.05	0.91
5[N]	-0.62	0.05	1.12
6[N]	0.15	0.05	0.98
7	-1.10	0.08	0.91
8	0.12	0.07	0.93
9[N]	-0.16	0.07	1.02
10	0.10	0.07	0.98
11[N]	-0.83	0.07	1.21
12	0.58	0.07	1.17
M	0.00	0.06	1.01
SD	0.64	0.01	0.12

Note. [N] non-native rater; Reliability: 0.99; Separation: 10.68; Fixed chi-square: 1415.9 (*d.f.*=11; $p < .00$)

Task measurement report. Table 10 presents detailed information for the five open role-play situations including relative task difficulty in logits, model error, and fit statistics. For the *task* facet, the logit values indicate task difficulty. The most difficult task was *Request a recommendation letter* from role-play with a professor with the highest logit value of 0.56, and the easiest task was *Decide a meeting mode for class*

project from role-play with a classmate with lowest logit of –0.20. Another role-play task with a classmate (*Decide a meeting time*) was also similarly easy (logit = –0.18). Thus, generally speaking, the role-play task with a professor was more difficult than the role-play task with a classmate. The high reliability of 0.99 of the separation index (9.07) indicates that the role-play tasks were not identical in terms of difficulty, confirmed by the significant chi-square statistic (χ^2= 361.4, df = 4, p< .00). The reliability index for the task facet also indicates the replicability of tasks' spread along the continuum if these same tasks were given to another sample of the same size that behaved the same way. The model errors were all negligible, which is ideal. The infit statistics in the last column were all within the acceptable fit range of 0.79 and 1.19, based on M±(2×SD) (i.e., 0.99±(2×0.1)). These acceptable fit values for the tasks also "help to determine whether the item estimations may be held as meaningful quantitative summaries of the observations" (Bond & Fox, 2007, p. 35). That is to say, the stable fit values indicate that the tasks contribute to measuring one construct (i.e., unidimensionality), which is an important assumption of Rasch model.

Table 10
Measurement Report for Open Role-play Tasks

Task	Difficulty Logit	Model Error	Infit Mean Square
1. Request a recommendation letter [a]	0.56	0.03	1.03
2. Request time to discuss a project [a]	-0.04	0.04	0.90
3. Refusal to professor's request [a]	-0.14	0.04	0.87
4. Meeting Time [b]	-0.18	0.04	1.12
5. Meeting Mode [b]	-0.20	0.04	1.03
M	0.00	0.04	0.99
SD	0.32	0.00	0.10

Note. [a] Role-play with professor; [b] Role-play with classmate; Reliability = 0.99; Separation: 9.07; Fixed chi-square: 361.4 (*d.f.*=4; p < .00)

Rating criteria measurement report. Table 11 presents detailed information on the five categories in the rating criteria. Each rating category was different in terms of levels of difficulty, indicated by a high reliability index of 1.00 with statistical significance (χ^2=1346.9, $df = 4$, $p < .00$). Among the five categories, *Language Use* was the most difficult with the highest logit (0.85) while *Engaging with Interaction* was the least difficult with the lowest logit (-0.74). Generally speaking, *Language Use* and *Content Delivery* were rather demanding categories. Between the two categories for interactional competence, *Turn Taking* (logit = -0.36) was more difficult than *Engaging with Interaction* (logit = -0.74). The infit statistics in the fourth column show that all categories were within the acceptable fit range of 0.78 and 1.22, based on $M \pm (2 \times SD)$ (i.e., $1 \pm (2 \times 0.11)$). Regarding the two categories for interactional competence, *Turn Organization* and *Engaging with Interaction*, it is worth to note that they had different levels of difficulty with the stable infit statistics. These findings indicate each rating category tapped into a distinct aspect of interactional competence and the raters were able to differentiate and use them consistently.

Table 11
Measurement Report for Five Rating Criteria

Rating Criteria	Difficulty Logit	Model Error	Infit Mean Square
1. Content Delivery	0.46	0.03	0.91
2. Language Use	0.85	0.03	0.87
3. Sensitivity to Situations	-0.21	0.04	1.16
4. Engaging with Interaction	-0.74	0.04	1.02
5. Turn Organization	-0.36	0.04	1.02
M	0.00	0.04	1.00
SD	0.64	0.00	0.11

Note. Reliability = 1.00; Separation Index: 18.24; Fixed chi-square: 1346.9 (*d.f.*=4; $p < .00$)

Three Monologic Tasks

The two different monologic task types, monologic speaking tasks (preference, choice) and monologic pragmatic task (giving comments to a classmate), were included in the second FACETS analysis. Compared to the role-play tasks, more iterations were needed to calibrate the facets in the monologic tasks. As two different rating criteria with a different number of rating categories were used for each task type (three rating categories for the monologic speaking tasks and four categories for the monologic pragmatic task, see Appendices F and G), the averaged composite scores were used. Thus, three facets (*examinee*, *rater*, and *task*) were included in the analysis. Unlike the role-play tasks, zero was awarded for either irrelevant responses or too limited performance. Figure 3 presents the three facets in a single set of relationship. Each asterisk represents two examinees and a dot indicates one examinee. More examinees are shown above the zero logit, indicating that more examinees are above intermediate ability levels. A wide gap between elements within each facet is also shown.

```
-----------------------------------------------------------------
|Measr|+Examinee|-Rater            |-Task       |Pref |Choic|Comme|
-----------------------------------------------------------------
+   8 + **      +                  +            + (3) + (3) + (3) +
|     | .       |                  |            |     |     |     |
|     |         |                  |            |     |     |     |
+   7 + .       +                  +            +     +     +     +
|     | *       |                  |            |     |     |     |
|     | **      |                  |            |     |     |     |
+   6 +         +                  +            +     +     +     +
|     |         |                  |            |     |     |     |
|     | ***     |                  |            |     | --- |     |
+   5 + *       +                  +            +     +     +     +
|     | ******* |                  |            |     |     | --- |
|     | **      |                  |            |     |     |     |
+   4 + *       +                  +            +     +     +     +
|     | *.      |                  |            |     |     |     |
|     | **      |                  |            |     |  2  |     |
+   3 + ***.    +                  +            +     +     +     +
|     |         |                  |            |     |     |     |
|     | ***.    |                  |            |     |     |     |
+   2 + ***.    +                  + Preference + --- +     +  2  +
|     |         |                  |            |     |     |     |
|     | *.      | R10              |            |     | --- |     |
+   1 + **.     + R1    R5N   R8   +            +     +     +     +
|     |         |                  |            |     |     |     |
|     | .       |                  |            |     |     |     |
*   0 * **      * R12   R2    R9N  *            *  2  *     *     *
|     | .       | R6N   R7         |            |     |     | --- |
|     | *       | R4               | Comments   |     |     |     |
+  -1 + *.      + R3N              +            +     +     +     +
|     | *.      |                  |            |     |     |     |
|     | **.     |                  | Choice     |     |     |     |
+  -2 + *.      + R11N             +            + --- +     +  1  +
|     |         |                  |            |     |     |     |
|     | *       |                  |            |     |  1  |     |
+  -3 +         +                  +            +     +     +     +
|     |         |                  |            |     |     |     |
|     |         |                  |            |     |     |     |
+  -4 +         +                  +            +     +     + --- +
|     | .       |                  |            |     |     |     |
|     |         |                  |            |     |     |     |
+  -5 +         +                  +            +     +     +     +
|     | .       |                  |            |     |     |     |
|     |         |                  |            |     |     |     |
+  -6 +         +                  +            + (1) + (0) + (0) +
-----------------------------------------------------------------
|Measr| * = 2   |-Rater            |-Task       |Pref |Choic|Comme|
-----------------------------------------------------------------
```

Figure 3. FACETS summary of monologic tasks

Note. R1 = Native Rater 1, R6N = Non-native Rater 6; Pref = Preference Monologic Speaking Task, Choic = Choice Monologic Speaking Task, Comme = Giving comments

Examinee measurement report. A wide spread of examinees was observed with logit values ranging from –5.38 and 7.75 (M = 2.56, SD = 2.92), with a total spread of 13.13 logits, which is quite wide compared to the spread shown in the open role-play tasks (a spread of 6.32 logits). This is due to two examinees whose abilities are much lower than other examinees, and they are shown at the bottom of the second column in Figure 3. The separation value was 3.64 with its reliability index of .93, analogous to Cronbach alpha. Although the reliability was not as high as the role-play tasks, these findings indicate that the three monologic tasks reliably separated examinees' varying ability. However, four examinees with quite high ability logits were identified as extreme results that did not generate infit statistics. Based on the fit criteria, $M \pm (2 \times SD)$ (i.e., 1.81~0.21=1.01±(2×0.40)), six examinees were over the upper-control limit (i.e., overfit)

and no examinees were below the lower-limit (i.e. misfit). The full result of examinee measurement report is presented in Appendix K.

Rater measurement report. Table 12 presents the 12 raters' severity measures and fit statistics. Raters' severity logits noticeably differed, based on the separation index of 4.49 with a high reliability of 0.95, supported by statistical significance (χ^2=283.5, *df* = 11, *p*< .00). Error values for the first six raters (Rater 1 to Rater 6) were a bit higher (0.24) than those (0.16) from the next six raters (Rater 7 to Rater 12). This might be due to a smaller amount of ratings observed from Raters 1 to 6 for the monologic tasks compared to other raters, as Raters 1 to 6 rated 30 examinees while Raters 7 to 12 rated 67 examinees. In terms of fit statistics, Rater 1 (infit = 0.60) was slightly outside of the lower-control limit (0.61=0.99-22×(0.19)) and Rater 5 (infit = 1.38) was slightly greater than the upper-control limit (1.37=0.99+2×(0.19)). Other raters, however, showed stable fit statistics within the range of 0.61 and 1.37.

Table 12
Measurement Report for Raters (Monologic Tasks Only)

Raters	Severity Logit	Model Error	Infit Mean Square
1	0.96	0.24	0.60
2	0.04	0.24	0.83
3[N]	-0.87	0.24	0.98
4	-0.58	0.24	1.01
5[N]	1.02	0.24	1.38
6[N]	-0.18	0.24	1.03
7	-0.42	0.16	1.10
8	0.98	0.16	1.05
9[N]	-0.14	0.16	0.99
10	1.30	0.16	0.77
11[N]	-1.97	0.18	1.15
12	-0.15	0.16	1.02
M	0.00	0.20	0.99
SD	0.95	0.04	0.19

Note. [N] non-native rater; Reliability: 0.95; Separation Index: 4.49; Fixed chi-square: 283.5 (*d.f.*=11; *p* < .00)

Task measurement report. The three monologic tasks' difficulty logits are presented in Table 13. The three tasks distinctly differed in difficulty based on the separation index of 19.17 with a high reliability of 1.00, supported by statistical significance (χ^2=723.0, *df* = 2, *p*< .00). The *Preference* task was the most difficult and the *choice* task was the least difficult. The infit values for all tasks were all within the acceptable range of 0.82~1.18 based on 1±(2×0.09), indicating that the tasks contribute to measuring one construct.

Table 13
Measurement Report for Monologic Tasks

Task	Difficulty Logit	Model Error	Infit Mean Square
1. Preferece [a]	2.04	0.10	1.05
2. Choice [a]	-1.52	0.10	0.90
3. Giving comments [b]	-0.53	0.09	1.04
M	0.00	0.04	1.00
SD	1.84	0.00	0.09

Note. [a] Monologic Speaking Tasks; [b] Monologic Pragmatic Task; Reliability = 1.00; Separation: 19.17; Fixed chi-square: 723.0 (*d.f.*=2; *p* < .00)

All Tasks Combined

As the final FACET analysis, this section reports results for both open role-play and monologic tasks combined based on composite scores (i.e., average scores from different categories in the rating criteria) as each task type was scored by the different rating criteria. Figure 4 presents a FACETS summary. Three facets (*examinee*, *rater*, and task) were calibrated. The 102 examinees' abilities are widely distributed in the second column. As observed in the previous two FACETS analyses, many examinees are located above the logit value of zero, indicating that many examinees' levels are above the

intermediate level. Wide gaps between elements for the *rater* and *task* facets indicate each element differed in terms of severity and difficulty, respectively.

```
-----------------------------------------------------------------------------------------------------------------------------
|Measr|+Examinee |-Rater      |-Task                                      |ReqLt|ReqTm|Refus|Time |Mode |Pref |Choic|Comme|
-----------------------------------------------------------------------------------------------------------------------------
+   6 + *        +            +                                          + (3) + (3) + (3) + (3) + (3) + (3) + (3) + (3) +
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
|     |          |            |                                          |     |     |     |     |     |     |     |     |
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
|     | **       |            |                                          |     |     |     |     |     |     |     |     |
+   5 + ****     +            +                                          +     +     +     +     +     +     +     +     +
|     |          |            |                                          |     |     |     |     |     |     |     |     |
|     | **       |            |                                          |     |     |     |     |     |     |     |     |
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
+   4 + **       +            +                                          +     +     +     +     +     +     + --- +     +
|     | ***      |            |                                          |     |     |     |     |     |     |     |     |
|     | *****    |            |                                          |     |     |     |     |     |     |     |     |
|     | ********* |           |                                          |     |     |     |     |     |     |     | --- |
|     | ****     |            |                                          |     |     |     |     |     |     |     |     |
+   3 + **       +            +                                          +     +     + --- +     +     +     +     +     +
|     | ***      |            |                                          |     |     |     |     |     |     |     |     |
|     | ******   |            |                                          |     | --- |     |     |     |     |     |     |
|     | *****    |            |                                          |     |     |     | --- |     |     |  2  |     |
|     | *****    |            |                                          | --- |     |     |     | --- |     |     |     |
+   2 + *        +            +                                          +     +     +     +     +     +     +     +     +
|     | ***      |            |                                          |     |     |     |     |     |     |     |     |
|     | ***      |            | Preference                               |     |     |     |     |     | --- |     |     |
|     | ***      | R1         |                                          |     |     |     |     |     |     |     |  2  |
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
+   1 + **       +            +                                          +     +     +     +     +     +     + --- +     +
|     | **       |            | Request ltr                              |     |     |     |     |     |     |     |     |
|     | ***      | R10   R2   |                                          |     |     |     |     |     |     |     |     |
|     | *        | R12   R8   |                                          |     |     |     |     |     |     |     |     |
|     | **       | R6N        |                                          |     |     |     |     |     |     |     |     |
+   0 + **       + R3N        + Time                                     +  2  +  2  +  2  +  2  +  2  +  2  +     +     +
|     |          | R4         | Mode                                     |     |     |     |     |     |     |     | --- |
|     | **       | R9N        | Comments      Refuse        Request time |     |     |     |     |     |     |     |     |
|     | ***      | R5N        |                                          |     |     |     |     |     |     |     |     |
|     | ****     |            |                                          |     |     |     |     |     |     |     |     |
+  -1 + ****     + R7         +                                          +     +     +     +     +     +     +     +     +
|     | **       |            | Choice                                   |     |     |     |     |     |     |     |     |
|     | **       | R11N       |                                          |     |     |     |     |     | --- |     |  1  |
|     | *        |            |                                          |     |     |     |     |     |     |     |     |
|     | **       |            |                                          |     |     |     |     |     |     |     |     |
+  -2 + **       +            +                                          +     +     +     +     +     +     +  1  +     +
|     | *        |            |                                          | --- |     |     |     | --- |     |     |     |
|     |          |            |                                          |     |     |     | --- |     |     |     |     |
|     |          |            |                                          |     | --- |     |     |     |     |     |     |
+  -3 +          +            +                                          + (1) + (1) + (1) + (1) + (1) + (1) + (0) + (0) +
-----------------------------------------------------------------------------------------------------------------------------
|Measr| * = 1    |-Rater      |-Task                                      |ReqLt|ReqTm|Refus|Time |Mode |Pref |Choic|Comme|
-----------------------------------------------------------------------------------------------------------------------------
```

Figure 4. FACETS summary of all tasks combined

Note. R1 = Native Rater 1, R6N = Non-native Rater 6; ReqLt = Request letter, ReqTm = Request time, Refus = Refuse, Time = Meeting time, Mode = Meeting mode, Pref = Preference Monologic Speaking Task, Choic = Choice Monologic Speaking Task, Comme = Giving comments

Examinee measurement report. A wide spread of examinees' abilities was observed with logit values ranging from –2.24 and 7.48 (M = 1.92, SD = 2.12), with a total spread of 9.72 logits. One examinee, who was the most proficient (logit = 7.48), was identified as an extreme result. A separation value of 4.62 with extremes (4.88 without extremes) with a reliability index of .96 suggests that all tasks reliably separated between 102 examinees' abilities. The significant finding of the chi-square test (χ^2=3326.8, df=100, p< .00) rejects the hypothesis that all examinees' abilities are equal. Based on the fit criteria, M±(2×SD) (i.e., 1.46~0.54=1.00±(2×0.23)), three examinees were slightly

over the upper-control limit (i.e., overfit) and two were below the lower-limit (i.e. misfit). The full result of examinee measurement report is presented in Appendix L.

Rater measurement report. Table 14 presents a measurement report for the 12 raters in their ratings of all tasks. Rater 1 (logit = 1.37) was the most severe and Rater 11 (logit=-1.40) was least severe. The model errors, although not excessively high, are slightly higher than those from the open role-play tasks only, perhaps due to the higher errors observed from the monologic tasks. Considering the acceptable fit range of 0.67~1.31 based on 0.99±(2×0.16), all raters, except for Rater 5, showed acceptable infit values. The infit value observed for Rater 5 was 1.39. Although this is still acceptable compared to a less conservative fit range (0.5~1.5), it is a bit higher than the upper-control limit of 1.31, indicating that Rater 5 showed some unpredictability. In fact, Rater 5's infit value from the open role-play tasks was stable. Thus, such unpredictability is most likely from Rater 5's performance in rating the monologic tasks.

Table 14
Measurement Report for Raters (All Tasks Combined)

Raters	Severity Logit	Model Error	Infit Mean Square
1	1.37	0.10	0.95
2	0.55	0.11	0.84
3[N]	0.09	0.11	0.93
4	-0.26	0.11	1.09
5[N]	-0.55	0.11	1.39
6[N]	0.15	0.11	0.91
7	-1.03	0.12	1.05
8	0.42	0.11	0.90
9[N]	-0.39	0.12	0.88
10	0.64	0.11	0.81
11[N]	-1.40	0.13	1.06
12	0.41	0.11	1.06
M	0.00	0.11	0.99
SD	0.77	0.01	0.16

Note. [N] non-native rater; Reliability: 0.98; Separation: 6.76; Fixed chi-square: 491.6 (*d.f.*=11; $p < .00$)

Task measurement report. All tasks' characteristics are presented in Table 15. Each task differed in its difficulty based on the separation index of 9.22 with a high reliability of .99, supported by the statistical significance (χ^2=683.6, *df*=7, *p*< .00). What is noteworthy is a *preference* (monologic) task was the most difficult with the highest logit value of 1.60,while the least difficult was a *choice* (monologic) task. Thus, no distinct patterns emerged between the open role-play and monologic tasks in terms of the difficulty. Perhaps, the task difficulty is influenced by the content of the task rather than the task format. Additionally, two choices were given in the *choice* task, but no options were given in the *preference* task resulting some examinees' minimal responses within a 45 second response time, which can explain the higher degree of difficulty in the latter task. All tasks' infit values are within the range of 0.76~1.24 1±(2×0.12), indicating that all tasks contribute to measuring one construct stably.

Table 15
Measurement Report for All Tasks

Task	Difficulty Logit	Model Error	Infit Mean Square
1. Request a recommendation letter [a]	0.88	0.09	0.93
2. Request time to discuss a project [a]	-0.45	0.10	1.10
3. Refusal to professor's request [a]	-0.40	0.10	1.23
4. Meeting Time [b]	-0.01	0.10	1.02
5. Meeting Mode [b]	-0.16	0.10	0.96
6. Preference [c]	1.60	0.09	0.87
7. Choice [c]	-1.16	0.08	0.85
8. Giving comments [d]	-0.31	0.08	0.99
M	0.00	0.09	1.00
SD	0.86	0.01	0.12

Note. [a] Role-play with professor; [b] Role-play with classmate; [c] Monologic speaking task; [d] Monologic pragmatic task; Reliability = 0.99; Separation Index: 9.22; Fixed chi-square: 683.6 (*d.f.*=7; *p* < .00)

Summary of Three FACETS Analyses

The open role-play tasks reliably differed in their levels of difficulty. The role-play task with a professor was more difficult than the role-play task with a classmate, which can be explained by the higher degree of imposition (e.g., *request a letter*) and the higher power (e.g., *professor*) involved in the role-play task with a professor. However, a noticeable difficulty difference did not appear across task types (role-play and monologic tasks). The open role-play tasks contributed to generating enough variance to differentiate the 102 examinees' L2 pragmatic abilities.

In terms of the raters' performances, generally speaking, the raters showed more stable performances in rating the open role-play tasks, as evidenced by a fewer number of iterations, the smaller error values, and the stable fit statistics compared to those from monologic tasks. Potentially, this result is due to differences in the rating criteria of the two task types. Specific examples and evidence described in the rating criteria for the role-play tasks might have helped the raters' performances, which was not the case for the rating criteria for the two monologic general speaking tasks. Furthermore, all five rating categories for the role-play tasks differed in their degrees of difficulty and were stable in terms of the fit statistics, indicating that they contribute to measuring the construct of L2 pragmatic competence. In particular, the two rating categories (i.e., *Engaging with Interaction*, *Turn Organization*) that tapped into interactional competence in the role-play performance were found to be distinct in terms of it difficulty. This finding is noteworthy as it supports the CA findings and indicates that each aspect deserves separate attentions in operationalizing interactional competence.

Different degrees of raters' severity were found across task types. Figure 5 illustrates differences in the raters' severity across task types. Higher logit values indicate more severe raters. Many raters showed similar levels of severity for both role-play and monologic tasks, such as Rater 1, Rater 2, Rater 4, Rater 6, Rater 9, and Rater 12. A few raters, however, distinctly differed in their degrees of severity. For example, Rater 5 and Rater 10 displayed more severity in the role-play tasks compared to monologic tasks. Rater 11 was lenient in general, but more lenient in rating the monologic tasks.

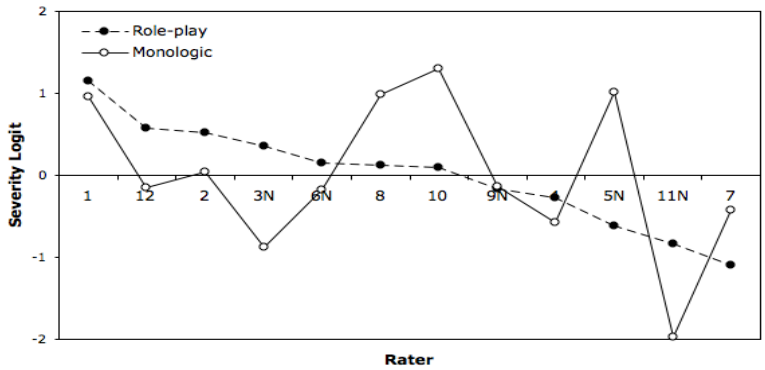


Figure 5. Raters' severity logits across task types

In terms of the raters' consistency in their ratings, as evidenced by the fit statistics, no rater showed misfit or overfit in rating the role-play tasks. However, two raters showed slight misfit and overfit respectively in monologic tasks, which potentially resulted in a slight overfit in one rater (Rater 5) in the results of all tasks combined. Figure 6 illustrates the fit statistics across task types.

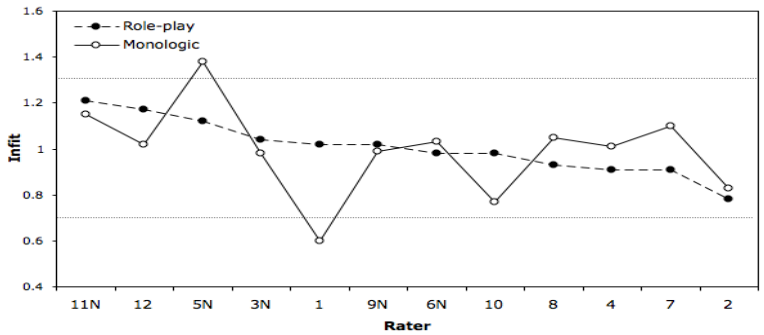


Figure 6. Raters' infit statistics across task types

Regarding the comparability of four rater groups' severity, Table 16 and Figure 7 present results. As a fully-crossed rating design was not employed, ensuring the comparability among rater groups is an important issue. Ideally, the averaged severity logit for each rater group needs to be similar. However, as seen in Figure 7, Rater Group 1 for the open role-play tasks showed a higher degree of severity, while other rater groups were relatively similar.

Table 16
Averaged Severity Logit for Four Rater Groups

Rater Group	Role-play	Monologic	All task combined
1 (R1-R3)	0.68	0.04	0.67
2 (R4-R6)	-0.25	0.09	-0.22
3 (R7-R9)	0.38	0.14	-0.33
4 (R10-R11)	-0.05	-0.27	-0.12

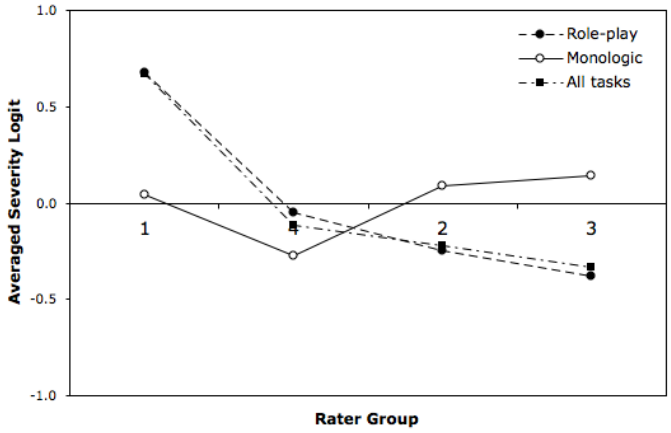


Figure 7. Raters' severity logits across rater groups

Correlation Analyses

In order to gather evidence for the *extrapolation* inference, which is investigated with Research Question 4 *To what extent are performances on open role-play tasks attributed to a construct of L2 pragmatic and L2 proficiency?*, three *Pearson* correlation

analyses were conducted. The first analysis compares the examinees' abilities logit values from open role-play tasks to their abilities logits from all monologic tasks combined. Next, in order to further investigate this relationship, examinees' role-play performances were compared with examinees' abilities logits from two monologic general speaking tasks only and those from monologic pragmatic task only respectively. For these correlation analyses, in addition to the main three FACETS analyses, two additional FACETS analyses, for the two monologic speaking tasks only and for the monologic pragmatic task only, were conducted to obtain the examinees' abilities logits separately.

Table 17 reports the correlation values. Interestingly, a less strong degree of relationship between open role-play tasks and monologic pragmatic task was observed (r = .81, r^2 = .66) compared to other two relationships (r = .90, r^2 = .81), respectively. Figures 8, 9, and 10 further demonstrate each of the three relationships. Some outliers in Figure 10, which examines the relationship between open role-play and monologic pragmatic tasks, are noticeable.

Table 17
Correlations across Tasks

	All Monologic Tasks	General Speaking Tasks	Monologic Pragmatic
Open Role-Play	0.90	0.90	0.81

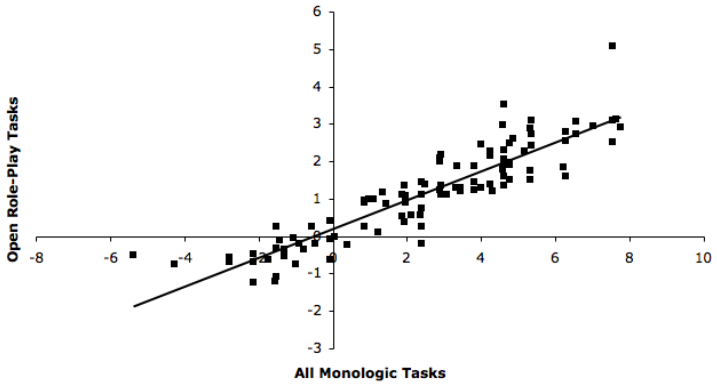


Figure 8. Relationship between open role-play tasks and all monologic tasks combined

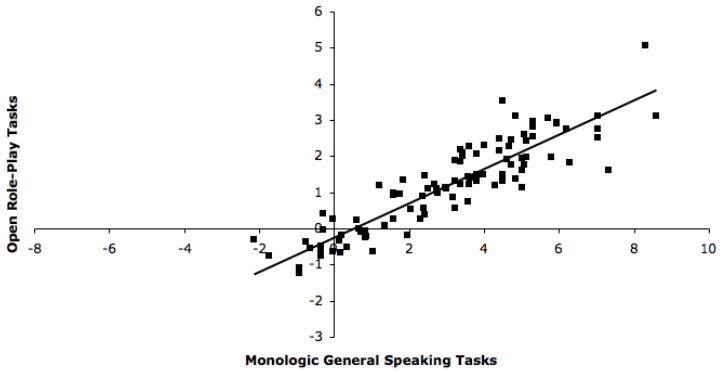


Figure 9. Relationship between open role-play and two monologic general speaking tasks

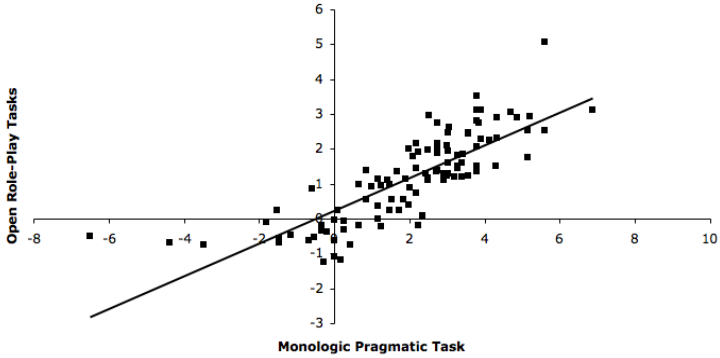


Figure 10. Relationship between open role-play tasks and monologic pragmatic task

CHAPTER 6

DISCUSSION

This section discusses the research findings in relation to the aforementioned research questions, focusing on the two types of arguments to discuss validity, *interpretive argument* and *validity argument*. Following Kane's (2006) argument-based approach to validity, as the first step, an *interpretive argument* specifies the proposed score uses and interpretations by laying out a chain of *inferences* and *assumptions*. Secondly, a *validity argument* provides an evaluation of the interpretive argument's coherence and plausibility. Backing for the assumptions that underlie the interpretive argument in this study provide the basis for the validity argument. Various inferences are inevitably entailed in language assessment when scores are derived from test performance and used for making decisions based on examinees' scores. As overviewed in Chapter 4, the chain of inferences along with the research questions guided the research steps and types of data collected in this study. Then, as Chapelle (2008) emphasizes, the data generated from the various inferences and steps involved in the present study needs to be then woven into a single account of the validity argument of the assessment tasks and the rating criteria, which is the focus of this chapter.

The *intended uses* of the open role-play tasks and interaction-sensitive data-driven rating criteria in this study were either for diagnostic or achievement tests in classroom assessments, in order to teach the L2 pragmatic abilities needed in an EAP context and, ultimately, to promote L2 pragmatic learning for ESL learners. In order to fulfill such uses, the score interpretations need to be meaningful for stakeholders. Namely, when a

score of '3' is awarded to an examinee, stakeholders need to be able to interpret the score in a meaningful way, including information on what types of pragmatics-related tasks examinees are successfully capable of performing or detailed diagnostic information on their performances. Thus, the meaning and relevance of the role-play task situations for stakeholders along with the analytical rating criteria with detailed descriptions are very important to satisfy the tasks' intended uses and their meaningful score interpretations.

Using Kane's (2006) argument-based approach to validity in assessment, four different types of inferences, *domain description*, *evaluation*, *generalization*, and *extrapolation*, were focused in this study. Each inference is related to each research question, which serve as the titles for the subsequent sections. Kane's validity argument framework is based on Toulmin's (2003) argument structure, consisting of various elements including *grounds*, *claims*, *warrants*, *backings*, and *rebuttals*. These elements are used to construct a chain of reasoning to build a particular conclusion based on the accuracy and relevance of a person's observations. Therefore, in order to draw valid conclusions about an examinee's ability, each inference is discussed by constructing a chain of reasoning and by establishing links between each of the argument elements. At the end of the chapter, the trustworthiness of mixed methods employed in the validity argument is discussed, which is addressed by the last research question.

Target Domain Description

Research Question 1*: What kinds of open role-play tasks can be developed to assess EAP L2 pragmatics in interaction? To what extent do the role-play tasks reflect*

stakeholders' pragmatic learning needs and measure L2 pragmatics in interaction in a valid way?

The first research question addresses the *target domain description* inference, which essentially links the examinees' assessment task performances in the target domain to their observed performances in the test domain. The observed performances in the assessment tasks need to reflect authentic language use in real life to claim meaningfulness to stakeholders. This inference is based on the *warrant* that the observed performances in the open role-play tasks reveal L2 pragmatic knowledge, skills, and abilities relevant to EAP domains. An *assumption* underlying the warrant is that the open role-play tasks require important L2 pragmatic knowledge and represent L2 pragmatics-related EAP domains. The validity of this warrant rests on the *backing* for the *assumption*, which is discussed below.

Assumption: Relevance and importance of EAP pragmatic tasks for stakeholders. In order to link examinees' performances in the tasks to the target domain, namely performance in real life, the assessment tasks should represent real-life situations. Additionally, the tasks should require certain targeted L2 pragmatic abilities and, consequently, the observed task performance should be the result of the targeted ability. As the targeted assessment tasks and rating criteria in this study are intended for classroom assessment in an EAP context, the assessment tasks need to be relevant to consider learning needs of the stakeholders. As the backing for this assumption, a large-scale needs analysis on L2 pragmatic learning needs in an EAP context was first conducted (Youn, 2010) and the findings from the needs analysis informed the development of the open role-play tasks. In order to ensure the relevance and importance

of each EAP pragmatic learning situation included in the questionnaire, qualitative interviews with stakeholders (students, instructors, and administrators) were conducted along with a literature review on L2 pragmatics. Twenty pragmatics-related situations were identified and a relatively high average score for each situation was found, indicating students' high EAP pragmatic learning needs. More details about the needs analysis findings and analyses are discussed in Youn (2010). The findings indicate that important interlocutors for the stakeholders in an EAP context were either professors or classmates and that students displayed great interests in learning the following situations: *making appropriate requests and refusals, writing e-mails, and how to appropriately make agreements/disagreements during discussion*. As the focus of the study was to measure L2 pragmatic performance in *interaction*, the role-play task format was chosen over a discourse completion task format due to the lack of interactivity inherent in the latter. Figure 11, adapted from Chapelle, Enright, and Jamieson (2007), illustrates the internal logic of the *domain description* inference, focusing on how backing (i.e., needs analysis results) supports the assumption that underlies the warrant for the *domain description* inference, leading to a claim that the observed performances provide the evidence of the targeted language abilities. Here, the validity of the warrant rests on the backing for the assumption that underlies it. Ensuring the *domain description* inference allows one to move onto the next inference, *evaluation*, and the claim made in the *target domain description* inference serves as the grounds for the next inference.

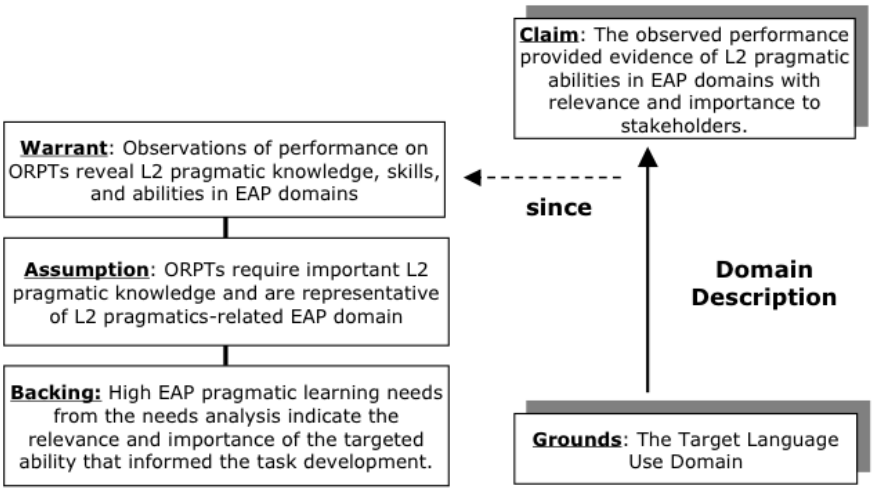


Figure 11. Domain description inference with an assumption and backing

Evaluation

Research Question 2*: What kinds of interactional features can be included in analytical rating criteria to assess examinees' L2 pragmatic performances in interaction? To what extent do conversation analysis findings of interaction data enhance the validity of the rating criteria?*

The *evaluation* inference describes how the observed scores accurately reflect the targeted language ability. This inference is supported by the warrant that observed performances in the open role-play tasks are evaluated to provide observed scores reflective of targeted L2 pragmatic abilities in interaction. Three different assumptions underlying the *evaluation* inference include (a) the open role-play tasks are designed to balance both authenticity and standardization, (b) the interaction-sensitive rating criteria include appropriate interactional features that can be objectively measurable, and (c) the raters received appropriate rater training to score the targeted performances. These are all important assumptions in investigating how well the targeted pragmatic performances are

evaluated to provide appropriate evidence of the targeted language abilities in relation to the intended uses and meaningful score interpretations. Figure 12 illustrates the relationships among the various assumptions and backings underlying the warrant, which lead to the conclusion that the observed performances were well evaluated and reflected important features of the targeted abilities. Details on how each assumption is supported by the backings are discussed below.

Assumption 1: Task design to balance both authenticity and standardization. Role-play tasks have been popularly used in measuring L2 pragmatic performances. However, the main criticism toward role-play tasks lie in their design, especially in their fixed interactional outcomes. The task design imposed certain interactional outcomes that were both known to all interlocutors, which jeopardizes authenticity and makes it questionable whether such a task can elicit authentic performance (Kasper & Rose, 2002). In other words, because one interlocutor already knows what another will say in certain situations, the task is somewhat contrary to conversations in real life. Thus, this kind of task design does not ensure meaningful targeted performance that reflect language use in real life, which is a threat to the *evaluation* inference.

Addressing such criticism, the open role-play tasks in the present study did not specify interactional outcomes for interlocutors but allowed them to choose interactional outcomes according to their personal preference, as in normal conversations. However, it should be noted that the goal of authenticity is hard to achieve, as the role-play task setting still invites certain degrees of inauthenticity. Nonetheless, the backing for this assumption includes the role-play cards given to each interlocutor, which include different contents. Additionally, certain information provided on the card was not shared

with an examinee to guarantee meaningful language performance. For example, for the role-play with professor, an examinee did not know the professor's schedule constraints when requesting a recommendation letter or an examinee did not expect the professor's request of rescheduling a presentation schedule (see Appendix B). For the role-play with a classmate, each examinee was given a different weekly schedule to naturally engage in deciding a good time to meet or were asked to choose meeting mode that she or he preferred, either face-to-face or through an online chat (see Appendix B). Endorsing the goals behind the role-play cards, many raters during the post-rating interviews agreed with the fact that most of the examinees, except for those with quite low language proficiency, seem to genuinely engage in the role-play conversations.

Assumption 2: Interaction-sensitive rating criteria. Well-designed rating criteria critically determine the consistency within and between raters and the meaningfulness of the observed scores. An important goal in this study was to develop rating criteria that tap into context-specific interactional competence as part of the targeted pragmatic abilities. The assumption of appropriate rating criteria is supported by the extensive efforts behind the development of the rating criteria. CA, one of the most systematic analytic programs in investigating spoken interaction, was employed to qualitatively analyze examinees' interactional features along with other task-specific features. Based on the detailed analyses, emerging patterns with concrete evidence and objectively measurable interactional features were identified across different role-play tasks and three different levels pertaining to examinees' L2 pragmatic performances in interaction, which then informed the detailed descriptions of each category. Although not all examinees' performances were examined, the predictable patterns were found in terms

of sequential organization, topic development, and language use due to the use of the role-play cards. In particular, CA findings on examinees' performances revealed that two different aspects, *Engaging with Interaction* and *Turn Organization*, did not go together especially for intermediate and low-intermediate examinees. Thus, such aspects deserve separate attention not only in developing the rating criteria and but also training the raters, and ultimately for teachers and students to pay separate attentions on them, considering the intended uses of all materials being classroom assessment. This observation becomes the rationale for including two different rating criteria, *Engaging with Interaction* and *Turn Organization*, to particularly tap into interactional competence. CA also enabled the researcher to identify diverse linguistic features used in delivering pragmatic meanings across different levels. Regarding the distinctness of each category, the raters indicated during the post-rating interviews that they all independently evaluated the targeted performance based on each category and did not have a difficult time differentiating between categories. Quantitative findings also provide convincing backing for the assumption of the rating criteria was appropriate for this study's assessment tasks. Stable fit statistics were found for the role-play rating criteria facet from FACETS analysis, indicating that the raters used them consistently and that the rating criteria contributed to measuring the targeted construct.

Assumption 3: Adequate rater training. Rater training is equally as critical as the well-designed rating criteria considering that the raters directly evaluate examinees' performances. Therefore, the assumption that the raters were adequately trained to accurately evaluate the targeted L2 pragmatic performances in interaction is critical in the *evaluation* inference. Backing for this assumption includes the extensive training sessions

along with diverse rating materials, such as example examinee performances and CA transcripts. Considering the complexity involved in examinees' performances for each role-play task, three separate training sessions, two for the two role-play tasks and one for the monologic tasks, were administered, as presented in Chapter 4. Focusing on three sequential steps, *familiarization*, *norming*, and *practice*, detailed explanations and materials were provided to each rater for each task type (see Appendix H for training materials). During the post-rating interviews, the raters mentioned that example performances for different levels discussed during the training and the CA transcripts used to exemplify the lack of interactional competence were greatly useful.

The various types of backing discussed above serve as validity evidence supporting the various assumptions underlying the *evaluation* inference as illustrated in Figure 12. Such extensive efforts made to generate the backings allow one to claim that the observed performances were accurately evaluated and reflected the important features of L2 pragmatic performance in interaction, thereby providing the foundation needed to continue to the next inference, *generalization*.

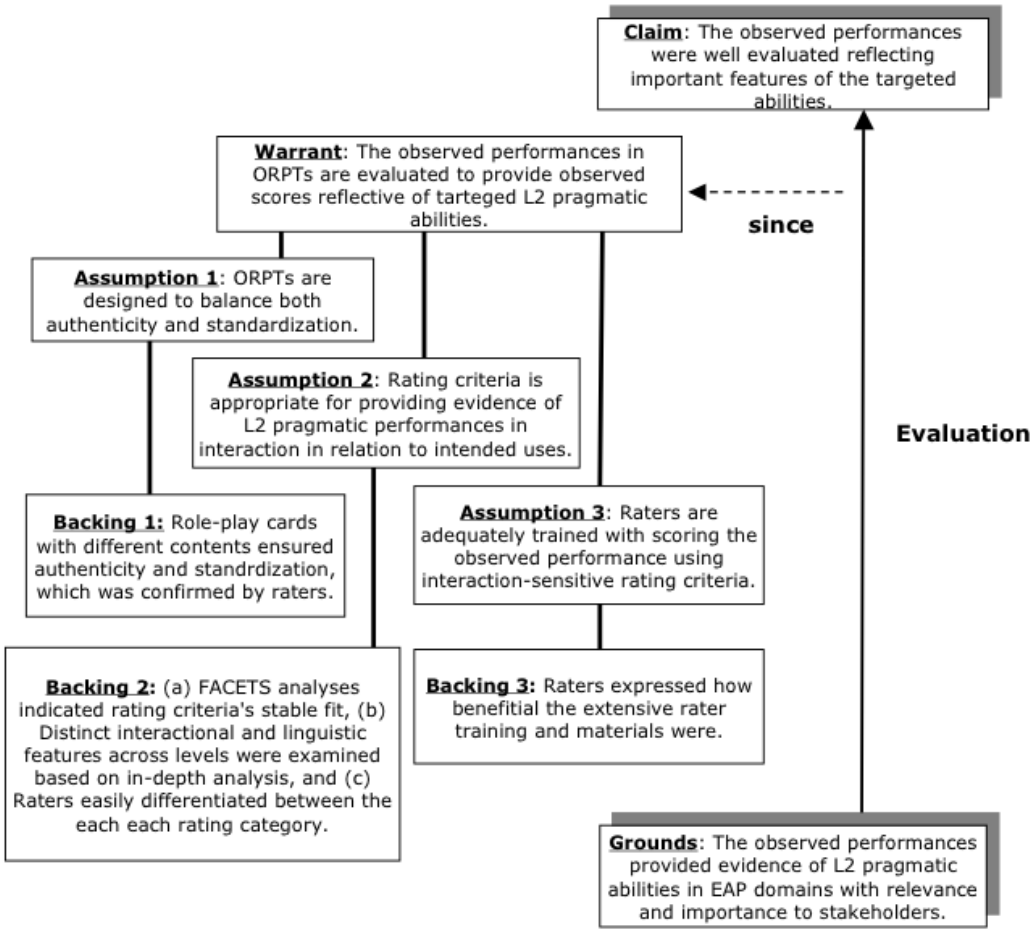


Figure 12. Evaluation inference with assumptions and backing

Generalization

Research Question 3*: How reliable is the rating process between and within raters' performances on rating the open role-play tasks using the data-driven interaction-sensitive rating criteria?*

The claim validated from the previous inference, *evaluation*, serves as the grounds for the next inference, *generalization*. As the observed scores are the result of the accurate evaluation of the targeted abilities, the *generalization* inference links the scores assigned to an examinee's performance in tasks and the same examinee's expected scores

on similar tasks under similar conditions. Thus, generalization is based on the warrant that observed scores are estimates of expected scores across raters and tasks. Underlying this inference is the assumption that two open role-play tasks provided stable estimates to measure the 102 examinees' pragmatic abilities and that the raters consistently used the rating criteria across examinees. As discussed below, in order to ensure the validity of this inference, assumptions on (a) examinees' performances on the tasks and (b) raters' performances are needed, and FACETS analyses provided the relevant backing.

Assumption 1: The role of the open role-play tasks. It is important to assume that the open role-play tasks provided stable estimates of the 102 examinees' pragmatic abilities in interaction in order to claim whether the examinees would receive similar scores on similar role-play tasks under similar task administration conditions. The backing for this assumption is found from FACETS analyses on examinees' abilities and task difficulty. A wide distribution of 102 examinees' pragmatic abilities measured by two open role-play tasks, consisting of five sub-situations, was reported with high reliability and statistical significance. Each task had a different level of difficulty, indicating that each task contributed varying degrees of variance to differentiate the examinees' abilities. Equally important, stable fit statistics were observed for all role-play tasks, which supports the important assumption behind the FACETS analysis that the tasks contribute to measuring one targeted construct. This support validates the claim that examinees would show similar performances in similar tasks and under similar rating conditions.

Assumption 2: Raters' performance. The assumption that the raters employed the rating criteria consistently across examinees and across tasks is also critical to the

generalization inference. Backing for this assumption includes various quantitative findings from FACETS analyses, particularly on the raters' performances. Although the raters showed varying degrees of severity, stable fit statistics for all raters, especially for the open role-play tasks, indicate that raters were internally consistent and that none exhibited unpredictable behavior. Although the 12 raters consisted of native and non-native English speakers, their different L1s did not seem to affect neither their internal consistency nor their severity. Additionally, some patterns emerged regarding variations within the raters' severity across task types.

However, two out of the 12 raters, one native and one non-native English speaker, showed unpredictability and lack of variation respectively when scoring examinees' performances in the monologic tasks. The degree of misfit and overfit was not serious, and the fact that those two raters showed internal consistency for the open role-play tasks indicates that the source of misfit was not due to raters themselves, but due to other factors, such as a lack of details in the rating criteria for the monologic speaking tasks. Additionally, some patterns emerged regarding variations within raters' severity across task types. For example, one rater was lenient in evaluating the examinees' performances in the role-play tasks while being more severe in his or her evaluations of the monologic tasks. This issue necessitates further analyses as it not only weakens the *generalization* inference and but also influences the next inference, *extrapolation*, which investigates the relationship between the performances in the role-play and monologic tasks.

Figure 13 illustrates the chain of reasoning that lead to this study's *generalization* inference and which provide the foundation for the next inference, *extrapolation*. The backing supports two assumptions of the warrant that observed scores are estimates of the

expected scores across raters and tasks and the claim that examinees would receive similar scores under similar rating conditions is made, which provides the foundations for the next inference, *extrapolation*.

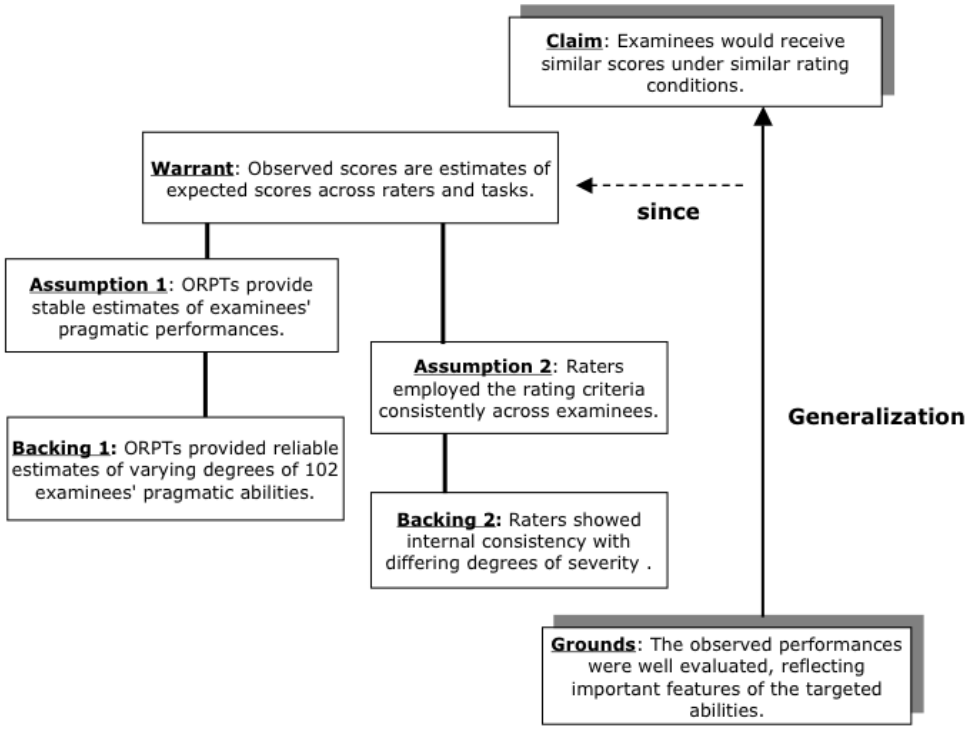


Figure 13. Generalization inference with assumptions and backing

Extrapolation

Research Question 4*: To what extent are performances on open role-play tasks attributed to a construct of L2 pragmatics and L2 proficiency?*

Based on the grounds that examinees would receive consistent scores under similar rating conditions obtained from the *generalization* inference, the last inference, *extrapolation*, seeks to claim that the consistently expected scores from the target tasks contribute to other parts of language proficiency. The *extrapolation* inference is supported by the warrant that expected scores are attributed to a larger construct of L2

pragmatics and L2 proficiency. Two assumptions underlying this warrant include the notion that L2 pragmatic performance in interaction is related to (a) other measures of L2 pragmatic performance and (b) other criteria of speaking proficiency.

Assumption 1: L2 pragmatic performance in interaction and monologic pragmatic performance. The assumption that the L2 pragmatic abilities being measured in this study are related to different types of L2 pragmatic performance can be supported a relationship between the targeted L2 pragmatic abilities and another measure of L2 pragmatics. As the backing for this assumption, examinees also completed a monologic pragmatic task (*giving oral comments on an e-mail written by a classmate*, see Appendix C). Each examinee's performance on the open role-play tasks and monologic pragmatic task were compared based on the examinee logits from separate FACETS analyses, presented in Chapter 5. A moderate degree of relationship (r =.81) was found, indicating about 64% of the variance in the open role-play task performances, is explained by the examinees' performance in monologic pragmatic task.

Assumption 2: L2 pragmatic performance in interaction and general speaking proficiency. Considering that pragmatic competence is part of language competence (e.g., Bachman, 1990; Bachman & Palmer, 1996), this study assumed that the L2 pragmatic abilities being measured are related to other measures of language competence. As the backing for this assumption, the examinees also completed the two additional monologic tasks (*preference* and *choice* tasks, see Appendix B) adapted from TOEFL iBT speaking tasks, which measure general speaking proficiency. A fairly high degree of relationship (r=.90) between the performances in the open role-play tasks and general monologic speaking tasks was found.

It is interesting to note that the stronger relationship was observed with general speaking proficiency (r=.90), rather than with the monologic pragmatic task (r=.81). This distinction might be due to the examinees' specific knowledge on the written genre (i.e., e-mail) involved in the monologic pragmatic task, as examinees needed to employ not only their L2 pragmatic knowledge in giving constructive comments but also genre-specific knowledge of writing an appropriate e-mail.

Figure 14 presents the chain of reasoning that lead to the *extrapolation* inference, which provides the claim that that L2 pragmatic performance in interaction is related to other criteria of language proficiency. The backing supports the two assumptions of the warrant that L2 pragmatic performance in interaction is related to both monologic pragmatic performance and general speaking proficiency. Table 18 provides a summary of backing, validity evidence, for each inference.

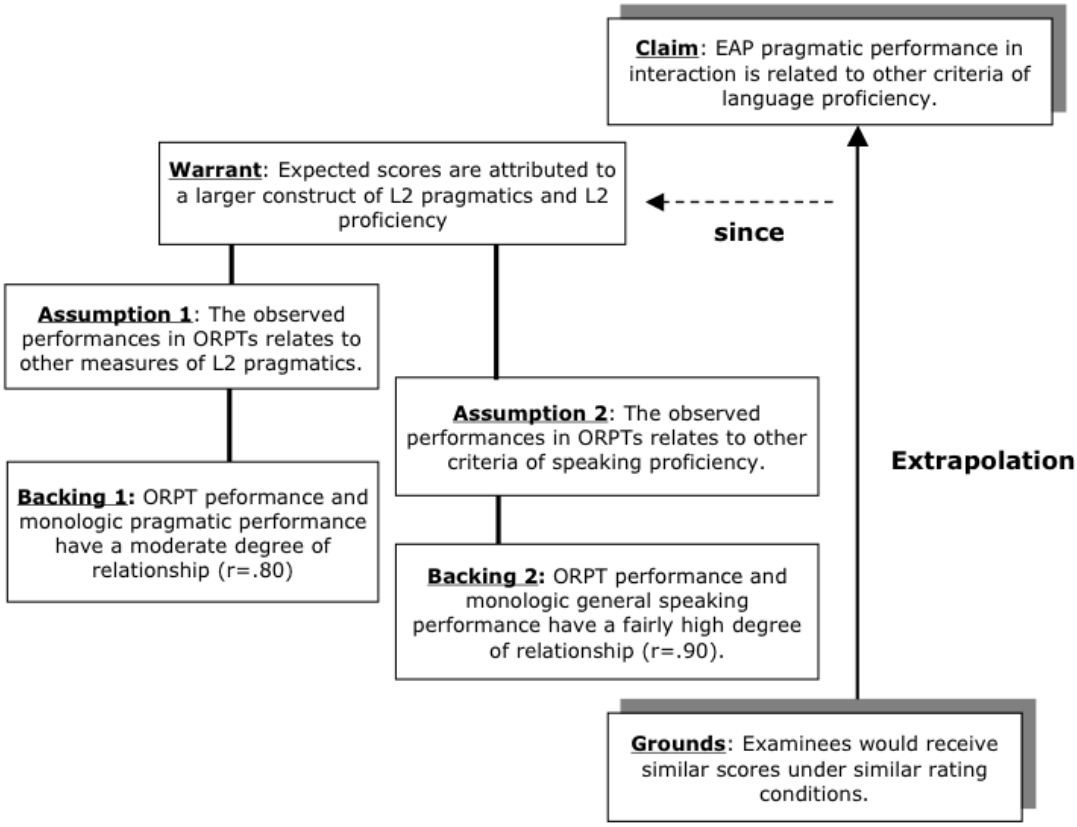


Figure 14. Extrapolation inference with assumptions and backing

Table 18
Summary of Validity Argument

Inference	Warrant Supporting the Inference	Backing for Inference
Target Domain Description	Observations of performances in ORPTs (open role-play tasks) reveal L2 pragmatic knowledge, skills, and abilities in an EAP context.	• Questionnaire items represented diverse L2 pragmatic learning tasks, developed based on qualitative interviews and literature on L2 pragmatics. • High needs in EAP pragmatic learning were found from a conducted needs analysis, which informed the development of the ORPTs. • High pragmatic learning needs in the interactions with aprofessor and classmate confirm the relevance and importance of ORPTs for stakeholders in an EAP setting.
Evaluation	Observations of performances in ORPTs provide observed scores that accurately reflect targeted L2 pragmatic abilities.	• A systematic task design using role-play cards ensured authenticity and standardization. Post-rating interviews also revealed that many raters evaluated that examinees' performances were also authentic and standardized. • Rating criteria was developed based on in-depth analyses of examinees' performances across different levels using CA which helped to identify interactional and linguistic features. Each rating category showed stable fit statistics quantitatively. • ORPTs showed varying levels of difficulty with stable fit statistics indicating each task contributes to measuring the construct of L2 pragmatics in interaction.
Generalization	Observed scores are estimates of expected scores across raters and tasks.	• 12 raters showed internal consistency in rating ORPTs using the interaction-sensitive rating criteria evidenced by stable fit statistics. • ORPTs provided reliable estimates of varying degrees of the 102 examinees' pragmatic abilities.
Extrapolation	Expected scores are attributed to a larger construct of L2 pragmatics and L2 speaking proficiency.	• A moderate degree of relationship between ORPTs and the monologic pragmatic task (r =.81). • A fairly high degree of relationship between the ORPTs and general monologic speaking tasks (r =.90).

Mixed Methods for Validity Argument

Research Question 5*: To what extent are findings from mixed methods trustworthy and how do they help to strengthen the validity argument?*

Sequential mixed designs (Teddlie & Tashakkori, 2006) in which qualitative and quantitative research occur chronologically were employed in this study to provide trustworthy backing to strengthen the validity argument. The validity argument presented in previous sections mainly concerned the development of open role-play tasks and interaction-sensitive rating criteria in relation to the intended uses and score interpretation. However, a separate argument for how mixed methods were employed in the study is also needed. For this purpose, the term, *legitimation*, has been coined by Onwuegbuzie and Johnson (2006) in order to investigate the degree to which mixed methods provide legitimacy, fidelity, credibility, trustworthiness by integrating the research strengths of qualitative and quantitative analysis.

As Brown (2013) emphasizes, a simple combination of qualitative and quantitative methods in mixed methods research is not enough. Instead, it is important to focus on what can be learned from them together. Assessing the validity of findings is complex in mixed methods research, as it employs both characteristics of qualitative and quantitative methods and, due to the mixing of paradigms and methods, results in the issue of commensurability (Teddlie & Tashakkori, 2003). Toward this challenge, researchers such as Brown (2013) and Onwuegbuzie and Johnson (2006) suggest the key issues needed for *legitimation* in mixed methods. Among the various suggestions, the legitimitation types and techniques of *sample integration*, *weakness minimization*,

convergence, and *divergence,* were deemed relevant to and selected for the current study, summarized in Table 19.

Sample integration. *Sample integration* is a legitimation type that refers to the extent that both qualitative and quantitative sampling design ensures quality meta-inferences. The sequential mixed design, employed in this study, in which conclusions made from each research strand lead to the formulation of questions, data collection, and analysis for the next strand. Thus, the qualitative sample analyzed needs to represent a random subset of the quantitative sample. The targeted population focused on in the current study were ESL learners using EAP in a university setting. The data for the qualitative analysis, which was conducted using a CA framework, included a small sample of ESL learners in a university-level EAP context with varying degrees of L2 proficiency. For the quantitative strand of investigation, 102 international students in a university-level EAP context were recruited as examinees. I argue that both this study's qualitative and quantitative strands employed a proper sample that represents the targeted population. Thus, the sample integration was appropriate.

Weakness minimization. A mixed methods approach requires the researcher's careful evaluation of each method's weakness in order to compensate for them through the strengths in the other approach. Additionally, careful research design to fulfill each method's potential is crucial. One limitations of the CA approach used in this study is that the findings themselves are inadequate to make generalizations of how well they can be utilized in differentiating between varying degrees of L2 pragmatic abilities, which was not only an important goal in this study but also the crucial validity evidence. The quantitative method itself also has limitations in relation to the research purpose in this

study, as a FACETS analysis cannot provide empirical grounds in operationalizing interactional competence, another important research purpose and validity evidence. However, each strand's strength compensated for the weakness inherent in the other. CA offered conceptual and empirical insights in explicating and explaining the interactional and linguistic features ESL learners employed in delivering diverse pragmatic meanings. FACETS analysis provided the researcher with generalizable findings, as the validity evidence, regarding how stably each category in the rating criteria differed in terms of difficulty, how consistently the raters utilized the interaction-sensitive rating criteria, and how stably 102 examinees' L2 pragmatic abilities can be measured using the rating criteria.

Sequential. A *sequential* legitimation type is particularly relevant to the sequential mixed method design as the order of sequencing itself can be a threat to legitimation. This study employed first a qualitative method, followed by the quantitative strand. The careful evaluation of each method's strength and weakness in relation to the research purpose influenced the decision regarding the order in which the method was used. An important research goal in this study was to develop data-driven interaction-sensitive rating criteria that measured L2 pragmatics in interaction, ultimately contributing to meaningful score interpretations and usefulness in classroom assessment. Considering the importance of well-designed rating criteria to generate meaningful conclusions on the examinees' language abilities, qualitative analysis was first required to explicate the complex nature of L2 pragmatic performance in interaction, rather than quantitative analysis in light of the lack of concrete foundations in developing the rating criteria.

Convergence. This technique refers to how quantitative and qualitative strands yield multiple data sources and shows how they support similar conclusions. Although the first strand influences the next strand chronologically in the sequential mixed design, resulting in interdependence between the two strands, convergence is not necessarily guaranteed. In this study, multiple data sources from the two strands support similar conclusions about examinees' L2 pragmatic abilities. This study's CA findings on distinct interactional and linguistic features across the diverse levels of examinees indicate that such features can differentiate varying levels of L2 pragmatics in interaction for the assessment purposes. For example, advanced-level examinees often employed complex grammatical structures, such as bi-clausal structures, which were not shown in low-intermediate-level examinees' performances at all. Regarding the interactional features of the examinees' performances, advanced-level examinees displayed sensitivity to making appropriate turns for dispreferred actions by employing meaningful pauses, hedges, and their situational knowledge. Thus, such features were included in the analytical rating criteria to serve as explicit descriptions of examinees' different L2 pragmatic levels. FACETS analysis findings also supported such a conclusion. In particular, the raters were able to easily understand different characteristics across examinees' performance and used the rating criteria accordingly with internal consistency, evidenced by stable fit statistics in the rater and rating criteria facets. Furthermore, the two rating categories for interactional competence (i.e., *Engaging with Interaction*, *Turn Organization*) showed a different level of difficulty which also indicates they functioned differently in separating the examinees' L2 pragmatic abilities.

This finding demonstrates that the quantitative findings support the qualitative distinctness in the two categories from a CA perspective.

Divergence. While it may sound contrary to the *convergence* technique, the *divergence* technique examines any contradictions and unexpected results that do not lead to a study's conclusions or to further insights. As one of the benefits of doing mixed methods is to learn more than what each method alone can yield, the *divergence* technique can provide us with more interesting results. This study's CA findings show that the examinees employed diverse resources and strategies to deliver pragmatic meanings politely. Examinees' utilization of these difference resources is not easily assessed quantitatively, as it is difficulty to evaluate one delivery as better than another. For example, despite their lack of grammatical resources, low-level examinees employed an indirect request strategy by orienting to seeking the professor's advice, instead of making a direct request employing specific linguistic resources. Additionally, examinees' sensitivity toward EAP context-specific knowledge were not always expressed through linguistic resources, but were employed through prosodic resources, such as tone. Such findings were not included in the development of the rating criteria despite their relevance and importance in making inferences about examinee's L2 pragmatic abilities. Nonetheless, the CA findings provide rich details on how those with limited linguistic resources still succeed interactionally, emphasizing the crucial role of CA in this study and in understanding the construct of L2 pragmatics in *interaction*.

Table 19
Summary of Legitimation in Mixed Method

Types of Legitimation (Brown, 2013; Onwuegbuzie & Johnson, 2006)	**Ways in which this type of legitimation was addressed**
Sample Integration: Integrating qualitative and quantitative samples	• Data for CA was drawn from examinees with varying degrees of proficiency in an EAP context. • Quantitative findings were based on102 examinees who represent a population in an EAP context.
Weakness minimization: Compensate for the weakness in some approaches with the strengths of others	• CA findings themselves are not enough to make generalizations of how well they differentiated varying degrees of examinees' interactional competence and how well raters employed the rating criteria. FACETS analysis itself cannot provide empirical grounds in operationalizing interactional competence. • Concepts and terminologies in CA literature and findings helped to provide structure in analyzing what was lacking in examinees' performance during the role-play tasks. • Stable fit statistics from FACETS analysis indicate that raters successfully differentiated between varying levels of interactional features following descriptions and concrete examples included in the rating criteria.
Sequential: The extent to which a potential problem is minimized by reversing the sequence of the quantitative and qualitative phases.	• The main purpose of employing the mixed methods approach in this study was to develop the interaction-sensitive rating criteria. As the quantitative method itself does not inform us *how* examinees display their interactional competence, starting with a qualitative phase to identify the interactional features serves the research purpose by minimizing the weakness of the quantitative phase.
Convergence: Bring multiple data sources together and show how similarly they support conclusions	• Distinct differences in language use and interactional features across different levels of examinees were shown from a CA perspective. • Distinct difficulty levels shown for the two rating categories for interactional competence support the CA findings. • Raters easily understood the qualitative evidence in both the interactional and linguistic features as found in the qualitative analyses, and were able to differentiate between examinees' varying abilities based on such understandings.

Divergence: Examine contradictions and anomalies to see if they do not lead to conclusions of their own or to further fruitful inquiries	• Not all interactional features identified from CA were included in the rating criteria due to the difficulty in giving weight to certain interactional features and politeness strategies. For example, various strategies utilized in delivering pragmatic meaning were not reflected in the rating criteria as it is difficult to judge one strategy as better than another, and the absence of such strategies alone cannot be penalized.

CHAPTER 7

CONCLUSION

In an effort to address the three research gaps from the perspectives of theory, methodology, and the narrow intended uses in previous L2 pragmatic assessment research, the present study investigated the validity for task-based assessment of L2 pragmatics in *interaction* in an EAP setting for classroom assessment. Following Kane's (2006) argument-based approach to validity, score interpretations from L2 pragmatic performance were based on building blocks of interpretive arguments composed of the four types of inferences and assumptions. Using the sequential mixed methods design (Greene, 2007; Tashakkori & Teddlie, 2003), qualitative and quantitative data was collected to use them as validity evidence to strengthen the validity argument. The findings from the large-scale needs analysis on stakeholders' L2 pragmatic learning needs ensured the relevance and meaningfulness of the two role-play tasks. Instead of following predetermined interactional outcomes, participants were asked to negotiate and interact naturally in the role-plays tasks with two interlocutors (a student and a professor). At the same time, standardization across examinees was ensured using the role-play cards.

A total of 102 adult ESL examinees completed the open role-play tasks along with monologic tasks. Four rater groups, consisting of 12 raters in total, scored each examinee's performance in the assessment tasks. Additionally, CA-informed analytical rating criteria (Fulcher, Davidson, & Kemp, 2011; Turner & Upshur, 2002) was developed to ensure the accuracy of the observed score and meaningful score

interpretations. As an empirical basis, the various features of examinees' example performances in the role-play tasks were analyzed using a CA perspective, which then informed the development of the rating criteria. As quantitative evidence, a many-facet Rasch measurement using FACETS was employed to investigate the examinees' abilities, raters' severity, task difficulty, and rating criteria's quality. The two open role-play tasks, consisting of five sub-situations, displayed different levels of difficulty with the stable infit statistics, indicating all tasks contributed to measuring one construct, namely L2 pragmatic competence. Furthermore, the role-play tasks reliably separated varying degrees of 102 examinees' L2 pragmatic abilities. The 12 raters showed varying degrees of severity, but no rater showed unpredictable rating patterns in rating the examinees' L2 pragmatic performance in the open role-play tasks using the interaction-sensitive rating criteria, although a few raters were a bit unpredictable in rating monologic tasks. The validity evidence gathered in the various stages of the study served as backing for a serious of inferences, warrants, and assumptions and they were woven into a single validity narrative that evaluates the plausibility of the interpretive argument.

Limitations of the Study

There are several limitations in this current study. The first limitation concerns interlocutors' performances. Due to logistic reasons, four interlocutors were recruited as the professor role in the role-play task. Although they received training to ensure the standardization, some variations existed across a few interlocutors in terms of conversational styles. For example, one interlocutor who had completed the role-play task with low-intermediate level examinees used frequent clarification questions, which

in a way is natural to engage in the conversation to make sure whether the low-intermediate level examines are following the conversation. However, such frequent clarification questions may have functioned as unintentional help or sources of confusions for some examinees. The raters also reported, during the post-rating interviews, that they noticed the variations across the interlocutor's conversational styles. However, the raters did not think that such factor greatly influenced the scoring result. Nonetheless, interlocutor variations in role-play tasks deserve more attention to advance the research on assessing interactional competence.

Second, not all features of L2 pragmatics performances in the role-play tasks identified in this study were included in the rating criteria mainly due to the difficulty of operationalizing them for assessment purposes. The CA findings revealed that the examinees, despite their limited grammatical resources, utilized various strategies in order to achieve interactional goals, such as indirect request and orientation toward being a good student, which were still relevant and meaningful component of L2 pragmatic competence. However, they were not included in the rating criteria except for concrete evidence that are objectively comparable across the examinees to ensure rater consistency.

Third, this study employed the partially-crossed rating design. As anchored data, all 12 raters at least rated the 30 examinees' performances. It was a pragmatic decision to reduce the amount of resources. However, the partially-crossed rating design resulted in the issue of incomparability between the four rater groups, which potentially weakens the *generalization* inference. Although the model was found to stable along with the

acceptable infit statistics, the fully-crossed rating design would have provided statistically more robust results.

Implications

Despite the limitations, the current study's findings provide the following implications for language assessment. First, the raters used the interaction-sensitive rating criteria for the role-play tasks successfully in this study. The importance of well-developed rating criteria over raters' experience has been reported (Jang, 2010; Knoch, 2007, 2009; Xi & Mollaun, 2011). Previous research has shown that well-designed rating criteria with concrete details is very important to ensure rater consistency, and they were more important than raters' previous rating experience (Barkaoui, 2010). Also, regarding the use of rating criteria with thick descriptions, concerns have been raised including the way that raters perceive the rich descriptions and employ them to make reliable decisions. Nonetheless, the present study showed that the raters used the rating criteria with thick descriptions with internal consistency and also appreciated the extensive training.

Second, the present study exemplifies interventional CA (Antaki, 2011) particularly for the practices of measuring interactional competence in language testing research. CA was instrumental in this study in designing role-play tasks to elicit authentic performance, developing rating rubric to provide concrete evidence of interactional features, and training raters to assist them to award consistent scores on interactional competence. Previous research on interaction-involved speaking assessment often used general descriptions in rating criteria for interactional competence, such as *the conversation is generally smooth* without concrete evidence, which becomes a source of

unpredictability in raters' performances or in rating criteria's quality. Although not all features of sequential competence were included in the rating criteria developed in this study, key interactional features, such as turn organization and the degree of engaging with interaction, statistically functioned well in measuring interactional competence. The distinct difficulty in each category supports not only the independence but also raters' capability of differentiating them along with the adequate rater training.

Third, the present study exemplified Kane's validity argument in the context of assessing L2 pragmatics using the mixed methods approach. The main reason for employing Kane's validity argument in the study was its explicit guidance, an internal logic, and structures in investigating the complex nature of validity, as argued in Chapelle, Enright, and Jamieson (2010). Specifying warrants and assumptions for the appropriate inferences helped to evaluate the adequacy of the intended uses and meaningful score interpretations and provided a concrete guideline to design research steps and gather appropriate validity evidence. Mixed methods also greatly contributed in providing appropriate backing in addressing the inferences, that either quantitative or qualitative research method alone cannot provide in explicating the complex nature of L2 pragmatics in interaction.

In addition to the implications for language assessment at large, this study also provides implications for L2 pragmatic assessment in particular. Assessing L2 pragmatics can be a contentious issue considering variations across contexts and speakers. The fact that the present study specified the intended uses of the open role-play tasks and rating criteria, and focused on specific language use in an EAP context greatly helped to operationalize L2 pragmatic performance. Additionally, the careful design of the open

role-play tasks following the TABL framework along with the use of role-play cards helped to elicit relatively controlled performances rather than generate too much variation, which helped to ensure raters' internal consistency and develop the rating criteria accordingly.

Additionally, the findings of the study provided an empirical insight on the importance of grammar, namely pragmalinguistics, in L2 pragmatics. Among the five categories in the rating criteria, the *Language Use* category, which taps into appropriate grammatical expressions in delivering L2 pragmatic meanings was found to be the most difficult. This finding suggests that grammar can be further utilized in predicting L2 pragmatic performance and thus further attentions on the role of grammar in differentiating varying degrees of L2 pragmatic performance will merit future L2 pragmatic assessment research.

Suggestions for Future Research

The following areas can be investigated in the future. Firstly, further research on training interlocutors and controlling potential effects of peer examinees in role-play tasks or interaction-involved speaking tests is needed. Although interlocutors-involved speaking tests can elicit ample opportunities for examinees to interact with each other and to demonstrate diverse ranges of interactional features, issues on the degree of examinee's contributions to the co-constructed nature of interactional competence still remain under-researched and require further research. Although the research on interviewer variations during oral proficiency interviews (e.g., Brown, 2003; Kasper & Rose, 2003, 2007) has been available, little attention on the interlocutor variations for

role-play tasks has been made, which can satisfy increasing needs of measuring interactional competence in large-scale speaking assessment.

More investigations on how raters perceive L2 pragmatic performance will be useful to explicate rating decision processes and rater variations in L2 pragmatic assessment. Previous studies on raters' perceptions have been available using diverse methodological attempts including stimulated verbal recall (Brown, 2003, 2005) or retrospective verbal reports (Brown, Iwashita, & McNamara, 2005). But, these studies did not focus on L2 pragmatic performance. Along with qualitative methods, such quantitative methods as G-theory and FACETS bias analysis will also allow more systematic investigations on the issue of rater variations.

Lastly, the utilization inference on usefulness and value implications of target assessment tasks in contexts was not directly addressed in this study. Considering the intended uses of the assessment tasks and rating criteria as classroom assessment, backing for the assumption that the open role-play tasks and interaction-sensitive rating criteria are useful as instructional materials in EAP classrooms is important. Following the framework of assessment for learning (Shepard, 2000, 2006; Torrance & Pryor, 1998), future empirical studies on how stakeholders utilize the open role-play tasks and rating criteria, and how they contribute to promoting L2 pragmatic learning in an EAP context will be necessary.

REFERENCES

Ahn, R. C. (2005). *Five measures of interlanguage pragmatics in KFL (Korean as foreign language) learners.* Unpublished Ph.D. dissertation. University of Hawai‘i at Manoa, Hawai‘i.

Alcón Soler, E., & Martínez-Flor, A. (2005). Editors' introduction to pragmatics in instructed language learning. *System, 33*(3), 381–384.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Antaki, C. (Ed.) (2011). *Applied conversation analysis: Intervention and change in institutional talk*. UK: Palgrave Macmillan

Arundale, R. B. (1999). An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics, 9*, 119–153.

Arundale, R. B. (2005). Pragmatics, conversational implicature, and conversation. In K. L. Fitch & R. E. Sanders (Eds.), *Handbook of language and social interaction* (pp. 41–63). Mahwah, NJ: Lawrence Erbaum Associates.

Arundale, R. B. (2008). Against (Gricean) intentions at the heart of human interaction. *Intercultural Pragmatics, 5*(2), 229–258.

Atkinson, J., & Heritage, J. (Eds.). (1984). *Structures of social action*. Cambridge: Cambridge University Press.

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453–476.

Bachman, L. F. (2007). What is the construct?: The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–71). Ottawa, Ontario: University of Ottawa Press.

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

Bardovi-Harlig, K. (1999). Exploring the interlanguage of interlanguage pragmatics: A research agenda for acquisitional pragmatics. *Language Learning, 49*(4), 677–713.

Bardovi-Harlig, K. (2001). Empirical evidence of the need for instruction in pragmatics. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 13–32). New York: Cambridge University Press.

Bardovi-Harlig, K., & Dörnyei, J. (1998). Do language learners recognize pragmatic violations? Pragmatics vs. grammatical awareness in instructed L2 learning. *TESOL Quarterly, 32*, 233–259.

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

Bialystok, E. (1993). Symbolic representation and attentional control in pragmatic competence. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatis* (pp. 43–59). New York: Oxford University Press.

Bilmes, J. (1993). Ethnomethodology, culture, and implicature: Toward an empirical pragmatics. *Pragmatics, 3*(4), 387–409.

Brindley, G. (1994). Task-centered language assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: Papers presented at the Annual International Language in Education Conference, Hong Kong* (pp. 73–94). Hong Kong: Hong Kong Education Department.

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

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt am Main: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tests. TOEFL Monograph Series MS-29. Princeton: Educational Testing Service.

Brown, J. D. (2000). Observing pragmatics: Testing and data gathering techniques. *Pragmatics Matters, 1*(2), 5–6.

Brown, J. D. (2001). Pragmatics tests. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–325). Cambridge: Cambridge University.

Brown, J. D. (in press, 2013). *Research for TESOL: Quantitative, qualitative, and mixed method*. Endinburgh, UK: University of Edinburgh Press.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics*, *43*, 198–217.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W., J. (2002). *An investigation of second language task-based performance assessment* (Technical Report #24). Honolulu, HI: Second Language Teaching and Curriculum Center.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. New York: Cambridge University Press.

Byrnes, H. (2002). The role of task and task-based assessment in a content-oriented collegiate foreign language curriculum. *Language Testing*, *19*, 419–437.

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

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chalhoub-Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815–831). Mahwah, NJ: Lawrence Erlbaum.

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

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York & London: Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple…, *Language Testing*, *29*, 19–27.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*, 3–13.

Chapelle, C. A., Enright, M. E., & Jamieson, J. M. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Compare TOEFL® scores. (2012). Retrieved December 11, 2012, from https://www.ets.org/toefl/institutions/scores/compare.

Creswell, J. W., & Tashakkori, A. (2007). Editorial: Differing perspectives on mixed methods research. *Journal of Mixed Methods Research, 1*(4), 303–308.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.). *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crystal, D. (1997). *Dictionary of linguistics and phonetics* (4th ed.). Oxford: Blackwell.

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

D'hondt, S. (2009). The pragmatics of interaction: A survey. In S. D'hondt, J.-O. Östman & J. Verschueren (Eds.), *The pragmatics of interaction* (pp. 1–19). Amsterdam: John Bejamins.

Doughty, C., & Long, M. H. (2003). The scope of inquiry and goals of SLA. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 3–15). Malden, MA: Blackwell.

Drew, P. (2005). Conversation analysis. In K. L. Fitch & R. E. Sanders (Eds.), *Handbook of language and social interaction* (pp. 71–102). Mahwah, NJ: Lawrence Erlbaum.

Drew, P., & Holt, E. (1998). Figures of speech: Figurative expressions and the management of topic transition in conversation. *Language in Society*, *27*, 495–522.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing, 26*(3), 423–443.

Duranti, A. (2006). The social ontology of intentions. *Discourse Studies, 8*(1), 31–40.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Frankfurt: Peter Lang.

Edwards, D. (2006). Discourse, cognition and social practices: the rich surface of language and social interaction. *Discourse Studies, 8*(1), 41–49.

Edwards, D. (2008). Intentionality and means rea in police interrogations: The production of actions as crimes. *Intercultural Pragmatics, 5*(2), 177–199.

Farhady, H. (1980). Justification, development, and validation of functional language testing. Unpublished Unpublished Ph.D. dissertation, University of California, LA.

Firth, A., & Wagner, J. (1997). On discourse, communication, and (some) fundamental concepts in SLA research. *The Modern Language Journal, 81*(3), 258–300.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT*

Journal, *41*, 287–291.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction, Language Testing, *Language Testing*, *13*, 208–238.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*, 5–29.

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*(7), 4–10.

Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English.* Unpublished Ph.D. dissertation. Teachers College, Columbia University.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *2*, 89–119.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*, 585–602.

Gass, S. (1998). Apples and oranges: Or why apples are not oranges and don't need to be. *The Modern Language Journal, 82*, 83–90.

Gass, S. (2004). Conversation analysis and input-interaction. *The Modern Language Journal, 88*, 597–602.

Gardner, J. (2006). Assessment and learning: An introduction. In J. Gardner (Ed.), *Assessment and learning* (pp. 1–5). London: Sage.

Garfinkel, H. (1967). *Studies in ethnomethodology.* Englewood Cliffs, NJ: Prentice-Hall.

Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press.

Goffman, E. (1983). The interaction order. *American Sociological Review, 48*, 1–17.

Golato, A. (2003). Studying compliment responses: A comparision of DCTs and recordings of naturally occurring talk. *Applied Linguistics, 24*, 90–121.

Grabowski, K. C. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking.* Unpublished Ph.D. dissertation. Columbia University.

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

Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13*(1), 93–98.

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

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7–22.

Grice, H. P. (1957). Meaning. *Philosophical Review, 66*, 377–388.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Gysen, S., & Van Avermaet, P. (2005). Issues in functional language performance assessment: The case of the certificate Dutch as a foreign language, *Language Assessment Quarterly*, *2*, 51–68.

Hall, J. K. (1995). (Re)creating our worlds with words: A sociohistorical perspective of face-to-face interaction. *Applied Linguistics, 16*, 206–232.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage.

Härmälä, M. (2010). Linguistic, sociolinguistic, and pragmatic competence as criteria in assessing vocational language skills: The case of Finland. *Melbourne Papers in Language Testing*, *15*, 27–69.

Haugh, M. (2008). Intention in pragmatics. *Intercultural Pragmatics, 5*, 99–110.

Hellermann, J. (2007). The development of practices for action in classroom dyadic interaction: Focus on task openings. *The Modern Language Journal, 91*, 83–96.

Hellermann, J. (2008). *Social actions for classroom language learning*. Clevedon: Multilingual Matters.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In M. J. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 299–345). Cambridge: Cambridge University Press.

Heritage, J. (1998). Oh-prefaced responses to inquiry, *Language in Society*, *27*, 291–334.

Heritage, J. (1990/1). Intention, meaning and strategy: Observations on constraints on interaction analysis. *Research on Language and Social Interaction, 24*, 311–332.

Hudson, T. D. (2001a). Indicators for pragmatic instruction. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 283–300). Cambridge: Cambridge University.

Hudson, T. D. (2001b). Self-assessment methods in cross-cultural pragmatics. In T. D. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp. 57–74). Honolulu, HI: University of Hawai‘i Press.

Hudson, T. D., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics* (Technical Report #2). Honolulu, HI: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Hudson, T. D., Detmer, E., & Brown, J. D. (1995). *Developing prototype measures of cross-cultural pragmatics* (Technical Report #7). Honolulu, HI: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Hutchby, I., & Wooffitt, R. (1998). *Conversation Analysis*, Cambridge, UK: Polity Press.

Hymes, H. D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Middlesex, Harmondsworh: Penguin.

Ishida, M. (2006). Interactional competence and the use of modal expressions in decision-making activities: CA for understanding microgenesis of competence. In K. Bardovi-Harlig, C. Félix-Brasdefer & A. S. Omar (Eds.), *Pragmatics and language learning* (Vol. 11, pp. 55–79). Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai‘i.

Jacoby, S., & Ochs, E. (1995). Co-construction: An introduction. *Research on Language & Social Interaction, 28*(3), 171–183.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* Unpublished Ph.D. dissertation. University of Illinois at Urbana-Champaign.

Jang, E. E., McDougall, D. E., Pollon, D., Herbert, M., & Russell, P. (2008). Integrative mixed methods data analytic strategies in research on school success in challenging circumstances. *Journal of Mixed Methods Research, 2*(3), 221–247.

Jang, S. Y. (2010). *The development and evaluation of a systematic training program for increasing both rater reliability and rating accuracy*. Unpublished Ph.D. dissertation. University of Illinois at Urbana-Champaign

Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: A meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 165–211). Amsterdam: Benjamins.

Johnson, M. (2000). Interaction in the oral proficiency interview: Problems of validity. *Pragmatics, 10*(2), 215–231.

Johnson, M. (2001). *The art of non-conversation. A reexamination of the validity of the Oral Proficiency Interview*. New Haven, CT & London: Yale University Press.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133.

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

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 17–64). Westport, CT: Greenwood Publishing.

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

Kasper, G. (1997). 'A' stands for acquisition: A response to Firth and Wagner. *The Modern Language Journal, 57,* 307–312.

Kasper, G. (2004). Participant orientations in German conversation-for-learning. *The Modern Language Journal, 88*, 551–567.

Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K.

Bardovi-Harlig, C. Félix-Brasdefer & A. S. Omar (Eds.), *Pragmatics and language learning*, Vol. 11 (pp. 281–314). Honolulu, HI: Second Language Teaching and Curriculum Center, University of Hawai‘i.

Kasper, G. (2008). Discourse and socially shared cognition. In J. Cenoz & N. H. Hornberger (Eds.), *Encyclopedia of language and education, 2nd ed., Vol. 6: Knowledge about language* (pp. 59–78). New York: Springer.

Kasper, G. (2009). L2 pragmatic development. In W. C. Ritchie & T. K. Bhatia (Eds.), *New handbook of second language acquisition*. Leeds, UK: Emerald.

Kasper, G., & Blum-Kulka, S. (Eds.). (1993). *Interlanguage pragmatics*. New York: Oxford University Press.

Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Malden: Blackwell Publishing.

Kasper, G., & Ross, S. (2003). Repetition as a source of miscommunication in oral proficiency interviews. In J. House, G. Kasper & S. Ross (Eds.), *Misunderstanding in social life: Discourse approaches to problematic talk* (pp. 82–106). New York: Longman.

Kasper, G., & Ross, S. J. (2007). Multiple questions in oral proficiency interviews. *Journal of Pragmatics, 37*, 2045–2070.

Kasper, G., & Schmidt, R. (1996). Developmental issues in interlanguage pragmatics. *Studies in Second Language Acquisition, 18*, 149–169.

Kim, Y. (2009). The Korean discourse markers –nuntey and kuntey in native-nonnative conversation: An acquisitional perspective. In H. t. Nguyen & G. Kasper (Eds.),

Talk-in-Interaction: Multilingual perspective (pp. 317–350). Honolulu, HI: National Foreign Language Resource Center, University of Hawai‘i.

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

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

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

Larsen-Freeman, D. (2004). CA for SLA? It all depens... *The Modern Language Journal, 84*, 603–607.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Lee, Y.-J. (2005). Construct validation of an integrated, process-oriented, and computerized English for Academic Purposes (EAP) placement test: A mixed method approach. Unpublished Ph.D. Dissertation. University of Illinois at Urbana-Champaign.

Lee, Y.-A. (2006). Towards respecification of communicative competence: Condition of L2 instruction or its objective? *Applied Linguistics, 27*(3), 349–376.

Lee, Y.-J., & Greene, J. C. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*, 366–389.

Leech, G. (1983). *Principles of pragmatics*. London: Longman.

Levinson, S. (2006). Cognition at the heart of human interaction. *Discourse Studies, 8*(1), 85–93.

Nguyen, H. t., & Kasper, G. (Eds.). (2009). *Talk-in-interaction: Multilingual perspectives*. Honolulu, Hi: National Foreign Language Resource Center, University of Hawai‘i.

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

Linacre, J. M. (2006). Facets Rasch measurement computer program, version 3.61.0. Chicago: Winsteps.com.

Linacre, J. M. (2012, January). Many-facet Rasch measurement: Facet tutorial [PDF document]. Retrieved from http://www.winsteps.com/a/ftutorial4.pdf

Linacre, J. M., Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*, 484–509.

Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic controversies, contradictions, and emerging confluence. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 163–188). Thousand Oaks, CA: Sage.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Liu, J. (2007). Developing a pragmatic test for Chinese EFL learners. *Language Testing, 24*, 391–415.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, *3*, 635–694.

Long, M. H. (1997). Construct validity in SLA research: A response to Firth and Wagner. *Modern Language Journal, 81*(3), 318–323.

Long, M. H., & Norris, J. M. (2000). Task-based teaching and assessment. In M. Byram (Ed.), *Routledge encyclopedia of language teaching and learning* (pp. 597–603). London: Routledge.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331–345.

Markee, N. (1994). Toward an ethnomethodological respecification of second language acquisition studies. In E. Tarone, S. M. Gass & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 89–116). Hillsdale, NJ: Lawrence Erlbaum.

Markee, N., & Kasper, G. (2004). Classroom talks: An introduction. *The Modern Language Journal, 88*(4), 491–500.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

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

May, L. (2011a). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*, 127–145.

May, L. (2011b). *Interaction in paired speaking test: The raters' perspective*. Frankurt, Germany: Peter Lang.

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

Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23.

McNamara, T. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics, 16*(2), 159–179.

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

McNamara, T. (1997). 'Interaction' in second language performance assessment: whose performance? *Applied Linguistics, 18*(4), 446–466.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333–349.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31–51.

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62.

Mori, J. (2002). Task design, plan, and development of talk-in-interaction: An analysis of a small group activity in a Japanese language classroom. *Applied Linguistics, 23*(3), 323–347.

Mori, J. (2007). Border crossings? Exploring the intersection of second language acquisition, conversation analysis, and foreign language pedagogy. *The Modern Language Journal, 91*, 849–862.

Mori, J., & Markee, N. (2009). Language learning, cognition, and interactional practices: An introduction. *International Review of Applied Linguistics, 47*, 1–9.

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

Niezgoda, K., & Röver, C. (2001). Pragmatic and grammatical awareness: A function of learning environment? . In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 63–79). New York: Cambridge University Press.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*, 337–346.

Norris, J. M. (2008). *Validity evaluation in language assessment*. New York: Peter Lang.

Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C. J. Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Cambridge: Blackwell.

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessment*. Honolulu: University of Hawai‘i Press.

Olshtain, E., & Blum-Kulka, S. (1985). Degree of approximation: Nonnative reactions to native speech act behavior. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 303–325). Rowley, MA: Newbury House.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools*, *13*, 48–63.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge: Cambridge University Press.

Pomerantz, A., & Heritage, J. (2013). Preference. In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 210–228). UK: Wiley-Blackwell.

Ross, S. J. (2007). A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics, 39*, 2017–2044.

Roever, C. (2005). *Testing ESL pragmatics: Development and validation of a web-based assessment battery*. Frankfurt: Peter Lang Publishing.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing, 23*, 229–256.

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(2), 165–189.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, *28*, 463–481.

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

Sanders, R. E. (1987). *Cognitive foundations of calculated speech*. Albany, NY: State University of New York Press.

Schauer, G. A. (2006). Pragmatic awareness in ESL and EFL contexts: Contrast and development. *Language Learning, 56*(2), 269–318.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schiffrin, D. (1987). Toward an empirical base in pragmatics [Review]. *Language in Society, 16*, 381–396.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–33). New York: Cambridge University Press.

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement*, *3*, 323–338.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: vol. 3. Speech acts* (pp. 59–82). New York: Academic Press.

Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective*. Oxford: Blackwell.

Shepard, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20*(6), 2–16.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14.

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational*

measurement (4th ed., pp. 623–646). Westport, CN: American Council on Education.

Shimazu, Y. M. (1989). Construction and concurrent validation of a written pragmatic competence test of English as a second language. Unpublished Unpublished Ph.D. dissertation, University of San Francisco.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275–302.

Tada, M. (2005). *Assessment of ESL pragmatic production and perception using video prompts*. Unpublished Ph.D. dissertation. Temple University.

Taguchi, N. (2007a). Task difficulty in oral speech act production. *Applied Linguistics, 28*, 113–135.

Taguchi, N. (2007b). Development of speed and accuracy in pragmatic comprehension in English as a foreign language. *TESOL Quarterly*, *41*, 313–338.

Takahashi, S. (1996). Pragmatic transferability. *Studies in Second Language Acquisition, 18*, 189–223.

Takahashi, S. (2001). The role of input enhancement in developing pragmatic competence. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 171–199). New York: Cambridge University Press.

Takimoto, M. (2009). The effects of input-based tasks on the development of learners' pragmatic proficiency. *Applied Linguistics, 30*(1), 1–25.

Tashakkori, A. (2009). Are we there yet?: The state of the mixed methods community. *Journal of Mixed Methods Research, 3*(4), 287–291.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and*

behavior research. Thousand Oaks, CA: Sage.

Tateyama, Y., & Kapser, G. (2008). Talking with a classroom guest: Opportunities for learning Japanese pragmatics. In E. Alcon & A. Martinez-Flor (Eds.), *Investigating pragmatics in foreign language learning, teaching, and testing* (pp. 45–71). UK: Multilingual Matters.

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*(3), 325–339.

Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing, 17*(2), 163–184.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp.3–50). Thousand Oaks, CA: Sage.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools, 13*(1), 12–28.

Timpe, V. (2012). Strategic decoding of sociopragmatic assessment tasks-An exploratory think-aloud validation study. *Second Language Studies*, *30*, 109–246.

Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Buckingham & Philadelphia: Open University Press.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review*, *56*, 555–584.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale marker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*, 49–70.

Upshur, J., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, *49*, 3–12.

Upshur, J., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. Language Testing, *16*, 82–111.

van Lier, L. (1989). Reeling, writing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508.

Wagner, J. (1996). Foreign language acquisition through interaction: A critical review of research on conversational adjustments. *Journal of Pragmatics, 26*(2), 215–235.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24*(2), 155–183.

Walters, F. S. (2009). A conversation analysis-informed test of L2 aural pragmatic comprehension. *TESOL Quarterly*, *43*, 29–54.

Weir, C. J. (2005). *Language testing and validation*. New York: Palgrave Macmillan.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Xi, X., & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning*, *61*, 1222–1255.

Yamashita, S. O. (1996). *Six measures of JSL pragmatics* (Technical Report #14). Honolulu, HI: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Yoshitake, S. (1997). *Interlanguage competence of Japanese students of English: A multi-test framework evaluation.* Unpublished Ph.D. dissertation. Columbia Pacific University, San Rafael, California.

Youn, S. J. (2008). Rater variation in paper vs. web-based KFL pragmatics assessment using FACETS analysis. Unpublished manuscript. Honolulu, HI: University of Hawai‘i

Youn, S. J. (2010). From needs analysis to assessment: Task-based L2 pragmatics in an English for academic purposes setting. Unpublished manuscript. Honolulu, HI: University of Hawai‘i

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3–42.

Young, R. F. (2000). *Interactional competence: Challenges for validity*. Paper presented at the 2000 Language Testing Research Colloquium, Vancouver, Canada.

Young, R. (2002). Discourse approaches to oral language assessment. *Annual Review of Applied Linguistics, 22*, 243–262.

Young, R. F. (2008). *Language and interaction: An advanced resource book*. New York: Routledge Applied Linguistics.

Young, R., & He, W. (Eds.). (1998). Talking and testing: Discourse approach to the assessment of oral proficiency. Philadelphia, PA: John Benjamins.

APPENDIX A: A survey questionnaire for EAP pragmatic needs (from Youn, 2010)

Language Use in English as Academic Purpose Setting

Currently taking ELI courses: _____ Years of studying in USA: _____
Academic status (circle): Undergraduate MA PhD Other: _____
Major: _____ **First Language**: _____

Please indicate extent of learning needs of following situations.

1= Not at all necessary 2= Not necessary 3= Necessary 4= Very necessary

A. Communication with peers in class

1. During discussion, you want to know how to appropriately disagree with what classmates are saying.	1 2 3 4
2. During discussion, sometimes you don't know how to initiate, clarify, or close the conversation. You want to know how to do these during the conversation.	1 2 3 4
3. During discussion, you want to know how to comment on or compliment classmates' opinions.	1 2 3 4
4. During discussion, you want to know how to properly ask a clarification question or a relevant question to classmates' opinions.	1 2 3 4
5. When you give peer-feedback to your classmates' writing or speaking, you want to know how to give comments or suggestions nicely and appropriately.	1 2 3 4

B. Communication with professors

6. Imagine that you're working with a professor, you need to refuse some requests that you received from a professor. You want to know how to politely refuse.	1 2 3 4
7. You want to know how to appropriately make a meeting appointment with a professor either by email or in person.	1 2 3 4
8. You need a recommendation letter to apply for a scholarship. You want to know how to politely write an email to professor to request a recommendation letter.	1 2 3 4
9. When you want to thank someone such as professors or classmates, you want to know how to write a thank you email or card appropriately.	1 2 3 4

10. About writing an email to your professor or someone who you haven't met, you want to know how to appropriately write an email and reply in general. | 1 2 3 4

11. Possibly you might want to suggest something new to your professor about class. So, you want to know how to politely express your opinion and suggest new ideas. | 1 2 3 4

12. Sometimes professors make cultural jokes that are related with your class contents during the class. You want to know how to understand cultural jokes. | 1 2 3 4

13. When you talk to your professor, you couldn't understand the implied meaning. You want to know how to politely ask your professor what exactly he/she meant. | 1 2 3 4

C. Communication in general on campus

14. Sometimes you need to nicely refuse your friend's requests such as to borrow class note or your laptop. You want to know how to nicely refuse. | 1 2 3 4

15. As an international student, sometimes you might have visa problems that school should help you with. You want to learn how to nicely and effectively explain your situation to receive help. | 1 2 3 4 N/A

16. Sometimes, you feel that you're overly apologizing to someone. So, you want to know how to appropriately apologize. | 1 2 3 4

17. You want to learn how to appropriately write a cover letter or resume that you might need in future to apply for a job. | 1 2 3 4

18. Sometimes, you don't understand when people say something indirectly. You want to know how to understand indirect meaning. For example, people indirectly say "I feel so cold in the classroom" to ask you to turn off an air conditioner. | 1 2 3 4

D. Cross-cultural communication knowledge

19. In the ELI reading class, you want to read articles about different cross-cultural communication styles to raise your awareness. | 1 2 3 4

20. You think you don't know much about culturally or academically appropriate English speaking and writing rules. So, you want to discuss this matter in classes. | 1 2 3 4

APPEDNDIX B: Open Role-play Tasks and Role-play Cards

Role-play 1 (with a professor)

Situation: You have an appointment with a professor Morgan Brown today to ask for a recommendation letter for a scholarship for international students from your department and to ask a few questions about a course project. Your professor is meeting with you outside of the office hour since you have a class during the office hour. Now you're about to visit your professor. You just enter to a professor's room.

Task : You will receive role-play cards that describe what you're going to tell your professor. Please have a conversation with your professor naturally.

For undergraduate participants: This professor teaches Economy 101 that you're taking this semester.

For graduate participants: This professor is one of the faculty members in your department. Although he/she is not your advisor, you've known this professor for about 1 year and you're currently taking a course from this professor.

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play Card #1-1: Request for a letter

Jessie	Professor
1. After greeting, **ask for a recommendation letter** for the department scholarship that you will apply. The letter is due in one week.	
Jessie	**Professor**
	Respond to the request. Inform the student that you will write a letter and ask when the due date is, if the student doesn't tell you. Inform students that you have a conference next week and you're leaving tonight. Tell him/her that you will do your best to submit the letter by the deadline, but ask the student if the letter can be submitted a bit late.
Jessie	**Professor**
2. Respond to what the professor says and **tell the professor that you will check with your department office**.	
Jessie	**Professor**
3. Inform the professor of two options of providing a letter, **hard copy or electronic submissions** through a website. **Ask for the professor's preference**.	Prefer an electronic submission as you will be traveling.

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play Card #1-2: Ask for more time for meeting

Jessie	Professor
1. You have a few questions about the course project for the class that you're taking from this professor. You cannot decide a good topic. **Ask professor whether she/he has time for this question.**	
Jessie	**Professor**
	1. Inform your student that you have a pre-scheduled meeting in about 5 min. Suggest meeting another time (e.g., after class this afternoon, next week same time)
Jessie	**Professor**
3. Respond to what the professor asks	

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play Card #1-3: Respond to professor's question

Jessie	Professor
1. You have a class presentation two weeks later. Listen to what the professor says about this.	1. Ask whether the student can do a presentation a week early (which is next week). The student who was going to present next week had to cancel the presentation as he is pretty sick.
Jessie	**Professor**
2. You will be really busy next week since you have a midterm exam next Friday for history class, and **will have no time for extra class work**. Respond t o what the professor asks.	2. Respond to what the student says.
Jessie	**Professor**
3. Respond to what the professor says & close the conversation	

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play 2 (with a classmate)

Situation: After class, you're going to talk with your classmate who is doing a class project (article presentation) regarding **when and how your group members will meet** to discuss the project. The third member (Tom) is absent today in class. Your presentation is next Friday.

Task : You will receive role-play cards that describe what you're going to tell your classmate. Please have a conversation with your classmate naturally.

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play Card #2-1: Meeting time

Jessie	Phoenix
	1. As approaching to Jessie, start a conversation about an upcoming class project (article presentation). **Suggest** discussing an appropriate meeting time. Propose one available time slot based on your schedule.
Jessie	**Phoenix**
1. **Look at your schedule.** Respond to a Phoenix' question.	
Jessie	**Phoenix**
	2. Respond to Jessie's time availability **based on your own schedule**.
Jessie	**Phoenix**
2. You need to leave soon since you have another class soon. So, whether you found a good time or not, **suggest** asking the third member (Tom)'s opinion to make a final decision.	
Jessie	**Phoenix**
3. Respond what Phoenix says	3. Respond what Jessie says

APPEDNDIX B (continued): Open Role-play Tasks and Role-play Cards

Role-play Card #2-2: Discussion mode

Jessie	Phoenix
1. Move the discussion to a discussion mode. **Suggest** discussing how you will meet all together to discuss a project. Propose one option between **face-to-face discussion** and **online discussion (e.g., chatting)** that you personally prefer.	
Jessie	**Phoenix**
	1. Respond to what Jessie proposes. **Choose one option that you prefer** and express your own opinion.
Jessie	**Phoenix**
2. Respond to Phoenix' opinion.	2. Respond to Jessie's opinion. **Suggest** that you want to ask the third group member (Tom) who is absent today to make a final decision about how members will meet.
Jessie	**Phoenix**
3. Wrap up the conversation	3. Wrap up the conversation

APPEDNDIX B (continued)

Jessie's Schedule

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
9am-1pm: Classes	Part-time Work (10am-5pm)	9am-1pm: Classes	Part-time Work (10am-5pm)	9am-1pm: Classes		Part-time Work (2-9pm)

Phoenix' Schedule

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
10am-3pm: Classes	No class	10am-1pm: Classes	10am-3pm: Classes	Appointment with an advisor at 2pm	BBQ party with friends at 5pm	

APPENDIX C: Monologic Tasks

Monologic speaking task

Instruction: Please choose one question that you feel comfortable to respond. You will have 15 seconds to prepare an answer, and have 45 seconds to respond.

Personal preference: (preparation time: 15 sec, Response time: 45 sec)

1. People have different ways of escaping the stress and difficulties of modern life. What is the best way to reduce stress for yourself? Include details and examples to support your explanation.

2. Think of a place that makes you feel relaxed and peaceful. Describe it and explain why it is relaxing and peaceful for you. Include details and examples to support your explanation.

APPENDIX C (continued): Monologic Tasks

Instruction: Please choose one question that you feel comfortable to respond. You will have 15 seconds to prepare an answer, and have 45 seconds to respond.

Choice: (preparation time: 15 sec, Response time: 45 sec)

1. Which do you think is better: living on campus in dorms or off campus in apartments? Explain with reasons and examples.

2. Which kind of job would you prefer: A job that is uninteresting but has a high salary, or a job you really enjoy with a moderate salary? Explain with reasons and examples.

APPENDIX C (continued): Monologic Tasks

<u>Monologic Pragmatic Task: Give constructive peer-feedback to your classmate's e-mail</u>

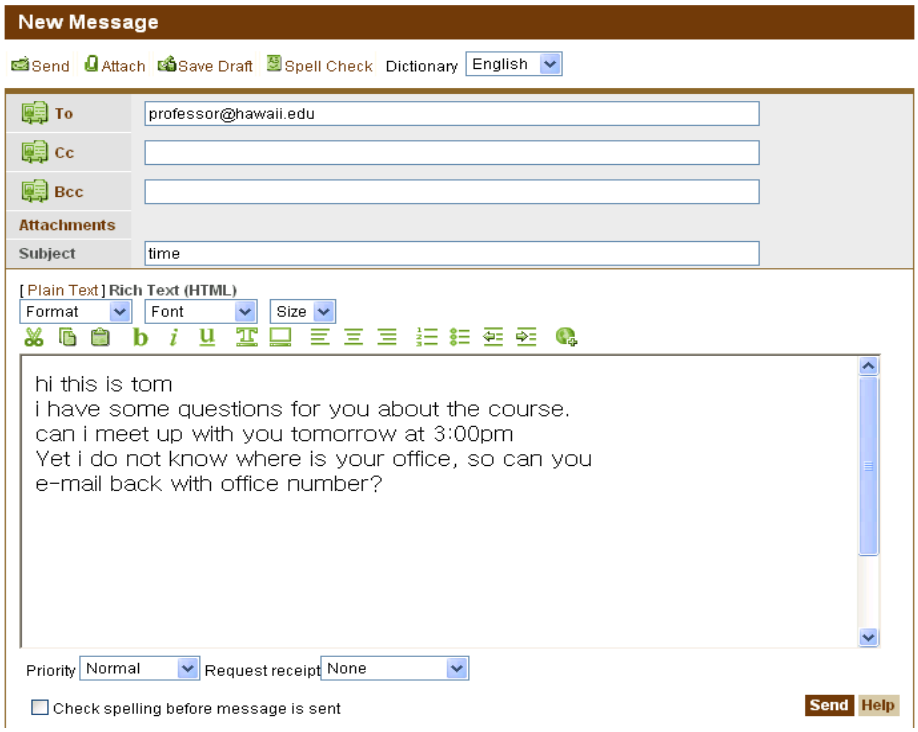
Situation: In your class, you're discussing how to write appropriate e-mails to your academic advisor in a University academic setting. The whole class was required to write an e-mail to a professor to request for a meeting. Now, your classmate sends you the e-mail that he has written. He wants your feedback on the e-mail. Now, please give constructive comments to your classmate.

Task: The e-mail below is written by your classmate (his name is Tom). Read this e-mail and give appropriate feedback verbally to your classmate so that your classmate can revise the e-mail.

Time: You have 10 minutes to complete this task.

Product: *You will prepare a short comment verbally* as if you're talking to your classmate face-to-face.

Your classmate's e-mail:



APPENDIX D: Rating Criteria for Role-play with a Professor

Score	Contents Delivery	Language Use	Sensitivity to Situation	Engaging with Interaction	Turn Organization
3	• Clear, concise, fluent (esp. speech act delivery) • Smooth topic initiations with appropriate transitional markers and clear intonations (i.e., smooth turn initiation) **#1**: a letter request & letter submission option **#2**: need for a meeting & decide a time **#3**: respond to professor's request & explain a situation	• Pragmatically appropriate linguistic expressions (**bi-clausal, conditional**: I was wondering if, I don't think I can; **modal verbs**, would, could, might) • Good control of grammar and vocabulary that doesn't obscure meaning	• Consistent evidence of awareness and sensitivity to situations exists in contents or tone **#1**: request along with explanations about the scholarship; acknowledge a short letter due **#2**: explanations for a meeting request **#3:** handle a face-threatening refusal with acceptable reasons or accept a request	• A next turn shows understandings of a previous turn throughout the interaction (i.e., shared understanding) • Evidence of engaging with conversation exists (e.g., clarification questions, backchannel, acknowledgement tokens) ***Note***: Non-verbal cues also serve as acknowledgement, so no need to heavily rely on the amount of discourse markers.	• Complete adjacency pairs (e.g., question & answer, granting a request & thank) • Interactionally fluid **without awkward pauses** or **abrupt overlap** ***Note***: Interactionally meaningful pauses include those before refusal and between disagreements
2	• Generally smooth, but **occasionally unclear**, or unnecessarily wordy • **Abrupt** topic initiation (in terms of contents) • Unclear transitional cues (e.g., unclear intonation and stress)	• Able to use modal verbs in **mono-clausal** (e.g., could, can, might), but doesn't or inconsistently use complex structures for pragmatic meaning • Linguistic expressions are occasionally inaccurate and a bit limited that sometimes obscure meaning	• Inconsistent evidence of awareness and sensitivity to situations (e.g., explain the letter request, but not acknowledge a short letter due)	• Some evidence of engaging with the conversation, but not consistent • A next turn doesn't sometimes show understandings of previous turns	• Some turns are **delayed** and a next turn is **absent** in adjacency pairs (e.g., absence of answers & thank) • Sometimes **abruptly cut off** a previous turn
1	• Delivery is choppy, fragmented, and minimal (due to lack of language competence)	• Expressions sound **abrupt, direct,** or **not polite enough** (e.g., I need, I want, I can't) • Linguistic expressions are inaccurate and quite limited that obscure meaning	• Little evidence of situational sensitivity (e.g., not acknowledge a short letter due, insist turning in the letter on time, lack of explanations for refusal)	• Noticeable absence of discourse markers • Evidence of not achieving a shared understanding	• Noticeably abrupt overlap or no pauses between disagreements and refusal • **Noticeably long pauses** or **noticeable cutoff** between turns

APPENDIX E: Rating Criteria for Role-play with Classmate

Score	Content Delivery	Language Use	Sensitivity to Situation	Engaging with Interaction	Turn Organization
3	• Clear, concise, fluent • Smooth topic initiations with transitional markers (i.e., smooth turn initiation) **Rating Phoenix**: *asking time for a meeting* & Phoenix' responses to Jessie's questions **Rating Jessie**: *asking how to meet* & Jessie's responses to Phoenix' questions ***Note***: Who initiates 'Asking Tom for a final decision' is not a crucial rating point, but focus more on delivery of follow-up contents.	• Pragmatically appropriate linguistic expressions (**bi-clausal, conditional, past progressive tense**: I was thinking, I don't think I can; **modal verbs**: would, could, might) • Good control of grammar and vocabulary that doesn't obscure meaning **Focus**: **asking questions, expressing different opinions** and **suggestions** ***Note***: No need to heavily rely on elaborated complex structures, but diverse grammatical structures for pragmatic meaning need to be observed for '3'.	• Consistent evidence of awareness and sensitivity to situations exists in contents **Examples**: what is needed for a team project (e.g., time negotiation, back up time slots for Tom), handle disagreement smoothly, explanations (at least brief) for time and meeting mode preference, pay attention to classmate's opinions ***Note***: Although not all examples need to be observed, a substantial amount of evidence needs to be observed for '3'	• A next turn shows understandings of a previous turn throughout the interaction (i.e., shared understanding) • Evidence of engaging with conversation exists (e.g., clarification questions, backchannel, acknowledgement tokens) ***Note***: Non-verbal cues also serve as acknowledgement, so no need to heavily rely on the amount of discourse markers.	• Complete adjacency pairs (e.g., question & answer) • Interactionally fluid **without awkward pauses** or **abrupt overlap** (especially between disagreement) ***Note***: Interactionally meaningful pauses include those before refusal and between disagreements ***Note***: Even with the elaborated language use ('3' in Language Use), this may not necessarily be done properly with a pause (esp. disagreement). Then, '3' in Turn Organization may not necessarily be awarded.
2	• Generally smooth, but **occasionally unclear** (which confuse a classmate), or **unnecessarily wordy** • **Abrupt** topic initiation (in terms of contents) • **Unclear transitional cues** (e.g., unclear intonation and stress)	• Able to use modal verbs in **mono-clausal** (e.g., could, can, might), but doesn't use various grammatical structures for pragmatic meaning • Linguistic expressions are occasionally inaccurate and a bit limited that sometimes obscure meaning	• Inconsistent evidence of awareness and sensitivity to situations (e.g., provide accounts for opinions, but do not necessarily handle the disagreement properly)	• Some evidence of engaging with the conversation, but not consistent (e.g., literally read the role-play card), • A next turn doesn't sometime show understanding of a previous turn	• Some turns are **delayed** and a next turn is **absent** in adjacency pairs (e.g., absence of answers) • Sometimes **abruptly cut off** previous turns
1	• Delivery is choppy, fragmented, and minimal (due to a lack of language competence)	• Expressions sound **abrupt** or **not polite enough** (e.g., I'm busy, I can't) • Linguistic expressions are inaccurate and quite limited that obscure meaning	• Little evidence of situational sensitivity (e.g., absence of providing accounts for *disagreements in particular*, handle disagreement awkwardly)	• Noticeable absence of discourse markers • Evidence of not achieving a shared understanding	• Noticeably abrupt overlap or no pauses between disagreements and refusal • **Noticeably long pauses** or **noticeable cutoff** between turns

APPENDIX F: Monologic Tasks (Preference & Choice) (Adapted from TOEFL Speaking Rating Rubric)

Score	Delivery	Language Use	Topic Development
3	Response is **clear**, **intelligible**, and **fluid** in terms of pronunciation and intonation pattern with minor lapses, which do not affect overall intelligibility.	Response demonstrates **effective and accurate use of grammar and vocabulary** with a **fairly high degree of automaticity**. Some minor errors are noticeable, but do not obscure meaning.	Response is **sustained** and **sufficient** to the task. It is generally **well developed** and **coherent**; relationships between ideas are clear (or **clear progression of ideas**).
2	Speech is **generally clear**, with some fluidity of expression, though **minor difficulties with pronunciation, intonation**, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates **fairly automated** and **effective use of grammar** and **vocabulary**. May exhibit **some inaccurate** use of vocabulary and grammar or **somewhat limited** in the range of structures used. This may affect overall fluency, but it **does not seriously interfere with** the message delivery.	Response is **mostly coherent** and **sustained** and conveys relevant ideas/information. Overall development is somewhat limited, usually **lacks elaboration** or **specificity**. Relationships between ideas may at times not be immediately clear.
1	Speech is basically intelligible, **though listener effort is needed** due to **unclear articulation, awkward intonation/rhythm/pace**, or choppy delivery; meaning may be obscured in places.	The response demonstrates **limited range and control of grammar and vocabulary**. These limitations often prevent full expression of ideas. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listening, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the **development of ideas is limited**. Mostly basic ideas are presented with **limited elaboration (details and support)**. At times **relevant substance may be vaguely expressed** or **repetitious**. Connections of ideas may be unclear.
0	**Not enough attempt or unrelated to the topic.**		

APPENDIX G: Monologic Pragmatic Task (Give constructive comments) (from Youn, 2010)

Score	Tone of giving comments/suggestions	Delivery	Language Use	Knowledge of writing an e-mail to professor
3	• Maintain **respectful, professional,** and **polite tone** throughout giving comments to classmate (e.g., not to sound too opinionated, strong, reproachful, and pushy) ***Note***: Despite elaborated language use, tone still can be disrespectful due to vocabulary choices (e.g., problem, wrong, weird) or laugh	• Comments and suggestions are **clear, intelligible,** and **fluid** in terms of pronunciation and intonation.	• Pragmatically appropriate linguistic expressions (e.g., **modal verbs, bi-clausal, conditionals**) in giving **suggestions** & **comments** (e.g., I think you can-, It would be a good idea-, If you do-, You could, I would, You might want) • Good control of grammar and vocabulary that doesn't obscure meaning.	**Show knowledge of most of the following elements of writing an e-mail in an academic context:** • E-mail format (e.g., subject, salutation, closing) • Not to assume that professor will be available at specific time • No need to ask for an office location • Keep professional and formal tone in the e-mail (e.g., modal verbs, punctuations, capital letters, accurate grammar)
2	• Maintain polite tone in general, but **inappropriateness is present** at times	• **Generally intelligible**, but **occasional difficulties** with pronunciation and intonation are noticeable.	• **Able to use modal verbs** (e.g., can, could) in **mono-clausal**, but **lack complex structures**. Do not sound polite enough (e.g., occasional uses of 'should') • **Occasionally inaccurate** linguistic expressions that obscure meaning	• Show knowledge of **some** elements mentioned above, but they are **not enough** or **lack enough explanations**.
1	• Sound strong, direct, and opinionated • Lack respectful and polite tone throughout the comment (e.g., laugh)	• Delivery is **choppy** and **minimal**. Listener effort is needed due to unclear articulation & awkward intonation.	• Use **pragmatically inappropriate linguistic expressions** that sound direct and opinionated (e.g., should, must) or **absence** of modal verbs • Limited range of grammar & vocabulary with noticeable errors	• Show knowledge of **very few** elements (e.g., **mainly grammar mistakes** & **punctuation,** do not show sensitivity to writing an e-mail in an academic setting).
0	**Not enough attempt or unrelated to the topic**			

APPENDIX H: Training Materials

Rater Training Workshop for Role-play Tasks (with a professor)

FAMILIARIZATION

I. Introduce the study & assessment tasks
   a. To develop and validate L2 pragmatic assessment tasks and rating criteria in order to measure pragmatic performance in *interaction* during various academic situations
   b. 2 role-play task & 2 monologic tasks
   c. Role-play cards for balancing between standardization & authenticity

II. Understanding *interaction* using **transcripts**
   a. Interactionally appropriate pauses & overlap
   b. Absence of next turn
   c. Incoherent turn development

III. Analytical rating criteria (please see the criteria for more examples)

Within a turn
   a. **Contents delivery**: the way that contents (actions) are delivered
   b. **Language use**: linguistic expressions related to pragmatic meaning
   c. **Sensitivity to situation**: sensitivity to various pragmatic situations (overall interaction)

Between turns
   d. **Engaging with interaction**: evidence of engaging with interaction (e.g., discourse markers, clarification questions, relevance of next turn, achieving a shared meaning),
   e. **Turn organization**: do *turn organizations* (e.g., adjacency pair, appropriate pauses) exist?, generally smooth interaction? Appropriate *pauses* are under **Turn Organization** criteria, as a timely pause is part of turn development; Does an examinee have an awareness of turn development (i.e., do examinees know when they have the floor?)

NORMING & PRACTICE

I. Prototype/Example performance & Rating points

Role-play with a professor:
Situation #1 (request for a letter with a short due date): how a request is done appropriately and how to respond to professor's time constraint
Situation #2 (request for time): how to request for time to discuss a class project and decide an alternative meeting time

Situation #3 (respond to professor's request): how to handle professor's request (either refusal or accept the request)

II. Noticeable patterns across levels
High: fluent, concise, clear, smooth interaction
Mid: somewhat fluent, relatively minimal and unclear speech act delivery
Low: noticeable absences of turns and pauses, fragmented sentences

III. Aspects that can be disregarded in the role-play rating

a. Role-play with a professor

Some participants use an interlocutor's real name
Some use their own major-related topics
Supposed to continue the conversation, but some start a new greeting in the second situation

b. Interlocutor effects (4 interlocutors as professors)

Slight variations exist (e.g., conversations with low-level examinees need more clarification questions)

IV. Rating Instructions & Strategies

a. Usually takes more time at the beginning, but will take less time after getting used to observing **important evidence** to rate
b. Start with the situation #1, then do situations #2 & #3
c. **Audacity**: a free software for sound recording and editing is useful to use
d. First listening: focus on aspects within turns (first three criteria) & Second listening: focus on aspects between turns (last two criteria)
e. Difference between accuracy and consistency in rating
f. Mark tricky cases and leave comments in rating Excel sheets
g. Evidence-based judgments: match evidence from the rating criteria descriptions to make a final decision

V. Example rating

VI. Due dates

APPENDIX H (continued): Training Materials

Rater Training Workshop for Role-play Tasks (with a classmate)

FAMILIARIZATION

I. Introduce the study & assessment tasks
   f. To develop and validate L2 pragmatic assessment tasks and rating criteria in order to measure pragmatic performance in *interaction* during various academic situations
   g. 2 role-play tasks & 2 monologic tasks

II. Understanding *interaction* using **transcripts**
   h. Interactionally appropriate pauses & overlap
   i. Absence of next turn
   j. Abrupt turn initiation

III**. Instructions** given to examinees for a role-play with a classmate
   k. Instructed to start a conversation naturally approaching to a group member right after class and both parties are aware of Tom's absence (but not necessarily instructed to mention Tom, unless specified in the role-play card)
   l. A weekly schedule for each examinee was given to find a time that works for them
   m. Instructed to talk about when to meet and how to meet separately
   n. Instructed to mention two meeting options (face-to-face, online) so that the other party is aware of the situation (as each party has different role-play cards)
   o. Instructed to choose a meeting mode that they personally prefer

IV. Analytical rating criteria (please see the criteria for more examples & details)

Within a turn
   a. **Content delivery**: the way that contents (actions) are delivered
   b. **Language use**: linguistic expressions related to pragmatic meaning
   c. **Sensitivity to situation**: sensitivity to various pragmatic situations (overall interaction)

Between turns
   d. **Engaging with interaction**: evidence of engaging with interaction (e.g., discourse markers, clarification questions, relevance of next turn, achieving a shared meaning)
   e. **Turn organization**: do *turn organizations* (e.g., adjacency pair, appropriate pauses) exist?, generally smooth interaction? Appropriate *pauses* are under **Turn Organization** criteria, as a timely pause is part of

turn development; Does an examinee have an awareness of turn development (i.e., do examinees know when they have the floor?)

NORMING & PRACTICE

V. Prototype/Example performance & Rating points
Role-play with a classmate:
Situation #1 (deciding time): clear and efficient negotiation of a meeting time
Situation #2 (deciding meeting mode): express agreement/disagreement. Sometimes, the agreement is achieved quite quickly, and then some continue to talk about where to meet (which can be considered for the rating)

VI. Aspects that can be disregarded in the role-play rating (with a classmate)
a. Instructed to talk when to meet and how to meet separately, but sometimes they talk about them together.
b. Some start a new greeting for the second situation (although they're asked to continue the conversation)
c. Some come up with situations for Tom (nothing was given about Tom except for the fact he was absent)
d. 'Asking Tom for a final decision' can be done by another party sometimes. Who really initiates this idea is not an important rating point, so focus on delivery of follow-up contents

VII. Instructions
a. Please start with an audio file named as RP1-1 (deciding time) then listen RP1-2 (meeting mode) for Pair #1. Then move onto next pair folder.
b. Using the *rating sheet*: Who starts the conversation (in terms of the contents, not an initial greeting) for each situation are marked along with a gender mark. Please double check them carefully.
c. Please listen to each file *at least four times*, as you're scoring two examinees performance for one file:
- First listening: Phoenix' performance focusing on aspects within turns (first three criteria)
- Second listening: Phoenix' performance focusing on aspects between turns (last two criteria)
- Third listening: Jessie's performance focusing on aspects within turns
- Fourth listening: Jessie's performance focusing on aspects between turns.
d. Mark tricky cases and leave comments in rating Excel sheets
e. One pair (female/female)'s conversations: difficult to discern, so transcripts are provided in the CD. If you have hard time differentiating voices in other pairs, a transcript will be available upon request.
f. Some examinees make a long turn to indicate time availability (a point deduction in 'Content Delivery')
g. Evidence-based judgments: match evidence from the rating criteria descriptions to make a final decision

VIII. Example rating

APPENDIX H (continued): Training Materials

Rater Workshop for Monologic Tasks

FAMILIARIZATION

I. Introduce a study & assessment tasks
   a. Monologic tasks (preference & choice, giving comments to a classmate)
   b. Role-play tasks

II. Internalizing rating scale
   a. Individual general speaking tasks (preference & choice)
      Delivery
      Language Use
      Topic Development
   b. Individual pragmatic task (giving comments)
      Tone of giving comments & suggestions
      Delivery
      Language Use
      Knowledge of writing an e-mail to professor

NORMING & PRACTICE

III. Prototype/Example performance
   a. Individual speaking
   b. Pragmatic task: Giving comments on an e-mail written by your classmate

IV. Aspects that can be disregarded in rating
   Preference & Choice:
   If responses are quite long, focus on scoring the response up to 50 sec
   Giving comments:
   Examinees often read sentences or a part of sentences from the e-mail in the task sheet which may sound incoherent.

V. Rating strategies
   a. Finish an individual task of *preference*, then move on to *choice* and *giving comments*
   b. Difference between accuracy and consistency in rating
   c. Listen to each file at least twice. First listening focuses on the first two criteria and second listening focuses on the last two criteria.
   d. Evidence-based judgments: match evidence from the rating criteria descriptions to make a final decision
   e. Awarding '0': Not enough attempt (e.g., too short response) or unrelated to the topic
   f. Mark tricky and borderline cases in the Excel rating sheet

VI. Practices

APPENDIX I: CA transcription

Transcription Conventions (Atkinson & Heritage, 1984)

:	Lenghtening of the preceding sound
-	Abrupt cutoff
(.)	Very short untimed pause
> <	Talk surrounded by this bracket is produced more quickly than neighboring talk
[	Point of overlap onset
=	No gap between adjacent utterances
<u>word</u>	Speaker emphasis
CAPITALS	Especially loud sounds relative to surrounding talk
° °	Utterances between degree signs are noticeably quieter than surrounding talk
(3.5)	Intervals between utterances (in seconds)

APPENDIX J: Examinee Measurement Report (Open Role-Play Tasks only)

Examinee	Measure	Error	Infit	Outfit	Examinee	Measure	Error	Infit	Outfit
1	5.08	0.52	1.00	1.06	44	3.07	0.15	1.01	1.07
2	3.11	0.26	0.97	0.99	45	0.57	0.10	0.79	0.80
3	0.96	0.20	0.90	0.91	46	2.98	0.15	0.89	0.85
4	2.01	0.21	0.91	0.93	47	-0.36	0.10	1.00	0.99
5	-1.09	0.22	0.95	1.00	48	-0.18	0.10	1.01	1.02
6	-0.62	0.21	0.92	1.04	49	1.78	0.11	1.02	1.06
7	3.11	0.26	1.26	1.45	50	2.92	0.15	1.12	1.15
8	2.15	0.21	1.18	1.18	51	0.55	0.10	1.15	1.13
9	1.11	0.20	0.66	0.68	52	-0.34	0.10	1.12	1.12
10	2.53	0.23	1.15	1.14	53	2.62	0.13	1.15	1.33
11	1.23	0.20	0.85	0.85	54	1.12	0.10	0.92	0.94
12	0.26	0.20	0.80	0.82	55	0.89	0.10	0.99	0.98
13	1.97	0.21	1.34	1.33	56	0.57	0.10	0.98	0.98
14	0.26	0.20	0.62	0.62	57	-0.75	0.10	1.05	1.07
15	0.26	0.20	0.97	0.97	58	-0.52	0.10	0.98	0.98
16	-0.31	0.21	0.92	0.95	59	1.40	0.11	0.91	0.91
17	1.51	0.20	1.31	1.32	60	1.10	0.10	0.90	0.89
18	1.51	0.20	0.76	0.76	61	2.27	0.12	1.34	1.61
19	1.23	0.20	1.11	1.12	62	0.10	0.10	1.13	1.12
20	-0.19	0.20	0.93	0.96	63	1.00	0.10	0.73	0.73
21	1.39	0.20	0.86	0.86	64	2.75	0.14	1.27	1.44
22	2.10	0.21	0.76	0.81	65	-0.04	0.10	0.92	0.92
23	2.75	0.24	0.89	0.86	66	0.25	0.10	0.77	0.77
24	1.88	0.21	0.93	0.94	67	1.51	0.22	0.96	0.93
25	0.92	0.20	0.82	0.82	68	0.75	0.20	1.08	1.08
26	-0.06	0.20	0.97	0.99	69	2.06	0.25	0.84	0.81
27	0.41	0.20	0.87	0.87	70	2.90	0.32	1.03	0.94
28	1.19	0.20	0.74	0.75	71	1.36	0.22	1.06	1.10
29	2.43	0.22	1.14	1.17	72	2.55	0.28	0.89	0.80
30	1.92	0.21	1.24	1.19	73	-0.74	0.20	0.75	0.77
31	2.29	0.22	1.17	1.16	74	-0.51	0.20	0.81	0.80
32	1.84	0.21	1.11	1.09	75	1.13	0.21	0.77	0.76
33	-1.24	0.23	1.00	1.02	76	2.81	0.31	1.17	1.11
34	-0.23	0.20	1.13	1.15	77	0.87	0.21	0.94	0.92
35	2.48	0.22	0.91	0.98	78	1.36	0.22	0.76	0.75
36	1.97	0.21	1.23	1.27	79	-0.67	0.20	1.21	1.20
37	1.87	0.11	0.96	0.91	80	-0.55	0.20	1.20	1.20
38	2.94	0.15	1.09	1.15	81	1.31	0.22	0.64	0.65
39	1.20	0.10	1.01	0.98	82	1.31	0.22	1.22	1.25
40	1.31	0.11	0.85	0.86	83	-0.47	0.20	1.02	1.02
41	-0.62	0.10	1.22	1.20	84	-1.19	0.20	0.92	0.91
42	-0.09	0.10	0.98	0.97	85	1.94	0.24	1.00	1.04
43	1.44	0.11	1.07	1.10	86	-0.01	0.20	0.90	0.91

APPENDIX J: Examinee Measurement Report (Open Role-Play Tasks only)

Examinee	Measure	Error	Infit	Outfit
87	0.38	0.20	1.14	1.15
88	1.61	0.23	1.37	1.44
89	1.61	0.23	1.00	1.02
90	1.46	0.22	1.19	1.19
91	2.32	0.27	0.98	1.05
92	1.36	0.22	1.24	1.32
93	2.47	0.28	0.95	0.93
94	1.77	0.23	1.10	1.14
95	2.18	0.26	0.93	0.92
96	1.00	0.21	0.66	0.66
97	-0.20	0.20	0.87	0.87
98	-0.67	0.20	0.99	0.99
99	1.22	0.21	1.08	1.05
100	3.12	0.34	1.00	0.90
101	3.53	0.40	1.05	1.34
102	1.13	0.21	0.93	0.91

APPENDIX K: Examinee Measurement Report (Monologic Tasks only)

Examinee	Measure	Error	Infit	Outfit
1	7.54	1.87	Maximum	
2	5.35	0.83	1.13	1.18
3	0.85	0.7	1.51	1.78
4	2.88	0.69	1.95	2.16
5	-1.53	0.75	0.89	1.63
6	-0.08	0.67	1.36	1.29
7	7.54	1.87	Maximum	
8	4.25	0.69	1.10	1.13
9	1.87	0.73	0.96	0.94
10	7.54	1.87	Maximum	
11	2.88	0.69	1.64	1.69
12	2.39	0.71	1.34	1.31
13	4.75	0.73	1.00	1.02
14	0.85	0.70	0.71	0.68
15	-1.53	0.75	2.05	2.48
16	-1.53	0.75	0.93	0.74
17	4.75	0.73	0.8	0.76
18	4.75	0.73	1.20	1.16
19	3.79	0.67	1.01	1.04
20	2.39	0.71	1.13	1.07
21	4.25	0.69	0.99	0.98
22	2.88	0.69	2.10	2.25
23	5.35	0.83	1.12	1.29
24	3.34	0.67	0.97	1.00
25	0.85	0.7	0.71	0.68
26	-0.08	0.67	0.59	0.57
27	-0.08	0.67	0.74	0.71
28	1.35	0.72	1.91	2.08
29	5.35	0.83	1.09	1.15
30	4.75	0.73	1.00	1.02
32	6.22	1.08	1.19	1.65
33	-2.15	0.83	0.49	0.33
34	0.37	0.68	0.67	0.64
35	4.75	0.73	0.88	0.93
36	4.75	0.73	0.80	0.78
37	3.80	0.35	1.06	1.19
38	7.01	0.74	0.94	0.50
39	4.29	0.36	0.98	1.23
40	3.32	0.34	0.85	0.85
41	-1.73	0.38	0.97	1.12
42	-1.44	0.37	1.69	1.46
43	3.80	0.35	0.72	0.68

Examinee	Measure	Error	Infit	Outfit
44	6.55	0.62	0.83	0.41
45	2.35	0.35	1.00	0.95
46	4.56	0.37	0.87	0.81
47	-1.31	0.37	0.82	0.71
48	-0.92	0.35	1.40	1.42
49	4.56	0.37	1.07	0.99
50	7.75	1.02	1.01	0.67
51	1.85	0.36	0.85	0.85
52	-0.80	0.35	0.93	0.92
53	4.84	0.38	0.97	0.89
54	3.08	0.35	0.69	0.71
55	1.97	0.36	0.73	0.77
56	2.10	0.36	0.76	0.72
57	-4.27	0.50	1.25	2.98
58	-1.31	0.37	0.83	0.77
59	2.48	0.35	0.81	0.83
60	1.97	0.36	1.22	1.20
61	5.15	0.41	0.88	0.77
62	1.22	0.35	1.61	1.67
63	1.09	0.35	0.77	0.72
64	6.55	0.62	1.03	0.80
65	-1.05	0.36	1.21	1.26
66	-0.56	0.34	0.88	0.89
67	5.30	0.88	0.65	0.40
68	2.40	0.70	0.93	0.88
69	4.61	0.80	0.75	0.56
70	5.30	0.88	0.96	0.70
71	4.61	0.80	0.75	0.56
72	6.26	1.12	0.74	0.34
73	-1.00	0.74	1.06	0.88
74	-5.38	1.04	1.91	2.42
75	2.40	0.70	1.38	1.33
76	6.26	1.12	1.37	1.94
77	1.44	0.68	1.31	1.20
78	2.91	0.72	0.25	0.24
79	-2.78	0.82	2.24	2.22
80	-2.78	0.82	0.68	0.47
81	3.44	0.74	0.42	0.36
82	4.00	0.76	0.84	1.02
83	-2.14	0.78	0.47	0.35
84	-1.56	0.75	0.69	0.53
85	4.61	0.80	0.41	0.32
86	0.03	0.70	0.58	0.50

APPENDIX K: Examinee Measurement Report (Monologic Tasks only)

Examinee	Measure	Error	Infit	Outfit
87	1.91	0.69	1.11	1.14
88	6.26	1.12	0.94	0.48
89	4.61	0.80	0.41	0.32
90	2.40	0.70	0.93	0.92
91	4.61	0.80	1.20	0.94
92	1.91	0.69	1.41	1.55
93	4.00	0.76	0.56	0.46
94	5.30	0.88	0.96	0.70
95	2.91	0.72	0.25	0.24
96	0.98	0.68	0.71	0.65
97	-0.48	0.72	0.36	0.31
98	-2.14	0.78	0.64	0.48
99	3.44	0.74	1.01	0.96
100	7.63	1.89	Maximum	
101	4.61	0.80	1.34	1.04
102	2.91	0.72	0.82	0.82

APPENDIX L: Examinee Measurement Report (All Tasks Combined)

Examinee	Measure	Error	Infit	Outfit
1	7.48	1.84	Maximum	
2	4.36	0.52	0.98	1.00
3	0.87	0.43	1.29	1.59
4	2.62	0.41	1.47	1.48
5	-1.86	0.43	0.83	1.07
6	-1.15	0.42	1.21	1.10
7	5.00	0.63	0.97	0.86
8	3.12	0.42	1.51	1.53
9	0.87	0.43	0.42	0.32
10	4.11	0.49	1.02	0.96
11	1.78	0.42	0.95	0.88
12	0.68	0.44	0.98	1.08
13	3.30	0.42	1.14	1.17
14	0.11	0.43	0.67	0.64
15	-0.97	0.42	1.19	1.12
16	-1.50	0.42	0.82	0.72
17	2.78	0.41	0.97	1.03
18	2.78	0.41	0.99	1.00
19	2.12	0.41	0.89	0.94
20	0.30	0.43	1.45	2.01
21	2.62	0.41	0.90	0.88
22	2.45	0.41	1.43	1.40
23	3.68	0.45	0.85	0.88
24	2.45	0.41	0.95	0.96
25	0.68	0.44	0.54	0.53
26	-0.80	0.42	0.69	0.64
27	-0.80	0.42	0.76	0.87
28	1.06	0.43	0.97	0.81
29	3.88	0.46	1.10	1.06
30	3.12	0.42	1.43	1.45
31	3.30	0.42	1.52	1.62
32	3.30	0.42	1.41	1.36
33	-2.24	0.44	0.59	0.51
34	-0.44	0.43	1.23	1.61
35	3.48	0.43	1.07	1.13
36	3.12	0.42	0.87	0.86
37	3.07	0.21	0.85	0.83
38	5.01	0.34	0.99	0.81
39	2.71	0.21	0.92	0.91
40	2.42	0.20	0.87	0.87
41	-1.48	0.22	0.99	1.04
42	-0.57	0.22	1.28	1.20
43	2.67	0.21	0.92	0.97
44	5.13	0.36	1.01	1.01
45	1.34	0.21	0.73	0.70
46	3.89	0.24	0.83	0.70
47	-0.91	0.22	0.93	1.03
48	-0.38	0.22	1.21	1.20
49	3.40	0.22	0.95	0.98
50	5.42	0.40	1.06	1.32
51	1.26	0.21	1.04	1.07
52	-0.62	0.22	1.10	1.20
53	3.95	0.25	0.92	0.81
54	2.21	0.20	0.79	0.82
55	1.64	0.20	0.83	0.84
56	1.43	0.21	0.81	0.81
57	-1.82	0.22	1.20	1.21
58	-0.71	0.22	0.85	0.92
59	2.21	0.20	0.81	0.82
60	1.76	0.20	0.99	0.99
61	3.83	0.24	1.11	1.44
62	0.51	0.21	1.45	1.55
63	1.47	0.21	0.88	0.90
64	5.13	0.36	1.03	0.89
65	-0.52	0.22	1.11	1.11
66	-0.05	0.22	0.85	0.88
67	3.45	0.46	0.92	0.82
68	1.72	0.40	0.98	0.98
69	3.93	0.52	0.84	0.77
70	4.58	0.63	0.78	0.46
71	3.05	0.43	1.27	1.30
72	5.05	0.75	0.94	0.79
73	-0.81	0.45	1.05	0.96
74	-1.99	0.44	1.26	1.12
75	2.20	0.40	0.97	0.95
76	5.05	0.75	1.06	1.05
77	1.56	0.40	0.86	0.85
78	2.20	0.40	0.58	0.60
79	-1.41	0.44	1.34	1.27
80	-1.21	0.45	1.00	0.87
81	2.70	0.41	0.66	0.68
82	2.70	0.41	1.09	1.11
83	-1.01	0.45	1.28	1.68
84	-1.99	0.44	1.08	0.90
85	3.68	0.49	0.79	0.68
86	0.18	0.44	1.11	1.26

APPENDIX L: Examinee Measurement Report (All Tasks Combined)

Examinee	Measure	Error	Infit	Outfit
87	1.07	0.41	1.08	1.12
88	3.45	0.46	1.28	1.37
89	3.25	0.45	1.01	1.12
90	2.36	0.40	1.16	1.18
91	3.68	0.49	0.94	0.98
92	2.04	0.40	1.43	1.43
93	3.68	0.49	0.78	0.66
94	3.68	0.49	1.02	1.11
95	3.45	0.46	0.63	0.58
96	1.56	0.40	0.87	0.87
97	-0.01	0.44	0.40	0.36
98	-1.01	0.45	1.13	1.30
99	2.70	0.41	0.92	0.87
100	5.80	1.03	1.04	0.99
101	4.58	0.63	0.83	0.50
102	2.36	0.40	0.88	0.91