Measuring L2 Oral Pragmatic Abilities for Use in Social Contexts:
Development and Validation of an Assessment Instrument for L2
Pragmatics Performance in University Settings

Naoki Ikeda

Submitted in total fulfilment of the requirements of
the degree of Doctor of Philosophy

September 2017

Linguistics and Applied Linguistics
School of Languages and Linguistics
Faculty of Arts
The University of Melbourne
Abstract

This study describes the development and validation of an assessment instrument of L2 oral pragmatics for university activities in English-medium contexts. Assessment of L2 pragmatics has been developed by conceptualizing the test construct as well as discussing task formats. Traditionally, the focus of pragmatics has been on offline knowledge of speech acts isolated from interaction. Recent research efforts to elicit speakers’ performances of L2 pragmatics by employing communicative tasks have expanded the argument to the perspective of interaction. Despite considerable groundwork in pragmatics research, the findings in the literature of L2 pragmatics and L2 pragmatic assessment are constrained by the narrowly defined construct and task design to restrict participants’ performances by pre-planned scenarios. Therefore, in the traditional methodology, conclusions drawn from the research of participants’ pragmatic abilities are inevitably limited, making the extrapolation of the elicited task performances to the real circumstances questionable.

The study is therefore designed to develop and evaluate a discursively-orientated instrument for L2 oral pragmatics integrating managing interaction, which allows us to examine and draw broader conclusions about their abilities to handle pragmatic demands. The language activity domain, from which task situations were created, was specified as English-medium university settings. In addition to the issues of test construct and task format, this study explored how the discursively-orientated assessment can be implemented practically while avoiding construct under-representation.

L2 speakers’ task performances on dialogue and monologue tasks, combined with actual speakers’ perspectives and their test scores were examined by a range of
analyses including qualitative discourse analyses, multi-faceted Rasch analyses, reliability estimation, a correlational analysis and group comparisons. Multiple sources of evidence were integrated to structure an argument (Chapelle, 2008; Kane, 2006; Knoch & Elder, 2013) for test score interpretation and use.

The results showed that the test takers (N=67), including university students and prospective students, were widely separated according to their pragmatic ability defined in the current study. The raters (N=3) showed internal consistency of rating although they showed differing severity in rating the test takers’ task performances. The two sets of test items (role play tasks; N=12 in total) attained high reliabilities and served also to elicit the test takers’ broader aspects of pragmatic features which separated the test takers. The test takers’ abilities estimated by the separate Rasch analyses based on the dialogue data and the monologue data respectively were highly correlated.

Overall, the test results were seen as useful for inferences about L2 students’ oral pragmatic abilities for university activities and for making decisions for educational purposes. It was also implied that the monologue tasks, which are less resource-intense than the dialogue tasks, can serve as an alternative to the dialogue tasks in separating and ranking L2 students according to their pragmatic ability.
Declaration

This is to certify that:

i. the thesis comprises only my original work towards the PhD except where indicated in the Preface,

ii. due acknowledgement has been made in the text to all other material used,

iii. the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices
Acknowledgements

I am enormously grateful for having being granted the wonderful opportunity to work on this PhD study, through which I have learned a great many things. More importantly, this study gave me the most pleasant opportunity of meeting my supervisors, my colleagues, and my family in Melbourne.

First and foremost, I would like to thank my supervisor, Dr. Carsten Roever, who provided me many insights into the study, who has been a constant force of encouragement, and who tirelessly provided me with opportunities to enhance my research skills. I would also like to thank my co-supervisor, Dr. Ute Knoch, who gave me advice most readily and willingly whenever I needed. Their critical advice, insights and passion for language assessment and for education in general have always been a source of inspiration for me. My thanks also to my thesis committee chair, Professor Lesley Stirling, who encouraged me with her most helpful feedback during the journey of my work. Professor Tim McNamara supported me by creating environments where I could interact both academically and socially with my post graduate colleagues at our university.

My heartfelt gratitude also goes to my past and present colleagues for their generous advice and support during several phases of this study, namely: Kellie Frost, Hannah Kunert, Agnese Bresin, Sharon Yahalom, Daniela Diedrich, Woranon Sitajalabhorn, Marzooq Aldossary, Dr. Ali Rastgou, Dr. Peter Hurst, Dr. Erin Moore, Dr. Hyejeong Kim, and Gwan-Hyeok Im.

I also extend my gratitude to Dr. Takanori Sato, who paved my way into studying at the University of Melbourne and who generously provided advice for my living and for my academic work, and to Professor Tomoyasu Akiyama for also tirelessly encouraging me.
The data collection was the heart of the study, which was made possible by the participants, to all of whom I owe deep and sincerely felt gratitude. My sincere gratitude is also extended to Mr. Nick Philippou, Mr. Makoto Abe, Dr. John Pill, Ross Smith, and Stephen Richmond, who provided enormous support during the data collection.

I am also grateful to Dr. Mike Linacre for his useful advice on my data analysis using FACETS.

This project was also possible thanks to the generous funds of the Melbourne Scholarships from the University of Melbourne, Small Grants for Doctoral Research in Second or Foreign Language Assessment from Educational Testing Service, Doctoral Dissertation Grant from the International Research Foundation for English Language Education, and M.A.K. Halliday Scholarship from the Applied Linguistics Association of Australia. I would like to acknowledge, with profound thanks, their financial support for this study and their encouragement.

My heartfelt gratitude also goes to my parents in Japan, and to my family in Melbourne, George, Carole, Chrissy, Elliott, Ashton, Arina and Helen. I thank my parents in Japan for believing in me, and for supporting me financially and emotionally. I thank my wife Helen and my daughter, Arina, whom I had never expected to meet before starting my study in Melbourne. I deeply appreciate their enormous support, laughter and love. I am deeply indebted to my father-in-law, George Theodoridis, academic, from whom I would like to learn much more. Finally, and most of all, I would like to thank Helen, whose love and support led me to the completion of the project.

To one and all mentioned above, your moral, intellectual and pragmatic support, I have appreciated beyond measure and for that I will be eternally grateful.
Table of Contents

Abstract ........................................................................................................................................... i
Declaration ............................................................................................................................... iii
Acknowledgements ................................................................................................................ iv
Table of Contents ....................................................................................................................... vi
List of Tables ........................................................................................................................... ix
List of Figures .......................................................................................................................... xi

Chapter 1. Introduction ................................................................................................................. 1
1.1 Background of the Study ........................................................................................................ 1
1.2 Goals of the Study .................................................................................................................. 3
1.3 Organization of the Thesis ..................................................................................................... 4

Chapter 2. Literature Review ...................................................................................................... 5
2.1 Introduction ............................................................................................................................ 5
2.2 The Construct of L2 Pragmatics ............................................................................................... 5
  2.2.1 Pragmatics from Speech Acts to More Discursive Orientation ........................................... 6
  2.2.2 Interactional Competence Compared with Theoretical Models of Communicative Ability ........................................................................................................................................... 9
  2.2.3 The Construct of Interactional Competence and Discursive Pragmatic Competence in the Literature and Empirical Studies ................................................................................................................. 11
  2.2.4 Development of L2 Pragmatic Abilities: Proficiency and Exposure to Target Settings .............................................................................................................................................. 18
  2.2.5 Role Plays as a Method to Elicit L2 Pragmatics ................................................................. 21
2.3 Testing of L2 Pragmatics ......................................................................................................... 28
  2.3.1 Pragmatics Under-represented in Proficiency Tests ......................................................... 29
  2.3.2 Studies of Testing of Pragmatics ........................................................................................ 31
2.4 Validity and Validation of Assessment Instruments ............................................................. 39
  2.4.1 Development of the Notion of Validity .............................................................................. 40
  2.4.2 Argument-based Approach to Validation ......................................................................... 42
  2.4.3 Validation Process ........................................................................................................... 45
  2.4.4 Application of Argument-based Approach to Testing of L2 Pragmatics ......................... 51
Chapter 5. Conclusion

5.1 Summary of the Key Findings

5.2 Implications

5.2.1 Educational Implications: Highlighted Significance of Practice of L2 Pragmatic Assessments for University Activities

5.2.2 Theoretical Implications

5.2.3 Methodological Implications

5.2.4 Implications for Research on Testing of L2 Pragmatics

5.3 Limitations and Suggestions for Future Research

5.4 Concluding Remarks

References

Appendix A. Task General Instruction to Test Takers

Appendix B. Role Play Task Set 1

Appendix C. Role Play Task Set 2

Appendix D. Task Prompts for Interlocutor

Appendix E. Questionnaire Items

Appendix F. Rating Rubrics

Appendix G. Transcription Conventions

List of Tables

Table 2.1 The Construct of Interactional Competence and L2 Pragmatic Competence

Table 2.2 Traits of Interactional Competence of More Proficient L2 Learners

Table 3.1 Summary of Domain Specifications

Table 3.2 A Framework of Tasks

Table 3.3 List of Possible Activities

Table 3.4 The Purposes and Outcomes of the Drafting and Piloting Phases

Table 3.5 Final Version of the Test Tasks

Table 3.6 Test Population L1 Distribution
Table 3.7 Grouping of the Test Takers…………………………………………………89
Table 3.8 Descriptions of Pragmatically Competent Test Takers.....................96
Table 3.9 Rating Order…………………………………………………………….100
Table 4.1 Backings and Sub-questions to Address the Domain Description
Inference…………………………………………………………………………107
Table 4.2 Response Distribution of Test Takers’ Awareness of Pragmatic
Demands……………………………………………………………………….110
Table 4.3 Mean Scores of Test Takers’ Awareness of Pragmatic Demands…111
Table 4.4 Backings and Sub-questions to Address the Evaluation Inference……113
Table 4.5 FACETS Descriptive Statistics: Separation………………………….201
Table 4.6 FACETS Descriptive Statistics: Internal Consistency………………202
Table 4.7 Criteria Measurement Report (Arranged by Criterion Difficulty)……202
Table 4.8 Rater Measurement Report (Arranged by Rater Severity)…………..205
Table 4.9 Backings and Sub-questions to Address the Generalization Inference…209
Table 4.10 Test Item Reliability……………………………………………………210
Table 4.11 Assignment of the Test Takers…………………………………………212
Table 4.12 Test Taker Measurements of the Two Separate Rasch Analyses……212
Table 4.13 Rater Measurements of the Separate Rasch Analyses
(Arranged by Rater Severity)……………………………………………….215
Table 4.14 Task Measurements of the Separate Rasch Analyses (Arranged by
Estimated Difficulty for Each Set)………………………………………….216
Table 4.15 Backings and Sub-questions to Address the Explanation Inference……219
Table 4.16 Test Taker Groups……………………………………………………221
Table 4.17 Test Taker Distribution According to their Estimated
Pragmatic Abilities……………………………………………………………….223
Table 4.18 Total Amount of Time Spent for Task Preparation……………………228
Table 4.19 Backings and Sub-questions to Address the Extrapolation Inference…..230
Table 4.20 Perceived Similarity between Task and Actual Performances………..232
Table 4.21 Test Takers’ Perceived Tiredness Affecting Their Performances……...237
Table 4.22 Test Takers’ perceptions of the Extent to Which Doing Six
Role Plays with the Same Interlocutor Affected their Performances………..238
Table 4.23 Mean Scores of Test Takers’ Self-assessment Items……………………241
Table 4.24 Distribution of Pragmatic Ability and Possible Cut-off Points………246
Table 4.25 Correlations between Dialogue and Monologue Abilities
List of Figures

Figure 2.1  DCT item ........................................................................................................... 22
Figure 2.2  An open role play format (Tran, 2014, p. 303) ..............................................26
Figure 2.3  An alternative role play format (Youn, 2013, p. 165) .................................27
Figure 2.4  Validity argument following Kane (2006) and Chapelle (2008)..............46
Figure 3.1  The research activities.....................................................................................64
Figure 3.2  The interlocutors and matters in the specified academic domain..............69
Figure 4.1  FACETS ruler map........................................................................................199
Figure 4.2  Probability curve: Language Use to Deliver the Intended Meanings..........207
Figure 4.3  Test taker distribution (task set 1) .................................................................213
Figure 4.4  Test taker distribution (task set 2) .................................................................214
Figure 4.5  Test taker distribution (all test takers) .........................................................222
Figure 4.6  Test taker distribution by task preparation time..........................................228
Chapter 1 Introduction

1.1 Background of the Study

Pragmatic competence is one of the components of major communicative competence models (Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale 1983; Canale & Swain, 1980; Purpura, 2004), and accordingly is a crucial part of human communication in real life (Roever, Fraser & Elder, 2014) and should therefore be “a key domain in language assessment” (Kasper & Ross, 2013, p.1). Testing of second language (L2) pragmatics is an area that has started to be explored relatively recently compared to other components of language competence (Roever, 2013; Youn, 2013) and is currently under-represented in English proficiency tests (Roever, 2011). However, it has received growing interest over the past decades.

While the literature pertaining to testing of L2 pragmatics has without a doubt, provided valuable insights for realization of language assessment that reflects more explicit human communication, it has highlighted the limitations of the previous studies and the challenges that they face (Roever, 2011; Youn, 2013, 2015). Traditionally, the focus of testing of L2 pragmatics has been on offline knowledge of speech acts (Searle, 1969, 1975) isolated from interaction since during the past decades, it was researched predominately in studies relating to pragmatics in general. In those traditional studies, L2 speakers’ pragmatic abilities were simplified and limited to speech acts examined under conditions that overly constrained their pragmatic performances by pre-planned scenarios. Therefore, the conclusions that those studies can draw about L2 speakers’ pragmatic abilities were inevitably limited. In this regard, although very limited, recent studies (Grabowski, 2009; Roever, et al., 2014; Youn, 2013, 2015) have broadened our perspectives beyond speech acts and have expanded the argument to include discursive
perspectives.

Integrating L2 speakers’ extended discourse performances and assessing a broader construct of pragmatics should be desirable as L2 speakers’ task performances elicited by communicative tasks allows for a broader conclusion of their pragmatic abilities, a stronger extrapolation of the task performances to reality, and most of all, richer information for test users (e.g., teachers and learners). A serious problem with this testing method, however, is that it negatively impacts instrument practicality. Despite its strengths, the discursive approach increases resource intensity and would consequently threaten instrument practicality, which hinders its utilization. Few implications have been provided for practicality of discursively-orientated tests of L2 pragmatics in order to reduce resource-intensiveness (e.g., costs for hiring role play conductors) while avoiding construct under-representation (Messick, 1989).

Only two studies (Roever, et al., 2014; Youn, 2013) in testing of L2 pragmatics have attempted extensive validation in an argument-based framework (Chapelle, 2008; Kane, 2006; Knoch & Elder, 2013). The argument-based approach to validation allowed these two studies to structure an argument to draw a conclusion and to logically demonstrate the extent to which, and how the conclusion was supported, rather than investigating and reporting different types of evidence (e.g., inter-rater reliability) separately and unsystematically. These two studies (Roever, et al., 2014; Youn, 2013) were methodologically not identical, and therefore the range and the types of validity evidence that was evaluated, their strengths and shortcomings, were different.

As Roever et al. (2014) pointed out, testing of L2 pragmatics has remained in the development stage for three decades and almost none of the pragmatic tests has been used in practice, although pragmatics is recognized as vital for successful communication in real life. However, the recent discursive approach to testing of L2
pragmatics, the arguments broadening our perspectives of the test construct, applications of more structured validation and highlighted importance of instrument practicality, indicates that testing of pragmatics has started to direct its interest to an actual utilization of an assessment battery. The realization of more meaningful and feasible assessment is a challenge facing the testing of L2 pragmatics (McNamara & Roever, 2006) since so far, none of the studies in this field have addressed these issues simultaneously.

1.2 Goals of the Study

With the insights and the guidance provided by the previous studies, the present study was designed to explore a meaningful and feasible assessment of pragmatics. More specifically, it aimed to (a) develop a discursively-orientated instrument to measure a broader construct of L2 pragmatic ability beyond speech acts, targeting L2 students’ oral pragmatic abilities for English-medium university activities and to (b) undertake validation within the argument-based frameworks (Chapelle, 2008; Kane, 2006; Knoch & Elder, 2013). A principle role of assessment is to inform stakeholders of the test takers’ abilities measured and to aid decisions which may lead to benefits for the stakeholders. A broader construct coverage of a pragmatic assessment can allow for a stronger conclusion of how L2 speakers are pragmatically competent in reality. The rationale for adopting an argument-based approach to validation is that it allows us to provide test users with a logical account for what the test results mean and how useful they are for decisions, based on the test results. Except a few examples, this approach has not yet been practiced for L2 pragmatic assessment.

Equally importantly, the study attempted to (c) provide implications for how assessment of discursively-orientated assessment of L2 pragmatics can be implemented
practically, while avoiding construct under-representation. Even if score interpretation and use was justified, resource-intensive tests cannot be utilized before the use of test and score is considered. Beyond a purpose of reporting results of data analyses, the research findings in this study are also intended for actual utilization and for realization of pragmatic assessment.

The present study is focused on the under-explored work mentioned above and the findings of the study are expected to help fill the research gap (Youn, 2013) identified in the testing of pragmatics and to help practitioners of classroom-based assessment of pragmatics with their assessment practices.

1.3 Organization of the Thesis

The remaining part of the thesis comprises six chapters. Chapter 2 reviews the literature relevant to theoretical construct of pragmatics and the testing of L2 pragmatics and validation, discussions which locate the context of the present study. It discusses issues challenging the testing of pragmatics, which has provided the rationale for the current study. The main research questions addressed in the study are also stated in this chapter. Chapter 3 describes the methodology in which an overview of the structure of the research is presented first. It also describes the design and development of the research instruments, the research participants, the procedures of implementing the assessment, as well as an overview of the data analyses. Chapter 4 reports and discusses the results for validation and aims to address instrument practicality. Chapter 5, the concluding chapter, summarizes the key findings and discusses the implications. It addresses some limitations of the current study and proffers suggestions for future research.
Chapter 2 Literature Review

2.1 Introduction

This chapter reviews the literature relevant to the assessment of L2 pragmatics including the theoretical construct of pragmatics, data elicitation methods, and empirical studies to situate the current research. This literature review is divided into four sections. The first section reviews how L2 pragmatics has been theoretically conceptualized and has been refined in empirical studies. This section also reviews data elicitation methods illustrated in the previous empirical studies. The second section introduces the literature pertaining to assessing L2 pragmatics. The third section reviews validity and validation. In the fourth section, I will highlight the issues challenging testing of L2 pragmatics, which provide the rationale of the current study. This chapter concludes by articulating the purposes of the study and by identifying the main research questions that need to be addressed.

2.2 The Construct of L2 Pragmatics

The definition and operationalization of the test construct and the configuration of test tasks are crucial for assessment. In this section, I will first review theoretically-informed construct of pragmatics, followed by the development of L2 speakers’ pragmatic abilities. These two aspects are particularly important for research on testing of L2 pragmatics, including the current study in terms of defining what it means to be pragmatically competent and exploring how it can be assessed according to test takers’ abilities. I will then review the task instruments to elicit L2 speakers’ oral pragmatic performances that are of interest to the current study.
2.2.1 Pragmatics from Speech Acts to More Discursive Orientation

Pragmatics is a multiple theoretical orientation, which is concerned with how a speaker uses linguistic resources to take necessary actions (Kasper, 2006) to achieve communicative goals in certain contexts (Mey, 2001). Pragmatics is thus a core component of human communication. Crystal (1997) captures the essence of pragmatics as follows:

Pragmatics is the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication. (p. 301)

L2 pragmatics has undergone conceptual change over the last quarter of a century with accumulated empirical studies on L2 pragmatics for research purposes and for testing purposes as well. Early frameworks of L2 pragmatics research were based on speech act theory (Austin, 1962; Grice, 1975; Searle, 1969, 1975) and politeness theory (Brown & Levinson, 1987) from the point of view of an individual speaker’s contribution in isolation. Speech acts (Searle, 1969) were described as a speaker’s employment of linguistic resources to realize actions intended for the context (e.g., making a request and expressing apology), which essentially constitute a core component of human communication, and have been comprehensively studied (e.g., Blum-Kulka, House, & Kasper, 1989; Cohen & Olshtain, 1993). Politeness theory (Brown & Levinson, 1987) argued that three contextual variables – relative power of the hearer over the speaker, social distance between the speakers, and degree of imposition theoretically account for speaker’s language use handling face-threats. Both speech act theory and politeness theory explain a speaker’s language production in social contexts.
Their approach can be viewed as more speaker-orientated theories, as reflected in the research design of Discourse Completion Test (DCT) for empirical studies on speech acts (discussed in section 2.2.5.1 later), in which the three contextual variables were operationalized in a task and a speaker is promoted to produce a speech act (e.g., request) for the context without physically interacting with an interlocutor simultaneously. Although speech acts and politeness construct limited aspects of pragmatics, these two aspects have played a key role in pragmatics as extensively discussed in the literature of pragmatics (Kasper & Ross, 2002; Roever, 2011).

Another perspective of pragmatics was discussed in Leech’s (1983) distinction between pragmalinguistics and sociopragmatics. The former is related to the linguistic resources or expressions employed to convey or to recognize the intended meaning for the situation, and the latter refers to linguistic choices characterized by social or cultural norms (e.g., the distance between the speaker and the hearer), which determine the speaker’s language production and the hearer’s recognition of the language production. Pragmalinguistics and sociopragmatics do not necessarily treat themselves as being independent from the central component of pragmatics above (Brown & Levinson, 1987; Searle, 1969), in view of the shared concern with language use for the given context.

Leech’s distinctively formulated aspects of pragmalinguistics and sociopragmatics are not completely discrete components. Rather, they are inter-related (Roever, 2005; Thomas, 1995) and used in combined manner to produce language performance in real world contexts. Sociopragmatic knowledge enables language users to adjust their language so as to deliver appropriate linguistic tools (concerned with pragmalinguistics) for the social context. In other words, sociopragmatic knowledge, in this case, is realized through language output or linguistic tools that pragmalinguistic ability allows
the language user to deploy. Also, even if a language user with a developed ability of pragmalinguistics has a variety of pragmatic expressions, the language user’s choice may be offensive or face-threatening without sociopragmatic knowledge (McNamara & Roever, 2006). These two aspects are consistent with the more language-use related part of Crystal’s (1997) statement in terms of language choice and the language use context that constrains the language choice.

The theories of speech acts, politeness, pragmalinguistics and sociopragmatics are in line with Crystal’s (1997) definition, in that pragmatics is the study of viewpoints of users, beyond universally applicable linguistic rules (e.g., syntactic rules, phonological rules). These theories are keeping with our understanding of a speaker’s language use for a given context in which the speaker employs linguistic tools to deliver the intended meaning and in which their linguistic choices are characterized by social norms. However, they do not necessarily account for Crystal’s more discursive end of pragmatics, “the effects their use of language has on other participants in the act of communication” (p. 301).

Other components of pragmatics include implicature (Bouton, 1999) and routine formulae (Coulmas, 1979). Implicature is characterized by indirect language use which requires a listener to infer the implied meaning underlying the literal meaning. Routine formulae refers to conventionally used fixed expressions, in a particular social situation. Both components are produced and received between speakers in social language use situations.

The theoretical construct of pragmatics discussed above describes a vital portion constituting language users’ pragmatic abilities. However, communicative situations where speakers need to handle pragmatic demands in real circumstances require them to use broader abilities, including abilities to manage extended discourse. Recent
discussions (Kasper, 2006; Roever, 2011), in this sense, have highlighted the perspective of sequential organization constructed by speakers in interactions as the relevance of interest to pragmatics (Crystal, 1997). In other words, the construct of pragmatics can be characterized by perspectives of interactional competence (Hall & Pekarek Doehler, 2011) utilized in social contexts, and pragmatic actions are accomplished in discursive sequences. Kasper (2006), for instance, illustrates that L2 speakers recognize pragmatic meaning and produce language actions in the ongoing context. This is also in accord with the views of Crystal (1997) whose definition of pragmatics encompassed discursive ability for use in social interactions where a speaker’s linguistic choices can affect the interlocutor’s ways of communication.

Pragmatic ability allows language users to construct discursive sequence from opening to closing (Roever, 2011), in order to handle pragmatic demands. An ability to produce speech acts is a vital part of pragmatic ability. However, as Kasper and Roever (2005) described pragmatic competence as the “ability to act and interact by means of language” (p. 317), handling pragmatic demands in real life situations requires speakers to perform broader tasks, including managing speech or interactions in order to realize intended actions.

2.2.2 Interactional Competence Compared with Theoretical Models of Communicative Ability

The concept of interactional competence was originally developed from social sciences outside of applied linguistics. In applied linguistics, the notion of language and communication was developed via several theories. The earliest conceptualization of communicative competence dates back to Hymes’s (1972) criticism of Chomsky (1965) for ignoring social aspects of language usage. Hymes (1972) argued that “there are rules of use without which the rules of grammar would be useless” (p. 278) and proposed
linguistic “ability for use” (p. 64) [emphasis in original] in social contexts. Hymes’s notion of ability for use encompassed language users’ ability to deliver their knowledge in practice of communication beyond simply possessing the knowledge. Canale and Swain (1980) and Canale (1983) integrated Hymes’ concept and proposed four competences within communicative competence: grammatical competence, sociolinguistic competence, strategic competence and discourse competence. Later, Bachman (1990) and Bachman and Palmer (1996, 2010), extending the previous models, also presented a communicative module with an assumption of language usage. They classified communicative competence into language competence, strategic competence and psychophysiological mechanisms and described how each component in this model interacts with each other for language output. In Chomsky’s theory, competence is what a single individual has in his or herself and not an ability to interact with others. In contrast, Canale and Swain (1980) and Canale (1983) added social aspects to the previous models by proposing sociolinguistic competence and strategic competence respectively. Yet, they do not clearly describe how these components are connected in interactions in social situations.

Given the concept of communicative competence stated in the above-mentioned models, it can be said that communicative competence is related to individual inner ability and does not always refer to ability in interaction. Indeed, Canale and Swain (1980), Canale (1983), Bachman (1990), and Bachman and Palmer (1996, 2010) shed light on social, functional, and strategic aspects of language abilities, which were not included in Chomsky’s language competence. However, they do not necessarily describe how participants interactively construct actions in social events in real life. Nor do they clearly explain sequential interactions of individuals who have their own communicative competence as described in the models above.
Therefore, the notion of competence in applied linguistics is focused on the individual and on “a relationship between an individual, the individual ability, and that individual performance” (Young, 2008, p. 93). Communicative competence resides in each individual and can be described as a characteristic of an individual in an actual situation, which does not refer to competence elicited in interactions with others and how the interaction affects their language performance. In contrast, interactional competence is displayed in interactions and is constructed in the interactions with interlocutors. According to the definition proposed by Hall and Pekarek Doehler (2011), interactional competence is the “ability to accomplish meaningful social actions, to respond to co-participants’ previous actions and to make recognizable for others what our actions are and how these relate to their own actions” (p. 1). It thus could be viewed as a reconfiguration of communicative competence. Yet, at the same time, it is different in that communicative competence is focused on individual characteristics even in interactions with others.

2.2.3 The Construct of Interactional Competence and Discursive Pragmatic Competence in the Literature and Empirical Studies

“Construct” refers to theoretical representations of invisible attributes, and as the components underlying the construct of competence are to be assessed in the case of testing, it is vital to define the underlying test construct for the present study. The construct investigated in this study is discourse-orientated L2 pragmatic ability, for use in speaking contexts, which can also be conceptualized as interactional competence, in terms of the ability to operate in social interactions. As shown in the previous sections, scholars of pragmatics have invoked multi-faceted theories to define what it means to be pragmatically competent. As reviewed earlier, pragmatics is characterized by how a
speaker manages to adjust language use by recognizing contextual variables underlying the context (Brown & Levinson, 1987). It also encompasses speakers’ employment of linguistic resources to achieve the communicative goals (Leech, 1983; Searle, 1969) in sequential organization and extended discourse (Kasper, 2006; Schegloff, 2007; Roever, 2011). Pragmatic actions are thus achieved in discourse context.

Table 2.1 shows the abilities underlying interactional competence outlined by Kasper (2006) and the components of the construct of L2 pragmatics for assessment proposed by Roever (2011). Kasper (2006), by reviewing what accumulated research adopting conversation analysis for second language acquisition, summarized interactional competencies. Roever’s (2011) list was intended for language assessment as measurable features of pragmatics integrating discursive perspectives.

| Table 2.1. The Construct of Interactional Competence and L2 Pragmatic Competence |
|-----------------------------------|------------------------------------------|
| Kasper (2006)                     |                                          |
| Recognition and production of social actions in their sequential contexts | Management of turn-taking in an organized fashion |
| Organization of actions including turn-taking and affective stance by utilizing several verbal, nonverbal resources | Conducting repair in conversation when problems in recognition and production are identified in interaction |
| Co-construction of social and discursive identities through sequence organization in interaction | Recognition and production of boundaries between activities, including management of openings and closings |
| Roever (2011)                     | Production & recognition of: Discourse structure, Speech styles, Contextualization cues, Sequence organization: pre-sequences, core sequences, post-sequences, Openings & closings, Repair, Response to first-pair parts, Effect on interlocutor |
| Comprehension of implicature       |                                          |

Although the meanings covered by each of the components in Kasper (2006) and Roever (2011) do not always explicitly correspond, the abilities underlying the construct are concerned with discursive and interactional activities and these features
are mostly identified by how speakers assemble and display linguistics resources in actual ongoing social interactions, how they manage the conversation jointly and how the sequential organizations in the conversation are developed.

Not much research has been done to date concerning the development of these interational features of L2 learners, but some recent studies have addressed this question. Pekarek Doehler and Pochon-Berger (2011) investigated the ways of performing disagreement of low and high proficiency learners of French and found that higher proficiency learners are more likely to delay disagreement by developing the conversation using several preliminary moves. This is also in accord with Al-Gahtani and Roever’s (2012) study, in which the advanced learners lay preliminary moves before transmitting to the request turn to handle pragmatic demands. By adopting Conversation Analysis (Heritage, 1984), they empirically identified pre-actions projecting the upcoming of the main actions. The findings of these studies are consistent with the flow of conversation described by Schegloff (2007) who argues, from the perspective of Conversation Analysis, that the request is not positioned at the first turn of conversation.

A closer investigation into the features of lower proficiency learners was made by Al-Gahtani and Roever (2012, 2013), who described how a speaker’s language performance affects an interlocutor’s ways of interaction. They found that learners at a lower proficiency are less likely to use introductions before making requests. This may be a tendency of lower learners but more interestingly, although not convincingly concluded, it seems to imply that the performances of the low level learners intentionally or unintentionally encourage the conversation partner to be involved in a “more collaborative role in the interaction” (Al-Gahtani & Roever, 2013, p. 413) and learners seem to give “the effect on the interlocutor” (Roever, 2011, p. 474).
Empirical findings from research on other interactional traits also offer implications for L2 interactional competence. Kim (2009) investigated the acquisition of two multi-functioned Korean discourse markers. She found the usage distributions of the two discourse markers varied depending on the participants’ proficiency levels. Advanced learners used the target discourse markers at several positions in conversational turns, whereas novice learners instead used another discourse marker, the equivalent of which in English (L1) was clearer than the two target discourse markers. This implies that the more proficient an L2 learner is, the more likely he or she is able to handle the diverse patterns of the multi-functioned discourse markers. While not explicitly concluded in this study, it could imply that more proficient learners are more likely to recognize the ongoing context and to be sensitive to contextualization cues (Gumperz, 1982).

Development of discourse markers as an indicator of development of interactional competence in Kim (2009) is in line with Ishida’s (2009) case study of one L2 Japanese learner. Ishida (2009) investigated the chronological change of the learner’s usage of a particle placed in a sentence, which is identified as “a versatile linguistic resource for engaging in conversation” (p. 382). Her longitudinal observation of the learner in the L2 community shows that the learner became more capable of using this multi-functioned particle in a wider placement of conversational turns more appropriately and began to engage in developing conversation more actively.

Another case study of an intermediate L2 learner of Japanese language was conducted by Ishida (2011), observing chronological development of the learner in the target community. Ishida’s (2011) study examined how the learner provides evaluative responses to the other’s storytelling, which is a signal showing understanding of the other’s story and is identified as a trait of interactional competence (He & Young, 1998).
She found that the participant became less likely to use commentary actions that may lead to closing the conversation and instead became more likely to initiate a new conversational sequence and to provide tokens to indicate understanding of the shared conversation and more informative comments in order to continue the conversation.

The findings of the studies above were gained from L2 learners’ interactions with L1 speakers or those presumably more proficient in the language than the learners. Development of L2 interactional competence can also be seen in interactions with other L2 speakers as shown in Galaczi (2013). She classified the learners in the study into four groups according to the descriptions of competence at B1, B2, C1, and C2 in the Common European Framework for Reference (Council of Europe, 2001) and identified several interactional traits separating the learners of English across proficiency levels. Higher proficiency learners showed confidence in developing the conversation mutually by managing both self-initiated topic and the topic initiated by the other by providing listener support and agreement involving backchannels and comprehension confirmation tokens (Galaczi, 2013). Lower level learners, on the other hand, showed a tendency to shift the topic and initiate a new topic. Also, compared to the performances of higher proficiency learners, the occurrence rate of extending the other-initiated topic was considerably lower for lower proficiency learners who more likely extend the topic initiated by their own.

As pointed out by Galaczi (2013), more limited linguistic resources may pose a challenge to lower level learners when they engage in the topic started by the conversation partner. Another thing to note is that the lowest proficiency participants in Galaczi (2013) are presumably more competent than the participants placed at the lowest level in the other empirical studies discussed here. They were equivalent to B1, described as being able to “initiate, maintain, and close simple face-to-face
conversations” (Council of Europe, 2001, p. 29). The limitation of interactional traits of the B1 learner needs to be investigated further as it was not clear whether their interactional performances were more related to limited linguistic resource or to underdeveloped interactional competence. However, what is clearly indicated in Galaczi (2013) is that the highly proficient learner’s greater mutuality and engagement in jointly constructed conversations.

Although research designs differ among the studies, some tendencies of the development of L2 interactional competence can be seen. Table 2.2 below summarizes the findings of performances of L2 learners with higher proficiency (Al-Gahtani & Roever’s, 2012, 2013; Galaczi, 2013; Kim, 2009; Pekarek Doehler & Pochon-Berger, 2011) and those with longer length of exposure to the target L2 setting (Ishida, 2009, 2011).

<table>
<thead>
<tr>
<th>Traits of Interactional Competence of More Proficient L2 Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>More actively involved in the conversation (Ishida, 2009, 2011; Galaczi, 2013)</td>
</tr>
<tr>
<td>Providing support to the interlocutor to develop the topic of the conversation (Galaczi, 2013)</td>
</tr>
<tr>
<td>Less dependent on the interlocutor’s support (Al-Gahtani &amp; Roever’s, 2012, 2013)</td>
</tr>
<tr>
<td>Less changing of the topic (Galaczi, 2013)</td>
</tr>
<tr>
<td>More able to develop the other-initiated topic (Ishida, 2011; Galaczi, 2013)</td>
</tr>
<tr>
<td>Less directly rushing to the main point (Pekarek Doehler &amp; Pochon-Berger, 2011; Al-Gahtani &amp; Roever’s, 2012, 2013)</td>
</tr>
<tr>
<td>More able to recognize a broader range of functions of a discourse marker/particle and use it more flexibly and appropriately (Kim, 2009; Ishida, 2009)</td>
</tr>
<tr>
<td>More able to recognize the ongoing context (Ishida, 2009, 2011)</td>
</tr>
</tbody>
</table>

These features were displayed in performances of higher proficiency learners, but were less likely to be displayed in those of learners at lower proficiency levels. Although the classification of proficiency level, target L2 and the numbers of participants in each empirical study are different, the results of these studies provide insightful implications for research on the development of L2 interactional competence.
and on what interactional features are measurable in the assessment of oral proficiency.

Features that constitute a broader construct of pragmatics, including discursive abilities in an educational setting were empirically investigated in Youn’s (2013, 2015) assessment criteria composed of Content delivery, Language Use, Sensitivity to Situation, Engaging with Interaction and Turn Organization. These criteria were identified in the oral extended discourse performances on role plays simulating university settings. The data were analyzed for an assessment purpose. Therefore, unlike a study which involved naturally occurring data from a small number of speakers to report every single speaker’s every single feature, Youn’s studies reported somewhat generalized features among a large number of L2 speakers and the features characterize L2 oral pragmatic abilities in the educational domain. The pragmatic and interactional features outlined in Table 2.1 and Table 2.2 do not necessarily refer to pragmatic abilities utilized in particular language activity domains (e.g., academic settings and business setting). However, Youn’s (2013, 2015) studies, particularly provided valuable insight for the current study, which had the immediate relevance of interests to L2 students’ pragmatic abilities for the educational domain.

The features listed in Youn (2013, 2015) inform that pragmatically competent students are able to:

- Clearly and smoothly deliver language contents intended actions including delivery of turns (Content Delivery)
- Control linguistic resources to structure pragmatic meanings (Language Use)
- Be aware of the given situations to take actions proceeding to the intended main actions (Sensitivity to Situation)
- Interact with an interlocutor by developing shared understanding with the interlocutor (Engaging with Interaction)
• Complete turns in a manner that a speaker recognizes pragmatic demands appropriately in the discourse context (*Turn Organization*)

Youn (2013, 2015) also analytically described features characterizing pragmatically less competent students. Although these features were identified in L2 students’ oral performances on role play tasks simulating university situations, they are not necessarily the features exclusively seen in the educational domain nor under role play conditions. Youn’s pragmatic features certainly capture the broader construct of pragmatics outlined by Kasper (2006) and Roever (2011) and are also in line with the empirical findings of the previous studies on interactional competence. Youn’s study highlights that interactional resources contribute to managing pragmatic demands beyond the production of intended speech acts. Youn empirically demonstrated operationalizing interaction as a medium where pragmatic actions are accomplished as well as the target of assessment.

2.2.4 Development of L2 Pragmatic Abilities: Proficiency and Exposure to Target Settings

Development of pragmatic abilities for second or foreign language learners is not always explained in the same way as it is for first language acquisition because in the context of second language acquisition, pragmatic abilities for the first language are already developed. The question that may then be asked is what can account for pragmatic abilities or what backgrounds of L2 speakers characterize their pragmatic performance? Although individual differences should be acknowledged in developmental patterns of L2 pragmatics (Taguchi, 2012), extensive discussions regarding literature on pragmatics published to date, have revealed factors nurturing
pragmatic abilities and pragmatic performances across features (e.g., speech acts and implicature), across medium of communication (e.g., oral communication and written communication) and across skills (e.g., receptive skills and productive skills). Research methodology such as how to group participants, how to define and quantify participants’ language competence and other background variables, target languages, and task situations was also diverse.

Even limiting the studies targeting English, many empirical studies in the pragmatic literature recognized language proficiency (e.g., Kobayashi & Rinnert, 2003; Grabowski, 2009; Roever, 2005; Roever et al., 2014; Taguchi, 2002; Timpe, 2013; Youn, 2013) and exposure to the target language community (e.g., House, 1996; Roever, 2005; Roever et al., 2014; Timpe, 2013) as factors accounting for development of pragmatic abilities. Taguchi (2002), for example, investigated how two proficiency levels differentiate English learners’ recognition of implied meaning of implicature and their strategies for it. Taguchi (2002) reported that the learners with the higher proficiency “identified more frequently the speaker’s intended purpose of using an implicature” (p. 151), a suggestion which is consistent with Roever’s (2005) study which showed proficiency levels aided English learners’ comprehension of implicature. With regard to the effect of target exposure, Roever (2005) identified substantial effect of length of exposure to the target setting on knowledge of routine formulae. A smaller number of empirical studies investigating interactional features such as those outlined in Table 2.2 above have shown that, regardless of the target language in those studies and the methodology (e.g., case study and group comparison), L2 speakers’ abilities for interaction were promoted by higher linguistic abilities and by interaction with others in the target language community, although these two factors were not always discussed separately.
As the number of empirical studies on L2 pragmatics taking on the discourse approach is limited, much evidence is sought to reveal the development of L2 speakers’ pragmatic abilities to engage in oral extended discourse. Although limited, empirical evidence has confirmed the role that proficiency plays on L2 speakers’ pragmatic performances in interactive settings. Youn (2013) reported considerably high correlation (r=0.90) between oral pragmatic performance elicited by role plays and proficiency. She measured the participants’ English (as L2) proficiency using monologue speaking tasks (independent monologue speaking tasks requiring test takers to produce an oral performance about a given topic) adopted from TOEFL iBT.

Timpe’s (2013) reported somewhat different results from Youn’s (2013) in the correlation between pragmatic competence and proficiency. Her study showed a statistically significant but weaker correlation (r=0.326 to 0.304) between pragmatic performance and proficiency with a different methodological approach. To measure the participants’ English proficiency, Timpe (2013) used Cambridge Placement Test, which is “an online adaptive test of general English proficiency that tests reading, use of English, and listening” (p. 115). She discussed the proficiency measurement employed in her study excluding speaking skills as a possible account for the weak correlation. Timpe (2013), by regression analyses, also identified learners’ experience of living in the target language environment as well as exposure to target language input (general learning backgrounds including exposure to target language input in the learners’ home country) as factors affecting pragmatic performances. The substantial role of proficiency level is consistent with the findings of Grabowski (2009), where the L2 speakers’ proficiency levels were informed by the in-house placement levels indicated by the affiliation to which the learners belonged. Her study showed that among the three groups of the learners, the group with the highest English ability pragmatically
outperformed the other groups.

As argued above, definitions of pragmatics and proficiency, methodology, and findings were not always consistent in the empirical studies in the literature. Thus, a careful interpretation is needed to draw conclusions from the evidence provided from different research contexts. The literature, however, seems to agree that the increased proficiency and experience of target exposure account for pragmatic abilities. Overall, the empirical evidence in the literature has provided a picture of L2 speakers’ pragmatic abilities, which are enhanced by higher proficiency and experience of target exposure although the development of pragmatic ability does not seem to be linear and individual differences seem to exist.

2.2.5 Role Plays as a Method to Elicit L2 Pragmatics

The present study aims to elicit and assess test takers’ productive performances of L2 pragmatics. This section reviews productive research instruments for L2 pragmatics, in particular oral role plays, which elicit L2 speakers’ performance of interest to the current study.

2.2.5.1 Role plays as a method to overcome the limitations of DCT

To date, several instruments have been employed to tap into learners’ pragmatic productive abilities in L2 pragmatics research depending on the construct to explore (e.g., speech acts, interactional abilities) and onto the aspect of language abilities (e.g., recognition, productive abilities in speaking or written contexts) as Kasper and Rose (2002) described comprehensively. These include ethnographic observation, recording and self-reporting of authentic discourse, DCTs, and extended discourse tasks such as role plays. Traditionally, a large amount of research has been conducted using DCTs to
elicit particular speech acts (Searle, 1969) such as requests and apologies (e.g., Blum-Kulka, House & Kasper, 1989) as well as to investigate directness and politeness (Brown & Levinson, 1987). DCT items traditionally give situational prompts and require a speaker (a research participant) to produce an utterance. The delivery of a task prompt and elicitation of a speaker’s response can be conducted in a written or oral form. Figure 2.1 below illustrates an example of a DCT item.

Sean borrowed ten books from a local library and they are all due tomorrow. His housemate, Erika, works at the local library. Sean is not able to go to the library tomorrow, so he has decided to ask Erika to return the books to the library.

Sean: ____________________________________________

Figure 2.1. DCT item

Although DCT items elicit language users’ language output, the findings of the accumulated research using a DCT indicated the limitation of this method, as it does not entail sequential organization of the conversation, which can be seen in interactions in the real world. Golato (2003), for instance, shows that the speakers’ ways of making compliments displayed in the DCTs are different from those in natural conversations. Conclusions drawn from the research of participants’ pragmatic abilities are inevitably limited, making the extrapolation of the elicited task performance to the real world highly questionable.

As argued in the literature (Kasper & Rose, 2002; Roever, 2011), extended discourse tasks, which can counter the limitation of DCTs, have been used as a data-gathering method for pragmatics research and are expected to elicit broader pragmatic features used by learners in social interactions in real world situations. Unlike DCTs, extended discourse tasks can elicit sequential organization of spoken interactions and
can be viewed as an important tool to elicit speech production, which can also be gained through observation of natural conversation. The ability to manage a conversation or a speech in social situations is an integral component of pragmatic competence to be measured (Roever, 2011), as pragmatics reflects language usage in social interactions with others (Crystal, 1997; Kasper, 2006). This aspect of language competence is most likely to be elicited in extended discourse provided through conversation turns and responses exchanged by a participant and an interlocutor in interactions. For example, Al-Gahtani and Roever (2012, 2013) used open role plays to elicit L2 learners’ sequential organization, and their Conversation Analysis revealed the participants’ interactional features across different proficiency levels and how they deploy linguistic resources in ongoing interactions, which would not have been drawn out in a DCT.

The different functions of role plays and DCTs are reported in some studies. Sasaki (1998), for example, compared DCTs (called “production questionnaires” in her study) and role plays by measuring speech acts of Japanese learners of English. She found that role plays are appropriate for investigating complex speech acts produced with a greater variety of strategies and a number of speech turns with longer responses, while DCTs were suitable for identifying strategies used for head acts of speech acts. She concludes that DCTs cannot always be an alternative method for role plays.

Unlike naturally occurring conversations, the stakes that interlocutors have in elicited conversations and role plays do not necessarily affect those in real life, which may be a limitation of role plays and a concern with extrapolation of the obtained role play data to the real domain. However, Huth (2010) argues that “elicited interaction may draw substantially from mechanisms structuring naturally occurring talk” (p. 549), several aspects seen in participants’ performance in role plays and those in daily conversation are not different (Huth, 2010; Okada, 2010), although language actions in
simulations can be more elaborate and exaggerated enough to be assessable (Stokoe, 2013). For example, compared to turn-taking patterns in interviewer-led situations in an oral proficiency interview (OPI), turn-taking features elicited from role played performance are more similar to those in natural conversations (Okada, 2010). Also, it is revealed that the data gained from role plays entails interactivity and is highly comparable to data from other methods (Bardovi-Harlig & Hartford, 2005).

Given the fact that role play performance is elicited fundamentally under conditions constraining participants’ performances to a varied degree (depending on the task design), role play performance should not be or will never be as authentic as the naturally occurring data. However, Stokoe (2013) suggested that research participants would produce richer data under role play conditions than that drawn from naturally occurring data. In the context of pragmatics research, role plays can elicit test takers’ discursive abilities to operate in social interactions from their engagement in communication. The interactively constructed language performance, which mirrors the nature of human activity, needs to be integrated in the task construct, even though extended discourse tasks are faced with the challenge of practicality as a method (McNamara & Roever, 2006). This will help to “broaden our view of performance in second language performance assessment to permit a renewed focus on the social dimension of interaction” (McNamara, 1997, p. 459).

2.2.5.2 Open role plays: characteristics and rationale for their use

Kasper and Rose (2002) made a distinction between closed and open role plays, although both types of role plays are simulations of communicative settings (Kasper, 2000). In open role plays, participants are provided with the goals and the roles that the participants play, “but the course and outcome of the interaction are in no way
predetermined” (p. 87). On the other hand, under closed role play conditions, participants’ performance is controlled as the actor is instructed to respond to the description of a situation. Kasper and Rose (2002) further stated the contribution of open role plays in comparison to closed role plays as follows:

Unlike closed role plays, open role plays can evolve over many turns and different discourse phases. Communicative acts evolve over multiple turns and their sequential organization is contingent on the interlocutor’s uptake. The conversational activity addresses interpersonal functions, such as politeness, and interactional functions, such as coordinating speaker and listener contributions through turn taking and back channelling. Open role plays thus allow researchers to observe those aspects of conversation that are fairly independent of particular contexts and goals but, unlike authentic discourse and elicited conversation, they also allow contexts and roles that are likely to elicit specific communicative events and acts – embedded in the role play as an activity in its own right. (p. 87)

Even within the framework of open role plays, the amount of information to provide for research participants (usually for both a speaker and an interlocutor) is not identical among studies, most probably due to different intentions of the researchers. There is a number of ways to control the amount of information in role play task prompts. For example, Tran (2014), investigating L2 speakers’ production of the speech act of suggestion and how the act was deployed in interaction, gave the role play cards to both speakers in the following ways (Figure 2.2).
As argued by Kasper and Rose (2002), the outcome of the scenario was not specified. Although the role play cards controlled the speakers in what to talk about, they allowed the speakers to interact as they wished. The role play design is similar to studies such as those by Al-Gahtani (2010) and Al-Gahtani and Roever (2012, 2013) regarding this matter.

An alternative design of role play was demonstrated by Youn (2013) as in Figure 2.3. Youn (2013) adapted the extended discourse approach using the role plays to explore L2 students’ oral pragmatics, integrating discursive features from assessment perspectives.
According to the definition of Kasper and Rose (2002), this type of role play can be categorized into open role play as the outcome of the scenario is not provided, although the two speakers were instructed to conduct actions specified in the task prompt. This design may possibly limit an extrapolation of the participants’ pragmatic ability as pragmatics is concerned with what to say, how to say it, to whom to say it and when to say it (Bardovi-Harling, 2013). It seems that the task design restricts the participants’ performances in what to do and when to do it, as well as their ways of developing their extended discourse performance.
As Tran (2014) claims, open role plays elicit extended discourse performances from participants and allow us to examine test takers’ targets features (e.g., speech acts) in a discourse context. Al-Gahtani & Roever (2012, 2013) and Tran (2014), for example, examined L2 speakers’ speech acts (request in the former studies and suggestion in the later study) and their interactional features to deploy the target speech acts. Other studies (e.g., Demeter, 2007) used role plays simply as a data collection method, but the focal feature to investigate was limited to production of speech acts, and interactional features underlying role play performances were beyond their scope.

Use of role plays as a method in the literature gives us awareness of why role play method is fundamentally necessary for the study. In this respect, Al-Gahtani and Roever (2012, 2013), Tran (2014) and Youn (2013) demonstrated their rationale for utilizing their role plays because these study investigated interactional features as their target features beyond simply extracting language users’ production of speech acts. In particular, Youn (2013) made significant use of what role plays allow for, because, as mentioned above, her study examined interactional features including how speakers organize the interactions when they handle the demands of pragmatics, something which is never made possible by a DCT.

### 2.3 Testing of L2 Pragmatics

Contribution of research on pragmatics, whether that contribution is a case study or a large-scale study, has served to describe how the target pragmatic features are characterized. One mission of research on assessment is to operationalize the findings from these empirical studies, as well as to bring the theoretically-informed insights into practice in a form of testing and assessment for test users. This section first introduces testing of pragmatics in the current practice of language assessment in general and then
reviews previous studies on L2 pragmatic assessment.

2.3.1 Pragmatics Under-represented in Proficiency Tests

As the recent discussions on testing of L2 pragmatics pointed out, no test of L2 pragmatics has been operated in real situations, as a part of the official English proficiency tests (Roever, 2011; Roever et al., 2014). Similarly, for in-house assessments, L2 pragmatic assessment has been rarely practiced in classroom assessment (Youn, 2013). The empirical research on pragmatics conducted in the past decades as well as the theoretical arguments conceptualizing pragmatics (Crystal, 1997; Mey, 2001; Leech, 1983) highlights pragmatic as a crucial component of human communication. In applied linguistics, pragmatics has been recognized as a part of the models of communicative competence (Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale, 1983; Canale & Swain, 1980; Purpura, 2004). However, it appears that the current practices of language education and language assessment under-represents what it means to use language.

The construct of pragmatics and interaction has not been fully reflected in widely used official proficiency tests (Roever, 2011) including TOEFL iBT (Educational Testing Service, 2014), the International English Language Testing System, IELTS (International English language test system, n.d.), and PTE Academic (Pearson, 2012). The speaking section of TOEFL iBT, for example, employs a monologue format where test takers are required to make a speech with no reference to specific contexts, which Kasper and Rose (2002) describe as an “oversimplified view on human interaction” (p. 56). The speaking section of the IELTS involves face-to-face interactions with an examiner but pragmatics and interaction were not explicitly defined in the scoring rubrics as the target of assessment.
Furthermore, it has been shown that language performances in oral proficiency interviews (OPI) are different from conversations in real life (He & Young, 1998). More specifically, in interview settings, turns are initiated and controlled by an interviewer (Okada, 2010), and the test takers are forced to produce a large amount of speech unnaturally (Johnson, 2001), something which is not shown in actual interaction and consequently raises the question, “Is OPI really a conversation?” (van Lier, 1989, p. 494). More empirical studies will need to be conducted to confirm the evidence, yet Okada (2010) indicates that in OPIs, test takers are unnaturally restricted from displaying interactional features involving non-verbal interpersonal communication, interactive listening and interactional management (Ducasse, 2010). Interaction in an asymmetric relationship between a test taker and an interviewer, and their different goals in the interview (Johnson, 2001) do not always provide convincing information on what test takers do in authentic interactions. Thus, even if interaction is involved in the test interview tests as a format of assessment, but test takers’ interactional features are not the target of assessment. Also, a test taker’s performance can be affected by the interviewer’s ways of managing the interview (Brown, 2005; Filipi, 2015).

Oversimplified or missing construct of pragmatics may lead to underestimating L2 learners’ language ability. Language users’ general proficiency can predict their pragmatic abilities from between to “some extent” to “a varied extent” as the previous empirical research revealed. However, even if test scores provided from general proficiency measurements can completely accurately predict the test takers’ pragmatic abilities, the test takers cannot be diagnostically informed as to how specifically they are pragmatically competent. As Kasper and Ross (2013) argued, pragmatics is “a key domain in assessment of second language” (p. 1). Testing of pragmatics may face a complex challenge (Roever, 2011), however the challenge should not negate the
significance of assessment of pragmatics (Grabowski, 2009).

2.3.2 Studies of Testing of Pragmatics

Given the current lack of assessment practices at a large scale (official proficiency tests) and at in-house scale (e.g., classroom assessment), testing of pragmatics has played a vital role in informing the literature on language testing, applied linguistics, and most of all, language educators and language learners of how L2 speakers’ pragmatic abilities are characterized and how they can be assessed both positively and negatively.

2.3.2.1 Pioneers

As the earliest studies of assessment of L2 pragmatics, Hudson, Detmer and Brown (1992, 1995) developed extensive assessment instruments for Japanese-speaking ESL learners including multiple-choice DCTs, written DCTs, oral DCTs, role plays, and two types of self-assessment, to elicit three speech acts: request, refusal, and apology. Three contextual variables to affect the participants’ performances identified by Brown and Levinson (1987)-relative power, social distance, and degree of imposition—were operationalized. Their methodological framework was adapted by some other studies targeting English and other languages as L2. Yamashita (1996), for example, used the translated and modified version of Hudson et al.’s (1992, 1995) methods for 47 English-speaking learners of Japanese language in Japan, employing mainly statistical validation. Her study shows how the participants’ proficiency and exposure to L2 target settings account for the results in each elicitation method. Yoshitake (1997) employed the original methods of Hudson et al. (1992, 1995) for 25 Japanese ESL learners in Japan and examined validity and reliability of each method and which methods are
more likely to discriminate the participants by the length of overseas experience. She also analyzed qualitatively how learners change their strategies for speech acts depending on the context in the tasks, whereas Yamashita’s (1996) analysis was more quantitatively-orientated. Ahn (2005) partially adapted Hudson et al.’s (1992, 1995) methods for 53 English-speaking KFL (Korean as foreign language) learners.

Brown (2001) and Hudson (2001) analyzed these studies and revealed that the instruments used have high reliability and validity but surprisingly, reliability of the multiple-choice DCT was low. A possible negative factor in Yamashita’s (1996) study is that translated task scenarios from Hudson et al.’s (1992, 1995) original battery would not fit into the Japanese contexts. How the translated context prompts affected the participants is not explicitly revealed, as the research did not include analysis of protocol, which may have brought a more convincing conclusion. In addition, Yoshitake’s (1997) study consisted of a small number of participants and the results may not be convincingly generalizable.

Hudson et al. (1992, 1995) and the spin-off studies used learners in different contexts, and provided guidance for further research. Multiple instruments including role plays eliciting test takers’ language production as well as self-assessment eliciting their perspectives were employed. The primary focus was on speech acts and politeness as they were predominantly targeted in the literature of L2 pragmatic research in general at that time.

Liu (2006) adapted a similar approach to Hudson et al.’s (1995) to measure Mandarin-speaking EFL learners’ production and reception of two kinds of speech acts: request and apology. He used written DCTs, multiple-choice DCTs and self-assessment. His research reported interesting findings, including high reliability for multiple choice DCT (indicated by Cronbach’s alpha of 0.88), given the surprisingly low reliability of
multiple-choice DCTs that Brown (2001) and Hudson (2001) discussed. The test construct targeted in Liu’s study was limited in the framework of speech acts as Hudson et al.’s (1992, 1995) and the spin-off studies observed. Also, targeted language production was limited to writing for the purposes of producing speech acts (request and apology). His study (as Roever, 2005 to be discussed next, did), however, indicated more emphasized effort for validation of the instrument by applying an existing validation template (Messick, 1989).

Another significant contribution was made by Tada (2005) who investigated 48 EFL learners’ oral production as well as perception of three types of speech acts: request, refusal and apology, by using oral DCT and multiple DCT. His test construct was thus within the speech act framework. Reliability (indicated by Cronbach alpha) for the perception items (multiple-choice DCT items) attained a satisfactory level of 0.75, which was lower than Liu’s (2006) findings but higher than the previous studies’ that Brown (2001) and Hudson (2001) reviewed. Although emphasis on validation may not be as explicit as Liu (2006), Tada (2005) demonstrated use of a computer-based media to deliver the test items with his consideration for presentation of the test materials. His study also presented use of a computerized procedure, which could facilitate the data collection. This test delivery mode was in line with Roever (2005, 2006), whose web-based test instrument raised awareness of instrument practicality (Bachman & Palmer, 1996).

2.3.2.2 Expanded construct of pragmatics and awareness of instrument practicality

Roever’s (2005, 2006) project on testing of L2 pragmatics was undertaken in the similar chronological frame as Ahn (2006), Liu (2006) and Tada (2005), but a different attempt was made in view of the test construct of pragmatics. Roever (2005, 2006)
expanded the components of pragmatics for assessment beyond speech acts to implicature (Bouton, 1999), which is characterized by indirectness of language usage and routine formulae (Coulmas, 1979), referred to as formulaic expressions. In his research, these three components (speech acts, implicature and routine formulae) were defined in the construct of the pragmalinguistic knowledge of ESL and EFL learners and different types of formats were used: multiple-choice DCT for both routine and implicature, and written DCT for speech acts. The descriptions of task prompts, distractors for multiple-choices, and terminology were carefully designed to suit his target participants. Data samples much larger than Hudson et al. (1992, 1995), Yamashita (1996), and Yoshitake (1997) were analyzed quantitatively. Unlike these previous studies, Roever’s project involved L2 test takers with a diverse first language backgrounds. This is of importance to the present study because it targets English-medium university context as the language activity domain where multiple L1 backgrounds of students pursue university studies. Roever also analyzed the test takers’ interview data qualitatively to examine their strategies for the test. Analysis of the learners’ background information combined with their test performances showed that exposure to L2 settings contributes more to routine formula than to implicature and speech acts, both of which are more attributed to English proficiency. With regard to reliability estimates, which were a concern in the previous studies reporting the inconsistent results, relatively high and stable results across the test sections were confirmed with 0.71, 0.80, and 0.89 for the sections of routine formulae, implicature, and speech acts, respectively. As argued above, the contribution of Roever’s study is demonstrating the extensive validation using the existing validation framework (Messick, 1989). Also, the study highlighted issues of instrument practicality while simultaneously attempting to expand the test construct of pragmatics beyond speech
acts.

2.3.2.3 Measuring the construct of L2 pragmatics in extended discourse

Since the testing of pragmatics itself is a relatively new area (Roever, 2013) in applied linguistics compared to other components of language competence and also testing of pragmatics traditionally has focused on a limited aspect of pragmatics (Crystal, 1997), interactionally-orientated tests of pragmatics were extremely limited. Two recent studies (Grabowski, 2009; Youn, 2013) have attempted to explore the construct of L2 pragmatics in sequential organization in social interactions. Both used extended discourse tasks to elicit the target construct from the participants’ speaking performances and analyzed the data both quantitatively as well as qualitatively.

Grabowski (2009) designed four role play tasks designed to measure grammatical and pragmatic knowledge in speaking settings. She explored the relationship between grammatical ability (accuracy and meaningfulness) and pragmatic ability defined in sociolinguistic, sociocultural and psychological appropriateness, and how these components affect the performances of the test takers at different proficiency levels. The total of 102 participants were divided into three groups according to their proficiency levels, and, also 10 native speakers of English were involved so as to compare the response patters of native speakers with those of the non-native speakers. It is still under discussion as to whether the development of grammatical knowledge promotes the development of pragmatic knowledge or vice versa (Kasper & Rose, 2002; Niezgoda & Roever, 2001), but Grabowski’s findings suggested that pragmatic knowledge can be tested for learners at all proficiency levels.

To examine the construct of grammatical and pragmatic knowledge targeted in the study, Grabowski (2009) used multivariate generalizability theory and many facet
Rasch measurement for quantitative parts of the analysis, also analyzing the test takers’ discourse quantitatively so as to compare sociolinguistic, sociocultural, and psychological meanings embedded in performances of native speakers and those of non-native speakers. Her study demonstrates that combining findings from discourse analysis with statistical results can serve to show “evidence of validity of the underlying test construct” (p. 129) defined in her study. In her study, the rating was conducted by two native speakers of English who were trained to judge the degree of appropriateness on the assessment features. Her rating criteria were adopted from Hudson et al. (1995) which relied on native speakers’ intuitions. The two raters in her study attained high a degree of agreement and consistency in their rating.

Youn (2013) also demonstrated that investigating underlying constructs of sequential organization in interaction by both quantitative and qualitative analysis provides more defensible insights and evidence for the validity of a pragmatics assessment instrument. She developed dialogue and monologue extended discourse tasks to elicit test takers’ pragmatics in interactional speaking contexts, and validated the assessment instruments by following Kane’s (2006) validation framework. The role play data were collected from 102 test takers. The design of the instruments and validation were conducted systematically and empirically as proposed in Kane (2006) to enhance the overall validity inference of the assessment tools. (For more comprehensive discussion of validity and validation, see 2.4 below). As pragmatic ability in interaction displayed in the language activities of university students are targeted for assessment in this study, she surveyed what language actions students are likely to perceive as necessary in communications in academic settings by questionnaire. Additionally, she examined the participants’ speaking data by Conversation Analysis. This served to specify the target traits of the pragmatics construct to be assessed in “the data-driven
interaction-sensitive rating criteria” (Youn, 2013, p. 41), and to develop the dialogue tasks that successfully elicit the target pragmatic performances, which were aimed to be extrapolated to authentic performances in the target domain. The tasks were carefully designed to allow for more appropriate interpretation of the test score. It was reported that results of the Conversation Analysis showed that the dialogue tasks could separate the test takers across different proficiency levels by both language choices and interactional traits, displayed in their performances.

Unlike Grabowski (2009), Youn (2013) employed three monologue tasks (one pragmatic task and two general speaking tasks) in addition to dialogue role play tasks. The performances displayed in monologues were used to examine how the dialogue role play performances are related to the construct of L2 pragmatics and L2 proficiency as well. For the rating of oral pragmatic performances, Youn (2013) used twelve raters including native speakers and non-native speakers of English and identified the variations of severity in rating the test-takers’ performances although they maintained consistency in rating. The rating was conducted under the partially-crossed rating design as she acknowledged that a fully-crossed rating design in which all raters scored all test takers’ performances was time-consuming. The raters were provided with the descriptors, which were developed based on Conversation Analysis of the test-takers’ task performances at different proficiency levels.

To my understanding, Youn’s (2013) was the first attempt of explicit and comprehensive application of Kane’s (2006) validation framework to assessment for pragmatic competence in interaction for academic purposes. Her research corresponded to the growing awareness of discursive practice in pragmatics (Kasper & Ross, 2013; Roever, 2011), and the data samples gained from extended discourse tasks undoubtedly expand our understanding of pragmatics and our interpretations of the test score as well.
At the same time, Youn’s (2013) research suggests further questions that were worth addressing, questions which may increase the validity for the designed assessment instruments. First, her research set academic context in general as the targeted domain, but if the domain was defined more specifically (e.g., on admission, after enrolment, in class settings, or outside class) but not too narrowly, the conclusions made based on the observed performances may have been more convincing because the target domain, and accordingly possible language activities, are more operationalized. Second, although dialogue role play tasks in her research were designed to induce speakers’ oral discourse in sequential organization, what each speaker was supposed to do and how the conversation was going to develop were provided to test takers before starting the tasks (as described in Figure 2.3 above). These structured tasks could help to restrict the test takers’ speaking performance as mentioned above, but some test takers may think they would not do as the given scenario instructed them in the authentic situation. Therefore, it would be interesting to investigate how the participants themselves think they would develop the conversation in the same situation in reality and the extent to which the participants perceive their performances in the task to be similar to what they were actually going to do.

Grabowski (2009) investigated the test construct of pragmatics in oral extended discourse. Youn (2013) conducted score interpretation within an argument-based framework (Kane, 2006) for assessment of pragmatics in an EAP context. The findings of these studies allowed us to draw broader conclusions about L2 speakers’ pragmatic ability in real settings, beyond an ability to produce speech acts isolated from interaction. However, it was beyond the scope of these studies to investigate how these performance-based assessments of oral pragmatics can be implemented more practically in real circumstances. Despite the strengths of these studies, resource intensity of their
tests unavoidably increase burden on assessment practitioners in administrating these tests. Solutions were not suggested nor addressed in these studies.

One of the most recent studies, Roever et al. (2014) attempted the expansion of the test construct while maintaining practicality. The testing tasks were developed from sociopragmatic perspectives and delivered online to a large number of test takers (191 for the pilot study and 567 for the main study). Roever et al.’s (2014) study innovated test items including (a) tasks requiring test takers to judge appropriateness of speech acts and politeness on five-point Likert scale, (b) tasks requiring test takers to make dichotomous judgment and to correct inappropriate utterances, (c) extended DCT items requiring test takers to produce speech acts in a written format in multiple turns and (d) items requiring test takers to judge two constructive dialogue passages and determine which one reflects a more successful communication and the reasons it does so. Test task (a) was similar to the traditionally used appropriateness tasks (Hudson et al, 1995). Tasks (b) and (d) elicited the test takers’ perception of communication in extended discourse contexts. Task (c), multiple DCT looked at the test takers’ productive ability.

Roever et al. (2014) limited the length of the extended discourse displayed to the test takers. In their test, the test takers worked on the productive test items in a written mode and the collected data did not involve their oral production. This may therefore limit the test construct of pragmatics, in particular language users’ oral sociopragmatic ability for use, but the test construct was as expanded as a web-based test could realize. Their study served to extend the argument to the perspective of interaction, while, at the same time, raising awareness of instrument practicality.

2.4 Validity and Validation of Assessment Instruments

This chapter has reviewed the constructs of pragmatics and tasks to elicit the
underlying features of pragmatics targeted in this study. Role play can be taken to be a valid measurement tool for pragmatic competence of L2 speakers (McNamara & Roever, 2006). However, there is still a dearth of research examining the validity for productive measurement of L2 pragmatics using role plays (Kasper & Dahl, 1991; Kasper & Rose, 2002).

Validation is not just examining different types of validity separately but to “prioritize, integrate, and evaluate collected evidence using various methods” (Xi, 2008, p. 177) and the validation argument is extended to the utilization of the test. In language testing, validity is considered to be the most important aspect when developing and using a test (Bachman, 1990) and thus must always be examined both during test development and in its use (Chapelle, 1999). The original concept of test validity addressed limited aspects of validity and accordingly, validation relied on individual pieces of evidence from content analysis and correlational analysis. In language testing, the definition of validity has changed over the past half century, and methods for validation have changed accordingly. This section reviews the notion of validity and the current argument-based approach of validation, and also reviews studies on testing of L2 pragmatics which applied the argument-based frameworks for their score interpretation and use.

2.4.1 Development of the Notion of Validity

The earliest discussion of validation in language testing dates back to Lado (1961) who argued for correlational procedures in the test validation process. In Lado’s (1961) definition, validity is a quality of and thus resides in the test:

Does the test measure what it claims to measure? If it does, it is valid. If it is a test of
The notion of validity and validation at that time was characterized by examining how much the content reflects what the test intended to measure and how much the score is correlated with data gained from other criteria in a similar domain. Test validity thus addressed limited aspects of validity and accordingly, validation relied on individual pieces of evidence from content analysis and correlational analysis to reveal the test characteristics. Language testing in this period did not capture multiple aspects involved in L2 performance including test takers’ internal cognitive aspects such as the strategies used to work on the tasks.

Current discussions (Kane, 2006, 2013; Messick, 1989), however, have begun to broaden the scope of the content validity and criterion-related validity. Different types of validity examined in the past were essential but not sufficient in themselves, capturing limited aspects of language users’ ability. By integrating more than content validity and criterion-related validity, the definition of validity has changed, and methods for validation have changed accordingly. In recent discussions provided by Messick (1989, 1996), Kane (2006) and others, validity is defined as “an argument concerning test interpretation and use: the extent to which test interpretations and uses can be justified” (Chapelle, 1999, p.25). The view of validity has thus shifted from the characteristics of the observation instruments themselves to what the score means and how the score is used by stake-holders (Messick, 1989). In other words, validity does not reside in the test itself but in how we interpret the test scores. Later, multiple types of validity were unified into construct validity (AERA, APA, & NCME, 1985; Messick, 1989), which was no longer an individual component of validity, being regarded as “the most important consideration in test evaluation” (AERA, APA, & NCME, 1985, p. 2).
Integrating Messick’s conceptual framework, Kane (2006) proposed extensive guidance for a more systematic and ongoing process of validation, which I will discuss in the following sections.

2.4.2 Argument-based Approach to Validation

With recent extensive discussion of the conceptualization of validity and logical validation procedure, validity is no longer simply justifying the degree to which the test measures what it aims to measure by examining contents of test tasks and correlations with other criteria. Rather, the current approach to validity integrates each piece of validity evidence by building an argument-based framework (Chapelle, 2008; Kane, 2006; Knoch & Elder, 2013) in order to interpret what the score means so as to support the conclusion that the test developer aims to draw. This concept is summarized in the argument of Messick (1989) 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 or other modes of assessment.” [emphasis in original] (p. 13). Integrating several types of evidence into the structure of an argument-based framework enhances the degree of plausibility of the proposed claims.

Kane valued Messick’s (1989) unified model of validity but at the same time was aware of the lack of a clear guidance for validation of score interpretations and uses (Kane, 2013). Kane (2006) outlines the interpretive argument, which serves to formulate a framework of “the chain of inferences from the observed performance to the conclusion and decision” (Kane, Crooks & Cohen, 1999, p. 6). Kane’s proposal of an interpretive argument provides a logically structured framework to justify articulated claims by examining each piece of evidence, how they are related, and how they can be
integrated to support the claim and ultimately to justify the proposed interpretation of the test score.

Kane’s argument-based approach to validation was adapted and further refined for several assessment contexts as exemplified by Chapelle (2008) and Knoch and Elder (2013). Chapelle (2008) tailored Kane’s systematic approach for second language assessment. She structured a validation argument for score interpretation and use, to discuss the extent to which and how the following claim is supported: the test score of TOEFL iBT reflects “the ability of the test taker to use and understand English as it is spoken, written, and heard in college and university setting. The score is useful for aiding in admissions and placement decisions and for guiding English-language instruction” (p.321). As the TOEFL iBT involved four sections (Chapelle, Enright & Jamieson, 2008), the validity evidence was sought comprehensively across the sections, which was integrated into an overall conclusion stated above. As suggested in Messick (1989) and in Kane (2013), validity is a matter of degree. Therefore, it is not a matter of whether or not the articulated claim must be absolutely justified. Chapelle’s logically structured validation argument for TOEFL iBT however, allows us to clearly identify what evidence enhances or undermines the inferences for the score interpretation and score use.

Similarly, although in a different assessment context, Knoch and Elder (2013), proposed an argument-based framework for a post-entry diagnostic language assessment in an Australian higher education context. In their model, Knoch and Elder (2013) integrated Kane’s (1992, 2001, 2004, 2006) argument-based framework and Bachman and Palmer’s (2010) assessment use argument to emphasize the decision made based on the test score and the consequences produced by the decision. Their test is administered to post-entry students for a diagnostic purpose and the test scores provide
important information for score users (e.g., students themselves and academic staff) in deciding to what extent students need support for their academic English skills. Thus their claim was that the test results provide useful information about students’ academic English skills in English-medium university domain and that decisions made based on the test score are beneficial to the stakeholders. Addressing the decision and consequential aspects of validation could provide an account for who or what groups of students need to take a post-entry diagnostic test, which under the current practice, is not always compulsory to students (Read, 2015) and for how a post-entry diagnostic test needs to be implemented in a way that would be beneficial to stakeholders. Unlike the official tests that international students take as a requirement for university entry (e.g., the TOELF iBT), the use of test score of post diagnostic assessment is not always widely recognized by policy makers and stakeholders (including students themselves) in educational contexts. The emphasis that Knoch and Elder’s (2013) put on educational policy in addition to evaluating test score interpretation, corresponds to the current practice of a post-entry assessment and indicates a strong rationale for improving the current practice of their assessment.

The significance of these two examples (Chapelle, 2008; Knoch & Elder, 2013) is at least twofold. Firstly, they have provided a guidance for validation tailored for second language assessment, although their contexts of assessment practices are different. Kane provided a logical structure to allow researchers to undertake validation systematically. He also provided useful examples (Kane, 2013), which should be helpful to readers to deepen their understanding of argument-based approach to validation, but the descriptions are not necessarily directed to a specific assessment practice. Chapelle (2008) and Knoch and Elder (2013) serve to provide a concrete example, presenting what type of evidence corresponds to what types of validity inference and how they
obtained and evaluated the evidence. Secondly, both Chapelle (2008) and Knoch and Elder (2013) clearly show the current scope, as well as the mission of validation, which integrates perspectives of test use. Both studies have provided stronger accountability for how a test is related to stakeholders’ benefit, beyond the scope of how the test itself measures what it intends to measure.

2.4.3 Validation Process

The preliminary process of validation starts with specifying the target domain on which the validation argument and the claim of score interpretation are made (Kane, 2006, 2013). The target domain will be, for example, academic settings and other settings of interest in the real world where authentic activities occur of which the task collects samples. Tests thus aim to collect samples of language production and reception, and to extrapolate the elicited performances to the broader performance assumed in the natural setting in the domain. Identification of the target domain is important as “the target domain reflect the educational outcomes that are of interest interpreting assessment scores” (Kane et al., 1999, p. 8). The target domain involves a broad range of activities and the domain can be further specified as “Universe of Generalization” (Kane, 2006, p. 33) which is a representative subset of the target domain and represents the trait instantiations that are of immediate interest to the assessment.

Next is the configuration of tasks identifying the representative traits in the domain. Due to the time limitation of test administration, covering all possible activities occurring in the domain in a test is impossible. Especially, unlike objectively scored tests, assessments involving oral productions are confronted with a limitation of the number of tasks administered at one time. Here, creating representative sample tasks of
the domain allows us to conclude that the completion of the tasks can indicate the completion of the real activities in the domain. Tasks intended to elicit oral extended discourse production, however, essentially involve practicality issues of administrating the task, and would put a limit to the number of the test tasks. This requires test developers to consider task specification and the number of the tasks possible to realize under the given circumstances in order to maintain representativeness of the Universe of Generalization.

Validation in argument-based framework provides a step-by-step guidance to investigate the following sequences of inferences: Domain Description, Evaluation (Scoring), Generalization, Explanation and Extrapolation, and Utilization (Chapelle, 2008; Xi, 2008). Figure 2.4 shows the steps of validation following Kane (2006) and Chapelle (2008). Each inference (e.g., Evaluation) needs to be supported by evidence to link each step of validation.

<table>
<thead>
<tr>
<th>Validation step</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target domain</td>
<td></td>
</tr>
<tr>
<td>Universe of Generalization</td>
<td>Domain Description</td>
</tr>
<tr>
<td>Observation</td>
<td>Evaluation/ Scoring</td>
</tr>
<tr>
<td>Observed Score</td>
<td>Generalization</td>
</tr>
<tr>
<td>Universe Score/ Expected Score</td>
<td>Explanation</td>
</tr>
<tr>
<td>Construct</td>
<td>Extrapolation</td>
</tr>
<tr>
<td>Target Score</td>
<td>Utilization</td>
</tr>
<tr>
<td>Test use/ Trait interpretation</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2.4. Validity argument following Chapelle (2008) and Kane (2006).*

The Domain Description inference allows us to infer the extent to which the designed assessment reflects the traits and the contents (e.g., skills to operate and
contexts where the skills were utilized) which are of interest to the assessment. In the Evaluation phase, the observed performance is converted to an Observed Score, based on the rating criteria. This inference is concerned with a systematic application of rating criteria, including issues of how rating criteria were developed, how test scores are assigned and whether or not raters appropriately apply rating criteria. In a case of providing diagnostic information of language use to a single test taker, it may be allowed to provide him or her with the task features descriptively instead of quantifying the task performance into score. However, this should not be the case for assessment targeting multiple test takers beyond the aim of giving feedback to a small number of test takers. In the process of converting a specific task performance into a score or Observed Score, we lose information of the task performance because a score is a more abstract representation of the observed task performance. The information lost in scoring process could well be ascribed to the above-mentioned issues. Therefore, the Evaluation inference requires investigation as to how the score was assigned (Brennan, 2013), by involving the characteristics of the employed resources (e.g., scoring procedures, scoring rubrics) and whether or not they are appropriately administered. In the cases of assessing speaking performances, scoring criteria are expected to discriminate test takers’ traits at different competency levels required in the domain. Valid rubrics for rater training need also be developed in order to maintain consistency of scoring.

The Observed Score is converted into Universe Score in the Generalization inference. Estimation of reliability and generalizability gained from test takers’ performances and raters’ judgments (McNamara, 1996) as well as test tasks are the major concerns in this stage, in particular, in cases of performance-based assessment (Kane et al., 1999). More specifically, the consistency of test takers’ performances on
similar tasks under similar conditions is examined. Consistency in test takers’ reactions in the sampled area in the target domain can be confirmed, yet output produced by statistical procedures does not guarantee that the test takers can perform in the broader area of the target domain. The inference from observed score to Universe Score is conducted on the assumption that performance samples are representative enough to draw generalized conclusions. Performances elicited by objectively scored tests such as multiple-choice tests, allow a number of items contained in a test, and samples from a sufficiently larger population enhance generalizability across tasks and raters on an assumption that the test taker samples should be representative of the target population. As discussed earlier, performance-based assessment eliciting test takers’ oral extended discourse performance would limit the number of tasks which would require examination of whether each test item reliably measures the same construct that represents Universe of Generalization.

Observations are subject to variations in, for example, test takers’ population, raters’ population, assignment of raters to observed test performances, sample of tasks, assignment of test tasks or test task sets to test takers, test conditions, or other contexts. Therefore, the Generalization inference is supported by evidence that the test score is obtained across assessment contexts, not by a particular context and consequently, the Generalization inference is enhanced by investigating the elements exemplified above, which may involve variations according to each observation.

In the framework of Chapelle (2008), the next step is Explanation, which links Expected Score (in Kane’s term, Universe Score) to the test construct. This step examinees the extent to which the test score is indicative of the construct of the target competence. In other words, an assumption to be made for the Explanation inference is that a designed assessment is able to detect the extent of test takers’ competence, which
should be reflected in test scores. For instance, higher scores can be achieved because
the test takers have higher abilities relevant to the test construct. Thus, the phase of
Explanation also requires us to consider how different levels of the target competence
(e.g., oral pragmatic competence) is characterized theoretically and empirically, in
addition to examining the extent to which the test score can be explained by the target
construct. This phase of validation was not explicitly separated from Extrapolation (to
be discussed next) in the framework of Kane (2006), but it is beneficial to include this
phase independently for validation for an assessment of language abilities, given the
complex nature of language abilities (Chapelle, 2008; Kane, 2013; Roever, et al., 2014).

In the next phase called Extrapolation, the Universe Score is extrapolated into
Target Score, representing traits in the target domain, which is to say, the test takers’
performance is extrapolated from that in the test situation to assumed performance in
the natural setting of interest. This phase examines not only language outputs but also
the process of how the output was produced, what abilities were employed, and to what
extent test takers perceived their performances as similar to their authentic activities.
This stage also involves correlation with performance elicited in other criteria in the
target domain, analysis of test takers’ cognitive process based on verbal protocol
(Roever, 2005) and simulated recalls. More convincing inference for Extrapolation
involves investigation of self-assessment of performance and assessment of domain
experts. This provides “a warrant that the processes that are engaged by the assessment
task are essentially the same as those defined in the construct” (Bachman, 2005, p. 26)
and enables us to infer that test takers can do similar activities in authentic situations of
the domain as successfully as they did on the test tasks. This phase also involves
investigating individuals’ internal process that characterizes language productions and
language performances affected by “context” or “other traits” (Kane, 2006, p. 33).
Test utilization (Chapelle et al., 2008; Xi, 2008) can also be a part of the overall inference for validity of the test. The information on the test score interpretation made by the previous stages of validation (Evaluation, Generalization, Explanation and Extrapolation) is provided and used by stakeholders of the test, who are involved in decision-making based on the test score and test consequences such as washback of the test (Cheng & Watanabe, 2004). The Utilization inference is thus supported based on investigation of decision-making process (e.g., cut score for selection and placement of test takers) and washback on learning and teaching. The Utilization inference also implies that engaging English learners in similar language learning activities as facilitated by the designed test tasks is deemed beneficial for English learners in developing learners’ language abilities for the target domain.

Performance-based assessment has its strengths in addressing the Extrapolation inference and the Utilization inference if the test performance was of direct interest to the test users. The stronger relevance of the skills used for the observed task performance to the expected performance in reality would facilitate extrapolation of the observed task performance to the reality. The relevance of the skills and the activities being assessed to the expected performance would also enhance the Utilization inference, which is concerned with consequential aspect of validity (Messick, 1989, 1994). The Utilization inference rests on an assumption that engaging in activities simulated in the test is beneficial for the test takers as well as that, test scores are useful for making decisions for educational purposes. However, the trade-off between Generalization and Extrapolation poses a challenge to performance-based assessment. The Extrapolation inference can be strengthened by using performance-based tasks, which are similar to those in the target domain and are of interest to the test developers and to the public involved in the assessment. However, labour intensity in the process of
implementing a performance-based assessment unavoidably limits the number of possible tasks that can be administered in a testing context, thus weakening the Generalization inference. The Generalization inference can be strengthened by using larger number of tasks such as multiple-choice tasks, test items, requiring test takers to produce short answers or highly standardized items, but these test items measure limited aspects of test takers’ abilities, weakening the extrapolation of test takers’ performance to reality. As such, the Extrapolation inference can be strengthened at the expense of the Generalization inference, and vice versa (Kane, et al., 1999) and this trade-off between them is unavoidable (Roever, 2011), challenging assessment of L2 pragmatics.

2.4.4 Application of Argument-based Approach to Testing of L2 Pragmatics

Currently, in the literature of testing of L2 pragmatics, studies exemplifying argument-based approach to validation are limited to Roever et al. (2014) and Youn (2013, 2015). These studies were reviewed in section 2.3.2.3 above, as the studies contributing to the discussions on a more discursive orientation of L2 pragmatic assessment. The integrated evidence for the studies was different according to the methodological differences.

Roever et al. (2014) followed Kane’s (2006)’s approach together with Chapelle’s (2008) model tailored for the context of second language assessment. As reviewed in section 2.3.2.3 above, they developed a sociopragmatically focused testing instrument. They structured an argument for the chain of inferences of Domain Description, Evaluation (Scoring), Generalization, Explanation, Extrapolation, and Utilization. Their target language activity domain was not restricted to any particular domain (e.g., medical setting), but they defined Universe of Generalization (Kane, 2006) as “everyday language use” (Roever, et al., 2014, p. 114). To support the Domain
Description inference, they provided an account for how they themselves defined everyday language use in the test task development stage. For the Evaluation inference, which is concerned with how test scores are assigned, they integrated their practice of selection of the test items (considering item discrimination) and their scoring practice as the relevant backings. To support the Generalization inference, reliability of the test items, inter-item correlations, and inter-rater reliability were evaluated. Their investigation of the Explanation inference also involved multiple backings: (a) the group comparisons to look at effect of the test takers’ proficiency and amount of exposure to the target language use, (b) the correlational analyses to examine the strengths of correlations between sections, (c) factor analyses to identify possible factors accounting for the test scores. In addition, they (d) examined correlation between the test takers’ test results on their test battery with those obtained in Roever’s (2005) pragmalinguistic test administered to the test takers separately from the Roever et al.’s (2014) main test.

Roever et al. (2014) explicitly separated the evidence for the Explanation inference from that for the Extrapolation inference in a sense of whether the obtained evidence offers an account for the construct underlying the test (Explanation) or links the test performances to the real life situations (Extrapolation). They also elicited the test takers’ perceptions on the extent to which they found communication in English easy or difficult as a backing of the Extrapolation inference under an assumption that test takers with higher scores on their test are more likely to perceive commanding language use in reality easier than those with lower scores. Their validation evidence was obtained largely quantitatively. They also elicited test takers’ perspectives on the dialogue choice items (requiring test takers to judge appropriateness of the extended discourse displayed on the screen), and examined the test takers’ perspectives
Roever et al. (2014) concluded that overall test scores generated from their test are adequately indicative of test takers’ sociopragmatic knowledge and that test scores are useful in making decisions for educational treatment for language users. The Utilization inference could not be directly evaluated as the test was not in the operational stage. Thus, they made suggestions for test use based on the findings in the process of validation discussing the confirmed strengths and shortcomings of their test items.

The other extensive application of argument-based framework was demonstrated by Youn (2013). As discussed above, Youn’s methodology was different from Roever et al. (2014). Youn’s study involved the test takers’ oral performance on open role plays. The target domain was defined as academics and integrated discursive features as well as language use as the target construct of pragmatics. Youn outlined the validation inferences somewhat differently from Roever et al. (2014) in that she integrated the results of a need analysis conducted proceeding to the study which empirically addressed the Domain Description inference. The need analysis also served to configure the role play tasks which reflect what mattered to L2 students in handling pragmatic demands and which required test takers to use pragmatic skills relevant to the target domain. The methodological consideration for the task design, development of the rating criteria as well as rater training were integrated and treated as the evidence for the Evaluation inference. The evidence for the Generalization inference was limited to the characteristics of the raters and the tasks indicated by the fit statistics (infit values) and the logit values estimated by a multi-faceted Rasch analysis. Also, the partially-crossed rating design may have weakened the Generalization inference as she acknowledged. Later in Youn (2015), the evidence from the raters’ fit statistics and severity
measurement for the Generalization inference was reframed and integrated into the argument for the Evaluation inference, which Youn (2015) concentrated on. In the Explanation and Extrapolation phase, Youn (2013) used evidence of strengths of correlation between (a) the pragmatic abilities estimated on the dialogue role play tasks and the test takers’ English proficiency levels measured in her study and (b) the pragmatic abilities estimated on the dialogue role play tasks and those on a monologue pragmatic task. Youn (2013) did not separate the Explanation inference from the Extrapolation inference and treat the resulting correlational analyses as the backing for Extrapolation. The backing that allows for the Utilization inference were mentioned but were beyond the scope of the study.

Youn’s (2013) validation was primarily quantitative for the phases addressing the Generalization inference and the Extrapolation inference. However, she also reported the findings of the test takers’ oral role play performances analyzed qualitatively by Conversation Analysis (Atkinson & Heritage, 1984). The qualitatively analyzed discourse findings were used to develop her data-driven criteria to assess the test takers’ performances and were integrated into the validity evidence of the Evaluation inference.

Youn’s study involved oral interactive data on role play tasks, which inevitably increases resource intensity on both task administration and on rating as acknowledged by Roever et al. (2014). This would inevitably constrain the number of the role play tasks to be administered to the test takers. Using a sufficient number of tasks that represent the target-domain activities and skills could enhance the Generalization inferences but it may be difficult to attain this for a performance-based speaking test like that by Youn (2013). Thus, it would not be possible to include as many test items as the previous studies (e.g., Liu, 2006) which limited the target construct of pragmatics
and did not require test takers to produce extended discourse performances. This limitation was not explicitly mentioned by Youn (2013). With the backing made based on the items for the Generalization, she relied on stable fit statistics of all tasks and different difficulties of the tasks (both estimated by a many facet Rasch analysis). She then integrated the backing into the Generalization inference by concluding as follows:

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. (p. 118)

For tests of L2 pragmatics using an oral performance-based format, a particular concern should be a trade-off between the Generalization inference and the Extrapolation inference (Roever, 2011). The broader construct coverage enhances the latter while the former may be weakened possibly because of an impact of human raters’ rating behaviours on test takers’ scores in performance assessment (McNamara, 1996) as well as of the above-articulated limitation (the number of the tasks to be delivered to test takers). Youn (2013) in this sense, demonstrated an example of addressing the Generalization inference and reported how the raters and tasks functions in an assessment of oral pragmatics targeting interactional features together with language use.

Despite the different scopes and the methodology, Roever et al. (2014) and Youn
(2013) demonstrated that argument-based frameworks are applicable to testing of L2 pragmatics. Their approaches allowed them to work on test validation more systematically than the previous studies, which in turn made it clear to the literature what the test score means and the reason why the researchers can claim it to be so. With respect to the test use, as discussed above, the Utilization inference could not be addressed in both studies, but they provided possible backings to address at their test operation in the future.

2.5 Issues of Testing L2 Pragmatics

The previous sections have reviewed the development of the testing of L2 pragmatics from earlier studies to the recent discussions about covering a broader construct of pragmatics. The earliest studies in this area defined pragmatics fairly narrowly, limiting the construct to appropriateness of speech acts and of politeness, which was in keeping with the traditional focus of pragmatics research in general. Methodologically, from the earlier stages of the study on testing pragmatics, multiple tasks were employed including oral role plays, multiple-choice tasks, oral and written DCTs to elicit speakers’ pragmatic productive abilities, knowledge and recognition. In spite of the attempts to integrate multiple types of tasks as a format of assessment, the target of assessment was limited to the narrowly defined pragmatics. This tradition first highlights an issue of how test developers define the test construct of pragmatics. As reviewed in the previous section, pragmatics has been conceptualized as the construct, the multi-faceted by sub-areas that addressed what human communication is about. Researchers defined pragmatics and pragmatic ability differently as the targets for their respective studies were different. This should be true for testing of L2 pragmatics, which needs to define the target construct of pragmatics assess. It would doubtless be
unrealistic for a single test to cover every aspect of multiple-theoretically constructed nature of pragmatics. However, it is equally clear that focusing on the traditionally oversimplified construct of pragmatics would considerably limit a conclusion of how speakers are pragmatically competent in real communicative settings.

The recent implementation of more interactionally-orientated tests has certainly expanded the discussion to a broader construct of pragmatics beyond speech acts. Testing of L2 pragmatics has thus started to swing the pendulum from assessing limited aspects of pragmatics practically, to an assessment adopting a more discursive approach covering a broader construct of pragmatics but a more resource intensive one. Integrating a discursive approach using performance-based tasks would, as a result allow pragmatic assessment researchers and developers to draw broader conclusions about L2 speakers’ pragmatic abilities in real settings. Interactionally-orientated tests using contextualized-performance based tasks (Grabowski, 2009; Youn, 2013) elicit test takers’ performances in a way that simulate real settings and those performances are of interest to the broader construct of pragmatics. The target of assessment is also broader than the traditional view of pragmatics. This approach would ultimately make it possible to provide English learners and teachers with richer information of L2 speakers’ pragmatic abilities.

As Youn (2015) argues, covering a broader construct and utilizing role play tasks should allow for more meaningful interpretation of the test scores as an indicator of L2 speakers’ pragmatic abilities. As such, it is desirable to implement performance-based assessments targeting a broader construct of pragmatics and to conduct further studies that follow Youn’s (2015) argument. However, it is equally important to consider instrument practicality because impractical instruments cannot be used despite their strengths described above. Even if scores interpretation were entirely justified, tests that
involve huge resource intensiveness (including human resources, financial resources),
or tests that considerably lower practicality, ultimately, lowers feasibility of its operation
or in an extreme case, the test is never operationalized before test use is considered.

The recent studies demonstrated broadening test construct of pragmatics, but
very few attempts up to date have been made to address the issue of maintaining
broader construct coverage considering instrument practicality in performance-based
assessment of pragmatics. Roever (2005) and Roever et al. (2014) addressed this
challenge, “Keep it social but practical” (McNamara & Roever, 2006, p. 63) that testing
of pragmatics has faced. However, their practical instruments did not allow them to
elicit online performances of oral extended discourse, structured all the way through
from the opening to the closing, which speakers themselves develop in real
communicative settings. Clearly, further studies are required to explore how we can
elicit and how to assess extended discourse performance that is of interest to the broader
construct of pragmatics more practically.

Another recent activity worthy of note in testing of pragmatics is test validation
adopting an argument-based approach to validation (Kane, 2006) as demonstrated by
Youn (2013, 2015) and Roever et al. (2014). It is eminently arguable that validity and
validation are vital to language testing. Previous studies reported results produced by a
range of quantitative and qualitative analyses (e.g., reliability estimates). Although the
validation in the previous studies was not systematically undertaken and thus the link
between their empirical findings and the conclusions made was not always logically
illustrated. However, their independent and overall findings provided us with useful
information in how their assessment instruments functioned and inspired subsequent
studies. Roever (2005) and Liu (2006) illustrated their explicit awareness of the
importance of validity for assessment of L2 pragmatics and demonstrated a more
systematic approach to validation, following Messick (1989).

Unlike the previous studies, Youn (2013) and Roever et al. (2014) demonstrated a concerted effort for validation within an argument-based framework (Kane, 2006). An argument-based approach enables us to draw a conclusion of what the test score means and to provide an account for that conclusion. In this approach to validation, pieces of evidence provided from separate analyses are integrated to structure arguments for multiple phases of validation as described in the previous section. Kane’s approach allows us to undertake validation more systematically than Messick’s (1989) validation template and can facilitate validation work with the clearer guidance. Following Kane (2006), both Roever et al. (2014) and Youn (2013) structured their validity argument and showed how useful test scores of their tests are and how the collected evidence supported their conclusion, more so than simply by reporting the results of different analyses.

As Roever et al. (2014) pointed out, despite the accumulated effort in the area of L2 pragmatics, almost none of the pragmatic tests have been used in practice and testing of pragmatics has been at research and development stage for more than three decades. However, the recent arguments and empirical studies would indicate that testing of pragmatics has started to direct its interest to an actual utilization of a pragmatic assessment battery. Grabowski (2009) and Youn (2013)’s discursive approach, for example, certainly could provide richer information of pragmatic abilities for test users and English learners. In addition, McNamara and Roever’s (2006) argument of practicality and construct coverage is vital for test administration to be realized by actual practitioners. Furthermore, Roever et al. (2014) and Youn’s (2013) validation approach highlights accountability of test developers to test users for how and why meaningful the test scores for inference as to L2 speakers’ pragmatic abilities.
the studies in the literature relevant to testing of pragmatics to date has addressed these issues simultaneously, something which should be a challenging but vital exercise to realize more meaningful and feasible assessment of L2 pragmatics.

### 2.6 Summary

This chapter has conceptualized the construct of pragmatics and has discussed the research methodology for L2 pragmatics and the testing of L2 pragmatics with relevance to the current studies. It also highlights the challenges that testing of L2 pragmatics faces. The literature reviewed in this chapter provided the rationale, theoretical and methodological frameworks for the current study.

### 2.7 Research Questions

The goal of the study is to develop and evaluate a test of L2 students’ oral pragmatic abilities relevant to language activities in the English-medium university domain. With the insights and the challenges provided by the previous studies, the researcher designed a discursively-orientated assessment instrument of L2 pragmatics and validated the possible use of the instrument by means of argument-based validation (Chapelle, 2008; Kane, 2006, 2013; Knoch & Elder, 2013). Unlike most validation studies, this study also addresses the issue of instrument practicality. The study is guided by the following main research questions:

1. To what extent does the assessment content reflect L2 students’ oral pragmatic abilities for use in university settings as the target domain? (Domain Description)
2. To what extent is the procedure for assigning test scores appropriate?
(Evaluation)

(3) To what extent does the assessment yield test results consistent across assessment contexts: task administration conditions, raters, and task sets?

(Generalization)

(4) To what extent can test takers’ test results be attributed to the construct of pragmatic abilities utilized for language activities at university? (Explanation)

(5) To what extent are the performances observed in the tasks indicative of the test-takers’ real pragmatic performances in the target domain? (Extrapolation)

(6) What are possible cut-off score points for educational purposes? (Utilization)

(7) How and to what extent can monologue tasks serve as an alternative to dialogue tasks? (Practicality)

Research questions (1) to (6) corresponds to the validity inferences, Domain Description, Evaluation, Generalization, Explanation, Extrapolation, and Utilization in the argument-based frameworks (Chapelle 2008; Kane, 2006; Knoch & Elder, 2013). Thus, the findings for the main questions (1) to (6) are intended to serve as the backings for warrants (Toulmin, 1958) to support each validity inference. These inferences supported with their underlying warrants are ultimately integrated to structure an argument to inform test users about:

- the usefulness of the test results for inferences about L2 students’ oral pragmatic abilities for university activities and
- the usefulness of the test results for test users in making decisions for educational purposes
Research question (6), which is concerned with test use, cannot be directly evaluated empirically as the current assessment instrument has not been used by test users. Therefore, research question (6) aims to make suggestions for future studies to empirically address these issues. Research question (7) is intended to explore instrument practicality.
Chapter 3 Research Design and Methodology

3.1 Introduction

This chapter describes the methodology adopted for the current study. I will first describe the structure of the research activities conducted in the study and in the subsequent sections I will introduce the instruments, the research participants, the procedures for the designed assessment and the data analyses.

3.2 Structure of the Research Activities

In this section, I will describe the overall research activities that involved the procedures, the analyses, the participants, and the data. Each of the procedures, the analyses, the participants, and the data is described in the subsequent sections. This section provides an overview of how these are connected for the aims of the study.

This study aimed to (a) develop an instrument to assess L2 students’ pragmatic abilities utilized in university contexts and to (b) investigate the collected evidence to interpret the meaning of the test scores and to (c) discuss the practicality of the instruments. First, the test tasks and the test administration procedure were developed through multiple stages of drafting and pilot studies. Additionally, the post-task questionnaire items were also developed to elicit the test takers’ perceptions. Developing test tasks may have been skipped if the study had employed the existing tasks available in the literature. However, in order to elicit meaningful performances from the test takers, the test tasks needed to be designed carefully by drawing on task situations simulating university settings and by adjusting pragmatic demands underlying the tasks. In this phase of the research activities, these instruments were initially drafted and then piloted prior to the main data collection.
The main data collection was subsequently conducted. The main data collection involved the administration of the developed test tasks and the post-task questionnaire items with individual test takers. The test takers provided their oral discourse data, their task preparation time and their perspectives. Prior to rating the test takers’ collected task performances, the rating criteria were developed based on the findings in the related literature and refined by the collected evidence in the above-mentioned data collection. The trained raters employed the rating criteria to convert the test takers’ task performances into test scores, which allowed for the Rasch analyses and further quantitative analyses using the Rasch analyses results.

The research activities are visually presented in Figure 3.1 below.

![Diagram showing research activities and data collected]

**Research activities**

- Drafting, piloting and revising the test tasks and the test administration procedure
- Developing questionnaire items to elicit test takers’ perspectives

The main data collection including:
- Task administration to the test takers
- Questions in the post task interviews

Rating including:
- Developing the rating criteria (based on test takers’ discourse performance applying conversation analysis)
- Training and Rating

- Quantitative analysis to explore test reliability, inter-rater reliability
- Many facet Rasch analyses
- Quantitative analyses using the Rasch analysis results

**Data collected**

- Oral discourse data
- Task preparation time
- Test takers’ perspectives

Test scores

Validation and addressing of practicality issues (Chapter 4)

*Figure 3.1 The research activities.*
3.3 Instruments

3.3.1 Test Task Design and Pilot Study

This section introduces the test task design, followed by the process of revising and developing the tasks. Test tasks play an important role as a device in language assessment to elicit meaningful task samples from test takers. Task design accordingly plays a crucial role in validation as we investigate and evaluate evidence for a test score interpretation and test score use based on the data obtained from the test takers’ test samples.

The overall specifications of the test tasks were essentially led by the rationale of the study. First and foremost, this study focused on assessing oral pragmatic features, integrating how speakers manage the extended discourse. Therefore, the tasks were expected to situate test takers in contexts where they handle pragmatic demands by the medium of oral extended discourse. The aims of the current study required oral role play tasks to elicit the target features of pragmatics from test takers. Simultaneously, the tasks needed to be administered logistically to a number of test takers.

As discussed in Chapter 2, role play tasks have been extensively and successfully utilized in previous empirical studies on L2 pragmatics. A strength of role play tasks is comparability of performances across speakers (Bardovi-Harling, 2013) and in eliciting discourse features (Roever, 2011). In the field of pragmatic assessment, role plays have received attention in recent studies as demonstrated by Grabowski (2009) and Youn (2013). These studies showed the strengths of utilizing role plays in reasonably constraining speakers’ oral extended discourse performance for assessment purposes while avoiding construct under-representation.

Equally important is the fact that the current study targeted and assessed L2
students’ oral pragmatic abilities for university activities. Applying Kane’s (2006) argument-based framework, the current study defined “Universe of Generalization” (p. 33) as L2 students’ language use to handle pragmatic demands in speaking contexts at English-medium university. Thus, the tasks were designed to simulate language activities at university. Test takers’ task performances on role plays simulating other situations (e.g., refusing a request from a client) may allow us to predict the target task performance, but the conclusion for the Domain Description inference would be made questionable.

Finally, the other fundamental purpose underpinning this study was to develop an instrument to measure the broader construct of L2 pragmatics (Roever, 2011) integrating discursive features in a valid way with practicality considered. An attempt to maintain construct representation inevitably decreases the instrument practicality, making the instruments difficult to utilize in reality. To address the issue of practicality required two modalities of oral role play tasks with different resource-intensiveness created for a comparison. Thus, it was decided to create tasks whose modality was considered to be less resource intensive for its administration and thus more practical and to investigate to what extent and how those tasks could be an alternative to dialogue role plays. These role plays were thought to be less practical but would be able to avoid construct under-representation.

Taking these facts into consideration, the dialogue and the monologue oral role play tasks simulating language activities where students handled pragmatic demands at university were developed. The assignment of test tasks to test takers were also considered for the aim of addressing practicality. In the following sections, I will detail (a) how the target language activities domain was specified, (b) how they were operationalized to pose a certain degree of pragmatic demands, (c) how the role play
tasks were operationalized to elicit oral extended discourse from the test takers as the study intended, and (d) how the tasks were assigned to test takers. These may appear like separate considerations in the task specification and are described in the following subsections independently. However, these were, in fact, interrelated. For example, if the pragmatic demands embedded in the task were not high, the task may be found easy for the test takers and therefore may not be able to elicit extended discourse performance. Also, the operationalization of the pragmatic demands reflected in the tasks was partly restricted by possible communicative situations in the target domain (e.g., who to address, for what purposes in that domain). Therefore, these fundamental considerations needed to be taken iteratively into the test task development.

### 3.3.2 Specifications of the Target Test Domain

According to Kane (2006), the score interpretations and the conclusion in validation arguments are made based on the specified target domain (e.g., academic settings, business settings) to which collected samples of performances are extrapolated. The present study set university situations in English speaking countries as the target domain. The academic domain has been widely targeted by existing tests developed for university admission (e.g., TOEFL iBT, IELTS), and was the subject of a recent validation study of pragmatics assessment (Youn, 2013, 2015), and was also identified by Norris, Brown, Hudson and Yoshioka (1998) as one of the social areas from which task-based tasks can be created. Norris et al. (1998) divided the university context into four subcategories: application to the university, registration at the university, in-class activities, and contact with lecturers and access to study materials. A wide range of situations and activities in general occur in the university context, but the target language activity domain was not specifically defined in TOEFL iBT, IELTS, and Youn
The university domain in this study was more constrained than those of TOEFL, IELTS, and the examples in the literature, in order to make the conclusions based on the observed performances more convincing by operationalizing conditions in the target domain. The target domain in this study was defined in terms of the chronological order of the events, in which the activities supposedly occur within or outside official class periods, what events students engage in, as well as with what groups of people the students converse. Language activities which can occur outside the class have been recognized as possible language activity settings at university (e.g., describing an event and explaining a personal preference related to social-interpersonal settings) (Xi, 2015). The target domain in this study was thus specified by the following points within, by, and in which the assumed activities are likely to occur (Table 3.1).

Table 3.1
Summary of Domain Specifications

<table>
<thead>
<tr>
<th>Specification</th>
<th>Target in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronological frame</td>
<td>During the semester</td>
</tr>
<tr>
<td>Where activities occur</td>
<td>Outside of official class period</td>
</tr>
<tr>
<td>People to address</td>
<td>Professors, Administrators, Peer students</td>
</tr>
<tr>
<td>Ultimate purposes of the concrete activities</td>
<td>To prevent failure and to promote academic performance in the subjects</td>
</tr>
</tbody>
</table>

In this domain, university students contact professors, administrators and other students, to successfully complete the academic work required for the subject. For example, students may be forced to request an extension from their teachers due to an unexpected and unpreventable event. This action would help prevent failure in a subject or the undermining of the professor’s evaluation of the student. University students are expected to engage in events and concrete activities for the purposes described in Figure 3.2, based on which the monologue and dialogue role play tasks were created. These possible events and concrete activities were identified, based on the experiences of the
researcher and his colleagues, all of whom had studied at a university in Australia at the
time of this brainstorming stage of the task development. In the brainstorming stage, a
range of language activities at university were listed, including activities that students
experience in, before or upon admission to a university (e.g., enquiry to the university).
However, it was considered that university students were most active on campus during
the semester. Researchers’ own experiences can be a rich source of information to
utilize to design test tasks as demonstrated in Timpe (2013) and these experiences were
realized in the current study.

Figure 3.2 displays how a university student, peer, administrator, and a professor
are related to each other by their need to negotiate certain matters. University students
discuss a variety of issues regarding their academic work with peers, administrators and
professors, respectively. This happens so as to prevent failure or to achieve better
performance in their chosen subjects.

In these situations, it was assumed that university students are required to take
various actions such as making requests, proposals to the listener, providing opinions,
showing considerations for the listener for a mutual agreement with the intended
addressee regarding the matter.

![Diagram of interlocutors and matters in the specified academic domain]

*Figure 3.2. The interlocutors and matters in the specified academic domain.*
It should also be noted that these tasks were assumed to occur in various university contexts in English speaking countries in general, not in a particular institution or educational program. The tasks were not tailored for students from particular countries nor for students in particular tertiary disciplines. These task situations simulated situations that students may possibly encounter as students. As well, the designed situations to deal with pragmatic demands were connected in terms of problem-solving activities to support academic work, not activities for entertainment (e.g., booking a table for a party for an end-of-the semester party). Activities which are distantly related to subjects such as matters dealing with scholarships and a party with classmates were also not included in the task situations. Even though these activities could be assumed to occur outside of a class period, they were supposed to be closely associated with the successful completion of the subjects that a student pursues.

3.3.3 The Specification for Test Tasks

The data targeted in the present study was online, oral performances of L2 students of English. This study employed two types of extended discourse tasks, namely dialogue and monologue open role plays. Being informed by the previous sub-section describing the specified target language activity domain, the communicative goals in the university settings were embedded in both types of tasks. The participants situated themselves in the specified context through input from the task prompts. In the tasks on both modalities, the test takers’ role was essentially fixed as a university student. In dialogue role plays, the test takers interacted with a trained interlocutor, who played the roles of a professor, an administrator or a peer student. The social context (e.g., situation, and the physical person to talk to) was also specified in each role play task. All of the tasks simulated real situations and were arranged in order to simulate the real
chronological flow of events from the beginning to the end of a semester at university.

This order provided more authentic chronological associations of the situations, allowing test takers to remember more easily what they are supposed to do. Twelve tasks were created. Table 3.2 presents the overall framework of the task arrangement.

Table 3.2
*A Framework of Tasks*

<table>
<thead>
<tr>
<th></th>
<th>Addressee</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Professor or</td>
<td>Dialogue</td>
<td>Monologue</td>
<td>Topic 1</td>
</tr>
<tr>
<td>2</td>
<td>Administrator or</td>
<td>Monologue</td>
<td>Dialogue</td>
<td>Topic 2</td>
</tr>
<tr>
<td>3</td>
<td>Peer student</td>
<td>Dialogue</td>
<td>Monologue</td>
<td>Topic 3</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Monologue</td>
<td>Dialogue</td>
<td>Topic 4</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Dialogue</td>
<td>Monologue</td>
<td>Topic 5</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Monologue</td>
<td>Dialogue</td>
<td>Topic 6</td>
</tr>
</tbody>
</table>

*Note.* This framework was applied through the pre-pilot studies and the pilot study, in which an original task was replaced by another and the task prompts were revised.

In order to situate test takers in the context where they handle pragmatic demands, in all of the tasks, three context variables, identified by Brown and Levinson (1987) - imposition, relative power of the interlocutors (e.g., professor or peer student), social distance (familiarity) and the degree of imposition - that can constrain test takers’ performances were operationalized. All of the situations were designed to be of a middle or higher imposition. Low imposition situations were not included, such as where a student asks a classmate to lend the classmate’s eraser. Relative power of the interlocutor over the speaker was set to be equal or higher and social distance between the speakers was middle or low. This operation allowed the tasks to draw out both the test takers’ linguistic resources required to deliver the intended meaning for the situation and their linguistic choices characterized by social and cultural norms (Leech, 1983).

The tasks were designed to make it possible to compare the dialogue performances constructed with an interlocutor and those based on individual
contributions in monologue performances with the same social context. Table 3.2 above shows that the dialogue tasks correspond to the monologue tasks in each line sharing the same topic and the person whom a test taker is to address. The difference between dialogue tasks and monologue tasks was the physical presence of the interlocutor to interact with in dialogic situations and its absence in monologues. In both types of tasks, the test takers situated themselves in the given contexts so as to recognize the embedded pragmatic demands, to handle and to structure an oral extended discourse performance from opening to closing (Roever, 2011). The elicited oral discourse performance therefore was also expected to reflect how the test taker recognized the context and to involve actions (e.g., explanation of the matter, making requests, expressing considerations for others) that they considered appropriate or necessary for the context.

With regard to the communication modes, dialogue assumed face-to-face conversations and monologue assumed communication through the medium of voicemail. Voicemail might not be necessarily as familiar to people in general as other modes in everyday communication. The decision on voicemail was based on the researchers’ personal observation of communication practices at university, where voice-recorded messages and delivering them through an electronic medium were found to be a common means used by academic staff to give feedback to students.

In the dialogue tasks, different instructions were given to the test takers and to the trained role play interlocutors (to be discussed later) respectively so that they engage in the conversation naturally and in such a way that a negotiation for mutual agreement occurred (for the task instructions to the test takers and to the interlocutors, see APPENDICES B, C, and D). The test takers described in the task prompt as a university student addressing the intended interlocutor were expected to lead and manage the interaction for the communicative goals. As the general instruction (see APPENDIX A,
which was provided for the test takers before they started the tasks) stated, what to say and how to say it were up to the test takers. Because the test takers were not given a pre-planned scenario, when to say it (what test takers want to say) in their extended discourse performance was also left up to the test takers. For the role play interlocutors, what to say and how to say it were controlled to minimize the difference among the interlocutors (in the extent to which an interlocutor should be accommodating towards the test takers and in how an interlocutor should impose pragmatic demands), but they were allowed to judge when in the ongoing discourse context to say it.

The purpose of this task was to elicit from the test takers’ several social actions while handling pragmatic demands so as to achieve the underlying communicative goals. To perform in the dialogue tasks, the interlocutors were trained to interact in ways that required test takers to interpret the interlocutor’s implied meaning (unwillingness to agree on the test takers’ request or being unsure of offering help for the test taker or inconvenience for the interlocutor). They played a role of extending the test takers’ discourse so that the discourse data reflected pragmatic features. As described above, the role play tasks simulated problem-solving situations. If the interlocutors were instructed to offer help before the test takers initiating an action to ask for help or to be too accommodating, the task performance would finish without being adequately extended. Conversely, if the test takers’ initiated or projected actions were firmly refused as soon as they addressed the interlocutor, the discourse performance might be terminated. Therefore, the instructions for the interlocutors to control their performances were carefully developed so as to elicit meaningful data from the test takers, even though the test takers’ performances were not as overtly controlled as those with the pre-planned scenarios.

For both dialogue and monologue tasks, the test takers were prompted to
recognize the pragmatic demands underlying the tasks and to initiate the opening of each task situation. Both modalities of the tasks allowed them to conclude the performance in their own ways when they thought it was finished. Therefore, to perform under both dialogue and monologue conditions, it was left to the test takers as to how they should construct the extended discourse from the opening to the end of the performance to achieve the communicative goals.

### 3.3.4 The Process of Drafting and Piloting the Tasks

This section introduces the process through which the initially drafted tasks were refined. After the initial phase of task development, three phases of pre-pilot studies and one pilot study were carried out. Each phase had different purposes and made distinct contributions to the task development. All phases of these task development activities were initiated and conducted by the researcher himself through face-to-face meetings with the participants.

#### 3.3.4.1 The initial phase of the test task development

As mentioned above, the initial brainstorming of the task situations was conducted based on the researcher’s experiences in the target context and on discussions with his colleagues in applied linguistics in Australia as L2 speakers of English. The discussions were made with respect to the authenticity of the situations, occurring in a university domain considering when, where, for what and with whom students in university environments would encounter the situation and the activity. Through the discussions, the possible topics were identified as listed in Table 3.3 below.
### Table 3.3
**List of Possible Activities**

<table>
<thead>
<tr>
<th>Chronological frame</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before or on admission, enrollment</td>
<td>application, enquiry</td>
</tr>
<tr>
<td></td>
<td>registration-related activities such as activation of account (e.g, email), registration of personal information, application for scholarship, registration of subjects</td>
</tr>
<tr>
<td></td>
<td>transferring credits from another university, request for exemption, enquiry</td>
</tr>
<tr>
<td>During a semester</td>
<td>self-introduction, discussion, auditing lectures (in class activities)</td>
</tr>
<tr>
<td></td>
<td>change of a subject, consultation about subjects</td>
</tr>
<tr>
<td></td>
<td>making a pair for a presentation</td>
</tr>
<tr>
<td></td>
<td>changing the assigned group for presentation</td>
</tr>
<tr>
<td></td>
<td>proposing a presentation plan and getting approval from teacher</td>
</tr>
<tr>
<td></td>
<td>preparation of necessary device for presentation</td>
</tr>
<tr>
<td></td>
<td>request/proposal/announcement of a self-study group</td>
</tr>
<tr>
<td></td>
<td>self-introduction of your interest, discussion at a self-study group</td>
</tr>
<tr>
<td></td>
<td>request for reschedule of meeting</td>
</tr>
<tr>
<td></td>
<td>request for extension of due date</td>
</tr>
<tr>
<td></td>
<td>request for re-mark of the assignment</td>
</tr>
</tbody>
</table>

Some situations involved multiple people at university. For example, an enquiry about an application can be made to an administrator if it was an administration-related matter. An enquiry may also be directed to academic staff coordinating a program, if it was related to the contents of the offered subject. Several practices should be possible across universities with regard to who conventionally manages the various university activities. At this stage, the brainstormed ideas included activities relatively distantly related to academic subjects, such as activation of account (e.g., email) and application for scholarship. The range of university activities were broader than the specific domain ultimately defined in this study (see Table 3.1 above), but these broad ideas which were brainstormed, made possible the pre-pilot as well as the pilot studies described in the following sub-sections.
3.3.4.2 Pre-pilot study 1

The first pre-pilot study was conducted to tentatively fix the designated number of tasks (twelve role play tasks created from six topic as in Table 3.2 above), according to the fundamental considerations for the test tasks argued in section 3.3.1. This phase of the pre-pilot study involved one ESL-certified teacher, a native speaker of English, who served as one of the interlocutors for the pilot study (the final phase of the series of the pre-pilot and the pilot studies described in section 3.3.4.5 later) and for the main data collection (see section 3.5.1 later), and as one of the trained raters (see section 3.4.3 later). When piloting the tasks, the researcher played the role of a test taker. The piloting of the tasks was conducted task by task. The role play performances were not audio-recorded.

Drafting of the tasks and piloting the role play tasks were conducted iteratively. The role play tasks were piloted primarily to confirm that each of the drafted tasks were able to elicit adequate amounts of oral extended discourse samples, representing pragmatic features. At the same time, the researcher had discussions with the interlocutor as to whether the dialogue role play instructions to the interlocutors was sufficient and clear (for the final version of the instructions to the interlocutors, see APPENDIX D). This required an iterative process of drafting and piloting, where the interlocutor and the researcher trialed the drafted tasks, simulating the testing situation. This was followed by discussions on how the task could be revised or whether it should be discarded. Most of the task topics used for the final version (see APPENDICES B and C) were fixed in this phase of the pre-pilot study. The selected tasks were further edited by the researcher in the form of task prompts for the next pre-pilot study.
3.3.4.3 Pre-pilot study 2

After twelve tasks in the two sets were tentatively fixed, pre-pilot study 2 was conducted with three doctoral students (L2 speakers of English) in applied linguistics. The purpose of pre-pilot study 2 was to confirm whether the task prompts could be clear enough for future test takers to produce speaking performances as intended by the researcher and to estimate approximate time necessary for administrating six tasks per test taker. In this phase, the researcher played the role of an interlocutor and the doctoral students were involved as the test takers. The thorough trialing of the tasks in pre-pilot study 1 made the researcher realize that the drafted tasks could at least function to elicit the expected discourse samples from the target test takers. It was also considered significant to pilot the tasks on participants who were not familiar with the drafted tasks. The researcher met each of the three participants individually on different dates.

In administering pre-pilot study 2, the researcher audio-recorded the task performances of the two participants, who agreed with the exercise of audio-recording. The purpose of this recording was not for the researcher to analyze the discourse data but to pilot the administration simulating the real test situation, which required the interlocutors to operate an audio-recording device as well as to act as a conversation partner for the test takers in the dialogue tasks and to give instructions to them. The three participants conducted six tasks in either task set 1 or task set 2. Thus, both of these two task sets were piloted at least once.

It was confirmed that no major modifications were necessary. The performances elicited by the tasks were within the researcher’s expectations. All the information, including the general instructions, the simulated university situations as described in the task prompts and what test takers were required to do in the tasks was clear to them. However, as they were all equally highly proficient in English as L2 speakers, it was
deemed crucial to pilot the tasks and the test administration with participants with a wider range of proficiency targeted for the main study. This led to the decision to conduct the pilot study (the final phase of the task development) with the 12 participants described in subsection 3.3.4.5 later. Prior to this final stage, it was fortunate to have another phase (pre-pilot study 3) where the tasks were reviewed by the domain experts and the experts on language assessment as described below.

3.3.4.4 Pre-pilot study 3

Pre-pilot study 3 involved eight group members of a research seminar at the University of Melbourne for language assessment. They consisted of academic staff on language assessment and doctoral students including the researcher himself on language assessment and on the area related to language studies. The same tasks used for the previous pre-pilot study were piloted. The eight members piloted the tasks in pairs (one participant played the role of a test taker and the other played the role of an interlocutor), followed by discussions with all of the members about their findings and suggestions for revising the tasks. The activities in this pre-pilot study were not audio-recorded as the primary aim of this pilot study was to obtain feedback on the test task contents from the language assessment experts’ point of view. Unlike pre-pilot study 2, all of the participants gathered in the same venue at the same time with the researcher.

Their feedback suggested that overall, the designed tasks functioned properly to elicit the intended discourse samples from the main test takers. Some suggestions for minor modifications were also made to enhance the clarity of the information written in the task prompts. In addition, following their advice, one task situation was replaced with another situation (the current Professor topic 1 in APPENDICES B and C) which was deemed to be clearer to the test takers and able to pose pragmatic demands on them.
so as to elicit extended discourse more naturally.

Equally significant advice from the experts with regard to providing general instructions to test takers to give an overall idea of what to do in the tasks. The final version of the general instruction is provided in APPENDIX A refined by the advice from the seminar members. In particular, the following instruction to the test takers was made possible by their advice:

Some of the role play situations are connected, but each role play is independent. Therefore, your performance in the previous tasks will not affect what you are going to do in the next tasks.

Their advice made sense as the task situations were in fact connected by a theme that a university student takes actions for an academic subject that he or she is taking during the semester. The activities described in the task situations also simulated the authentic chronological flow of events in university life. Therefore, the test takers would be aware that the next task situation was linked to the previous task situation. If a test taker found their performance insufficient on a task, he or she might lose motivation to work on the subsequent tasks.

The insights and experiences of the participants were fully utilized for the minor and major modifications on the task instruments to elicit more meaningful task samples from test takers. After one of the task situations was replaced by the final version of Professor topic 1, the newly introduced task was trialed with the interlocutor involved in pre-pilot study 1 before embarking on the pilot study (the final stage of the series of the pre-pilot and pilot studies) discussed in the next subsection.
3.3.4.5 The pilot study

After the tasks were tentatively fixed through the three stages of pre-pilot study, the instruments were piloted with L2 speakers as pilot test takers in Australia (N=12). Each test taker was compensated for their efforts with AUD$20.

The primary purpose of this pilot study was to confirm that the tasks were able to elicit performance from those test takers who were assumed to be closer to the target population of test takers for the main study, with respect to their background (e.g., English proficiency levels and whether they had any experiences in university studies). Different from the test takers in the proceeding stages of the task development, the final stage of the pilot study involved the test takers with a range of English proficiency levels and their academic backgrounds. Their proficiency levels were equivalent to near native speaker to IELTS 5 or below. Their social background and professional or study area were not restricted to language studies and language assessment. Except for one test taker, they were all university students, English-medium university graduates or language school students. The English proficiency of two test takers was far below university-entry level, having had no experience in higher education settings in English speaking countries. The number of participants for each level indicated by IELTS levels was well balanced according to their self-reported official proficiency test scores.

The two sets of the tasks (set 1 and set 2) as for the main data collection were piloted. Thus, each pilot test taker completed six tasks in either set 1 or set 2. The task arrangement, the number of the tasks assigned to each test taker, the assignment of the two sets of the tasks to each test taker (see Table 3.2 for detail) and the test administration were conducted in the same manner as the main data collection, in which the interlocutors as a task conductor played the interlocutor role in the role plays and audio-recorded the test takers’ task performances. The test takers were given three
minutes to prepare for each task. They were allowed to start the tasks whenever they thought that they were ready, without having to use all of the given preparation time. The interlocutors, playing the role of the interlocutor in the role plays and administrating the tasks, were two ESL-certified native speakers of English. One of them was the interlocutor described in pre-pilot study 1 (see section 3.3.4.2), who conducted the task administration for most of the test taker participants. The task administration was conducted individually for the test takers and their performances were all audio-recorded.

After implementing the tasks with all of the test takers in this pilot study, the collected speech samples were analyzed qualitatively. It was confirmed that in general, pragmatic demands underlying the designed tasks adequately elicited extended discourse from the pilot test takers and that the instructions for the interlocutors were clear to them except for the two least proficient test takers who had no previous experiences of language activities at English medium university. Another important finding was that the tasks were able to elicit the test takers’ pragmatic features clearly discriminating the test takers, which could have been confirmed by piloting the tasks with a varied background of test takers. However, at the same time, an issue was discovered in one task situation in current Topics 3 and 4 (Peer student situation). The oral speech of some test takers elicited by the original version was not as extended as the researcher expected. Therefore, another line of complication was added into the task prompt (for the final version, see APPENDICES B and C).

For the test administration, the time that pilot test takers’ time spent to prepare for the tasks resulted in the decision to increase the given preparation time from three minutes to five minutes per task, although five minutes were considered to be generous for test takers. Most of the test takers did not use the maximum amount of the given
preparation time for most of the tasks. However, some students used the full amount of preparation time for particular tasks. Considering the slightly increased information in one task as discussed above, and the ways of test takers’ using the task preparation time, it was deemed safer to give five minutes, in order not to overly constrain the test takers’ pragmatic performances. It was also assumed that a majority of the main test takers would not use the maximum amount of five minutes in most tasks, even if they were allowed to do so.

3.3.4.6 Outcomes of the pre-pilot and the pilot studies

An appropriately designed and developed set of assessment instruments was essential to obtain meaningful results and to allow for a more accurate measurement of the target pragmatics construct as well as for more convincing interpretations of the test scores. Thus, the pre-pilot and pilot studies played a crucial role in allowing this research to elicit more meaningful data. Equally significantly, they ensured the confidence in facilitating the task administration as intended by the study. The purposes and the outcomes of these phases of the task development activities are summarized in Table 3.4 below.
Table 3.4
The Purposes and Outcomes of the Drafting and Piloting Phases

<table>
<thead>
<tr>
<th>Phase</th>
<th>Purpose</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial drafting</td>
<td>List possible task situations</td>
<td>Brain stormed ideas from which test tasks were created</td>
</tr>
<tr>
<td>Pre-pilot study 1</td>
<td>Transform the brain stormed ideas into the concrete task situations and the task instruments</td>
<td>The tasks in the two sets (as in the framework of Table 3.2) were created</td>
</tr>
<tr>
<td>Pre-pilot study 2</td>
<td>Evaluate the feasibility of the tentatively selected tasks</td>
<td>Confirmation of the extended discourse performance as intended and approximate time needed for completing the tasks for test takers</td>
</tr>
<tr>
<td>Pre-pilot study 3</td>
<td>Have the test tasks reviewed by experts’ points of view</td>
<td>Revision of one task situation and configuration of the general instruction for test takers</td>
</tr>
</tbody>
</table>
| Pilot study      | Evaluate the feasibility of the test instrument by administrating it to a range of backgrounds’ pilot test takers | Decision to slightly increase the information in one task situation  
|                  |                                                                        | Decision to increase the task preparation time |

The feedback and the task performances of the pre-pilot studies and of the pilot study were encouraging. Each phase of these played a distinctively important role towards refining the test tasks. Although one task situation was replaced by another, as a result of the discussions conducted in pre-pilot study 3, the remaining task situations fixed at the stage of pre-pilot study 1 were chosen for the final versions of the test tasks. The major and minor revisions were made possible by the useful and encouraging suggestions obtained in the series of pilot studies.

In addition, these piloting activities helped to ensure feasibility of the test administration. During all the pre-pilot and pilot studies, the time spent for the administration of the six tasks per test taker was confirmed to be within approximately 40 minutes or less, including the time allocated to provide the general instructions to the
test takers. The consideration for the time required for test administration was particularly important for the current study. To administer the test, it was necessary to make compromises regarding the constraints of the resources needed for the study and the quality of the test takers’ test samples that the study expected. In order to ensure the quality of the test takers’ data, administrating multiple tasks was considered to be desirable in order to obtain richer data within the allocated quantity. It was also expected that sufficient amount of time was necessary for the test takers to understand and situate themselves in the given communicative situation. To reflect the pragmatic demand in the tasks, complications needed to be injected in the task situations. Furthermore, the study fundamentally targeted a broader construct of oral pragmatics beyond offline production of speech acts. In the case of the current study, oral pragmatics in test takers’ extended discourse was expected to consist of discourse performances from the opening through to the closing (Roever, 2011), since it was deemed odd to assess a performance which was abruptly cut off by a time limitation.

These considerations required a certain amount of time for the test administration to ensure that the test takers’ discourse data adequately reflected what Bardovi-Harling (2013) said about L2 pragmatics as “how-to-say-what-to-whom-when” (pp. 68-69). However, since the study expected a large number of test takers, the amount of time required for the test administration per test taker was also needed to be considered. If completing the test had required several hours per test taker, which would inevitably increase the resource-intensity, it would have exceeded the allocated research budget. Through the pilot study, it was confirmed that the time for administering six tasks did not exceed 50 minutes per test taker and that the length of time was generally reasonable.
3.3.5 Final Test

The test delivery required a face-to-face meeting of a test taker and an interlocutor, invigilated by the researcher. The overall framework (the number of test sets, the number of the tasks, the arrangement of the tasks) was the same as practiced for pre-pilot studies 2, 3 and the pilot study as shown Table 3.5 below.

<table>
<thead>
<tr>
<th>Addressee</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Professor</td>
<td>Dialogue (conversation)</td>
<td>Monologue (voicemail)</td>
<td>Get a signature to change classes</td>
</tr>
<tr>
<td>2 Administrator</td>
<td>Monologue (voicemail)</td>
<td>Dialogue (conversation)</td>
<td>Change the registration of the class</td>
</tr>
<tr>
<td>3 Peer student</td>
<td>Dialogue (conversation)</td>
<td>Monologue (voicemail)</td>
<td>Make pairs for a presentation</td>
</tr>
<tr>
<td>4 Peer student</td>
<td>Monologue (voicemail)</td>
<td>Dialogue (conversation)</td>
<td>Reorganize a presentation</td>
</tr>
<tr>
<td>5 Administrator</td>
<td>Dialogue (conversation)</td>
<td>Monologue (voicemail)</td>
<td>Solve a projector problem</td>
</tr>
<tr>
<td>6 Professor</td>
<td>Monologue (voicemail)</td>
<td>Dialogue (conversation)</td>
<td>Avoid losing marks for an assignment</td>
</tr>
</tbody>
</table>

*Note.* The title of the topics is identical to those on the task prompts (APPENDICES B and C).

The number of the intended addressees, (professor, administrator, and peer student) and the two task modalities (dialogue and monologue) were equally distributed in each task set and across the two task sets. The two types of task modalities shared the same topic for each of six topics. The assignment of each task situation in this framework (Table 3.5 above) was not decided randomly but was intended to simulate the real chronological order of the events in the target domain. The earlier tasks created from topics 1, 2 and 3 simulated university activities (class registration-related activities) which were likely to occur at the beginning of the semester. Topic 6 was intended to reflect the end of the semester. The test takers played the role of a university
student in these tasks. The task prompts, comprising the topics, the task situations and
the tasks were provided in APPENDICIES B and C. In the dialogue situations, the test
takers were instructed to initiate the conversation in each task. These task sets were
created so as to assign either task set 1 or task set 2 to each test taker, not both
modalities of the tasks sharing the same topic. For both task sets, the test takers were
prompted to the next task with the different task modality after completing one task.

For a general rule, the test takers were allowed to take notes during the
preparation time and to refer to the written notes while performing the tasks. They were
also informed that they could conclude the performance (in both modalities of the tasks)
in their own way, as specified in general instructions to the test takers (APPENDIX A),
so that their performances were not abruptly concluded by a time limitation.

3.3.6 Post-task Questionnaire to Elicit Test Takers’ Perceptions

As listed in the research activities (Figure 3.1 above), the main study also
encompassed the post-task questionnaire with each test taker conducted by the
researcher. The background questions included their scores of official English
proficiency tests (either IELTS or TOEFL iBT), the length of their exposure to higher
education settings in English speaking countries, and their academic disciplines.

The post-task questionnaire was also intended to elicit the test takers’
perceptions, which were expected to serve as a source of evidence to support the
validity argument (Chapelle, 2008; Roever, et al., 2014). The test takers were actually
engaging in language activities relevant to the academic environments as a student, or
pre-entry students who were studying to enter that domain. Therefore, their perceptions
were considered to be invaluable. The post-task questionnaire was structured so as to
allow for quantitative analysis. The post-task questionnaire covered (a) questions to
identify possible construct-irrelevant factors (Messick, 1989) in the test takers’ task performances, which asked the test takers to provide their perceptions dichotomously (Yes or No, allowing the test takers to report their answers in other ways as well). Likert scale questionnaire included (b) how similar they perceived their language use in their task performances to be to that in reality, (c) the test takers’ awareness of the constructs of pragmatics in communication for university activities, and (d) self-assessments of pragmatic performances in the tasks.

These questions and the pragmatic features described within them were written in plain language, avoiding such terms as “pragmalinguistics,” “speech acts,” and “construct-irrelevant factors,” so that the test takers could easily comprehend what was required of them. The pragmatic features described in the question (c) were the same as those in question (d). The items (pragmatic features) constructing questions (c) and (d) were intended to elicit the test takers’ perspectives about their language use in an academic environment when handling pragmatic demands (as specified in the task design). The items did not, admittedly, cover all of the features discussed in the literature on pragmatics. For the items, the view that test takers would be able to clearly understand the meaning of each item asked to make judgements on, was taken into consideration and was prioritized. The questionnaire items are shown in APPENDIX E.

3.4 Participants

3.4.1 Test Takers

A total of 67 international L2 students in Australian English-medium education participated in the main study as test takers. They included university students (N=44) and language school students (N=23) at a university preparation course at a language
school geared to satisfy English language requirement for university admission. The test takers’ English proficiency levels ranged from an equivalent of IELTS 5 level to near-native level. The pre-entry students were receiving an EAP education to acquire English skills required for university study. Successful completion of the course was recognized as satisfying English requirements for university admission.

The test takers were recruited by direct communication by the researcher himself and by teachers at university and at a language school. Either the researcher had personally visited classes at a university and at a language school, or teachers there distributed the information (the call for participants) to students directly. Those who contacted the researcher and agreed to participate were recruited. There was also a small number of recruits who were introduced to the test by previous participants in the study. The test takers’ length of stay in English speaking countries ranged from one week to 22 years. The test takers’ first languages (10 native language backgrounds were reported), their gender, their major, their length of stay in English speaking countries and the length of exposure that they had to higher education settings in English speaking countries and in other countries, were diverse. Table 3.6 below shows the distribution of the test takers’ first languages. The largest first language group was Chinese, followed by Japanese and Spanish. The test takers’ age range was from 18 to 46. Of the 67 test takers, 33% were male and 67% were female. Each test taker was compensated for their efforts with AUD$20.

Table 3.6
Test Population L1 Distribution

<table>
<thead>
<tr>
<th>L1</th>
<th>N</th>
<th>Subtotal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>Japanese</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Spanish</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Danish, Farsi, Indonesian, Korean, Russian, Urdu</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 3.7 presents the test takers’ demographic information by academic background (either university student or pre-entry student) and by English proficiency levels.

Table 3.7
*Grouping of the Test Takers*

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>Hypotheses about performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>University Student</strong></td>
<td><strong>Advanced university student</strong> (N=9)</td>
<td>Highly proficient in speaking and in general. Those whose self-reported IELTS overall or speaking band was 8 or higher and/or those judged by interlocutors intuitively to be the equivalent level to near native speakers in speaking proficiency. Pragmatically more competent than Pre-entry student because they are more proficient, have experiences at the target domain. Within the university students, Advanced university students are pragmatically more competent than Other university students as they are more proficient.</td>
</tr>
<tr>
<td><strong>Other university student</strong> (N=35)</td>
<td>Students who did not meet the condition of Advanced university students. Self-reported IELTS levels were 6 to 7. None exceeded Advanced university students in both general proficiency and speaking proficiency according to their self-reported IELTS levels.</td>
<td></td>
</tr>
<tr>
<td><strong>Pre-entry student</strong></td>
<td><strong>Pre-entry student</strong> (N=23)</td>
<td>Students studying at a language school to satisfy English requirement for university admission. Most test takers’ self-reported IELTS levels ranged from 5 to 6. Most of them had experience of 3 months or less living in English-speaking countries, a time span which also included sightseeing and various other activities. The Pre-entry students as a whole would not outperform university students pragmatically.</td>
</tr>
</tbody>
</table>

They were divided into three groups according to the two variables (experience of language activities at target settings and proficiency). The hypotheses were suggested by referring to the literature, which has identified proficiency level and experience of
target exposure as the major factors accounting for L2 speakers’ pragmatic abilities. Pre-entry students had substantially smaller lengths of stay in English-speaking countries than the two groups of university students, in addition to Pre-entry students’ lack of experiences of language activities in English-medium higher education. Most of the Pre-entry students (21 out of 23) reported three months or less of time spent in those countries. Periods of staying in English-speaking countries varied among the university students. The majority of the university students (35 out of 44) reported staying in those countries for four months or more. The group of Advanced university students as a whole and the group of Other university students as whole respectively had considerably longer periods of time spent in English-speaking countries than the group of Pre-entry students.

In the current study, the Pre-entry students were preparing to satisfy English language requirement as requisites to enter a university domain. Thus, it was hypothesized that (a) university students as a whole are pragmatically more competent than Pre-entry students as a whole, in view of both English proficiency level and experience of university activities in the domain (b) the Advanced university students are pragmatically more proficient than the Other university students as well as the Pre-entry students and (c), the difference between the Pre-entry students and Advanced university students was clearly seen in pragmatic competence. Although it was difficult to predict to what extent the university students were advantaged by the proficiency level and the experience of language activities in the target domain, it was assumed unlikely that the Pre-entry students outperform the Advanced university students. The test takers were classified into their respective groups in order to investigate whether

---

1 Simply staying in an English-speaking country does not mean that the person is actually exposed to English. It is possible to stay away from English-speaking environments even if he or she is in that country. It is, however, less likely that university students in English-medium academic environments avoid passive exposure to English and active use of English.
and the extent to which the designed assessment of pragmatics for the current study was able to detect the test takers’ pragmatic abilities as made by the assumptions above.

### 3.4.2 Interlocutors

The test administration required an interlocutor, who gave general instructions (APPENDIX A) to the test taker before the tasks, handed the task prompts (APPENDIX B or C) to them, audio-recorded each task performance, and played the role of a conversation partner in the dialogue tasks. Four interlocutors were involved in the main study. They were all native speakers of English who had a number of years of experience as ESL and/or EFL instructors. At the time of the data collection, one interlocutor was the holder of a PhD in language assessment. Another was a PhD candidate in linguistics and the other two were ESL teachers, one of whom was involved in pre-pilot study 1 and the pilot study (see sub-sections 3.3.4.2 and 3.3.4.5). After being informed of the background of the study, the interlocutors were trained to familiarize themselves to the interlocutors’ role for the test administration and how they should perform in the dialogue tasks as the conversation partner (see APPENDIX D for the prompts to interlocutors). No interlocutor participants reported acquaintanceship with the test takers assigned to them.

### 3.4.3 Raters

Three raters participated in the study including the researcher (Rater 1). Rater 2 was an ESL-certified teacher who served in pre-pilot study 1, the pilot and main studies as one of the role play interlocutors. Rater 3 was a postgraduate student in language studies at a university in Australia. Rater 2 (female) and Rater 3 (male) were both native speakers of English and educated at tertiary level in Australia. All three raters had
shared characteristics in having tertiary-level education in Australia, thus being familiar with the context from which the role play tasks situations were designed and in having experience of teaching English in the same EFL country. The raters were trained by the researcher (Rater 1) in a face-to-face meeting where all three raters gathered to familiarize themselves with the designed role play tasks. During this meeting the test takers’ audio-recorded samples were used so as to elicit the levels of the rating rubrics, the rating categories, the band levels, each descriptor and several features as indicators of each level and in each category. They were then given the practice of rating some audio-recorded samples, followed by a discussion to share their practice rating results and to have a shared understanding about which band level each test taker’s performance deserved and the reasons for the score assignments. The raters were also informed of the rating procedure described in section 3.5.3 later.

Following Roever (2005) and Timpe (2013), the researcher involved himself as a rater. Although both of the hired raters agreed to engage in the rating of all collected speech samples, their availability was beyond the researcher’s control. Therefore, the researcher’s rating could be used as the anchor data in case any parts of the rating of the remaining two raters were missing. In the end, all of the raters scored all test takers’ performances.

3.5 Procedures

3.5.1 Test Administration Procedure

The test was administered to each test taker individually through a medium of face-to-face meeting. Each test taker completed the test consisting of six role play tasks with one of the trained interlocutor participants. Assignment of the interlocutors to each
test taker was decided based on the availability of the test takers and of the interlocutors. A venue suitable for private use and for audio-recoding at the University of Melbourne was arranged for the test administration. Similar test conditions (for private use and for audio-recoding) were maintained for these test sessions.

For the task administration, either set 1 or set 2 (see Table 3.5 above) was assigned to each test taker. Assignment of test takers to either task set was decided by the researcher each time before the test administration to make sure that the test takers’ demographic information was equal, considering their background (proficiency levels as indicated by the test takers’ self-reported official test scores and university or pre-entry students), in order to have the average abilities of both test taker groups assumed to be the same. The whole test was administered by one of the interlocutors, who first provided and then read the instructions (APPENDIX A) to the test taker as they were. The interlocutor confirmed with the test taker that he or she understood what to do before starting the first task. Regardless of the task set, the test takers were assigned each task in the order indicated in the first column in Table 3.5 above. For the dialogue tasks, the interlocutor played all of the three kinds of roles (professor, administrator, and peer student). It might be more authentic that different roles in the role play tasks were played by different interlocutors, but it was logistically difficult to adjust multiple interlocutors’ schedules to meet each test taker’s availability.

In each task, the test takers were given the task prompt (APPENDIX B or C) and five minutes to prepare before starting the task. They were allowed to start the task whenever they judged they were ready, without having to use all of the given time. Unlike the pilot study, it was decided for this main test that the interlocutor could give an extra two minutes if the test taker found it necessary. The test takers were not informed, however, that they would have that extra two minutes until they had spent all
of the maximum five minutes. For task preparation, the test takers were allowed to make notes on the task prompts and, if necessary, to ask the interlocutor about the situation described in the task prompts and of the meaning of certain vocabulary so that they could understand the situation clearly. However, clues as to what and how to say it, in order to complete the task, were not provided nor suggested to the test takers as these were the very things that the study aimed to investigate. The test takers were also allowed to look at the notes in each task while they were performing in the role play tasks. There was no time limit for each task. To make the tasks more authentic, the test takers could conclude the conversation when they thought it was finished. Completing all of the tasks took approximately 40 to 50 minutes. All of the test takers’ performances were audio-recorded.

In total, four task performances produced by four test takers were re-taken as they misunderstood the basics of the tasks (e.g., who to talk to, on what modality to perform). For the analysis of discourse data and the rating, the re-taken performances were used because their first performances did not appropriately reflect the target features for the study. Except for these four cases, no major issues were confirmed in the administration of the test.

3.5.2 Post-task Questionnaire Procedure

Immediately following the test tasks, each test taker met the researcher for the post-task questionnaire. The presence of the researcher for the face-to-face meetings were considered necessary in order to have the intentions of each question item clearly understood by the test takers. For the question items, the test takers were given a written form in addition to the researcher’s oral explanation of what each item meant. To elicit the test takers’ responses, the test takers were allowed to report their answers to the
researcher about the given questions either in written or verbal form, depending on their preferences. The meetings with the test takers were conducted in English as their first languages were diverse. In the cases in which the test taker’s first language was the same as the researcher, both the test taker and the research were allowed to communicate in their first language. As far as the researcher conducted all meetings with the test takers, they were able to express their perspectives to the questions in English as adequately as the researcher expected. They were in reality proficient enough to pursue university studies in English-medium contexts or near that level. In addition, the questions and the response categories were to a large extent structured and were designed so as not to pose too high a demand upon the test takers for them to understand the meaning of the questions and to express their perspectives on each question.

3.5.3 Rating

The test takers’ performances were scored according to the rating criteria and the descriptors developed for the current study (shown in APPENDIX F), to facilitate the raters’ rating work. The rating criteria were intended to analytically assess L2 oral pragmatics in extended discourse. Although its intended utilization was not specified, the rubrics were expected to provide stakeholders (e.g., English teachers) with descriptors characterizing test takers’ pragmatic performances, in addition test scores for their decision-making.

The first four features defined in the current rating criteria, *Social Actions to Achieve the Communicative Goal*, *Facility with Language*, *Language Use for Mitigation*, and *Language Use to Deliver Intended Meanings*, were the criteria shared for assessment of both dialogue and monologue performances. The last two criteria,

---

2 How useful the descriptors and test scores are for stakeholders will be investigated in future studies addressing the Utilization inference.
Engagement in Interaction and Turn Organization, were for dialogue performance.

Table 3.8 below summarizes six assessment features of pragmatics defined in the rating criteria for the current study and how these features integrated the constructs discussed in the literature. The table below provides a summary of the competency characterizing each criterion. These criteria operationally defined what it means to be pragmatically competent for the current study.

Table 3.8
Descriptions of Pragmatically Competent Test Takers

<table>
<thead>
<tr>
<th>Rating criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Actions to Achieve the Communicative Goal</td>
<td>Being able to take actions (Al-Gahtani &amp; Roever, 2012; Schegloff, 2007; Youn, 2013, 2015) adequately, appropriately and without redundancy to handle pragmatic demands for the communicative goal from opening through to closing in a given context (Roever, 2011). The actions taken are not simply diverse but explicitly tailored to the situation.</td>
</tr>
<tr>
<td>Facility with the Language</td>
<td>Being able to smoothly and clearly deliver contents controlling sound variation (Gumperz, 1982; Youn, 2013, 2015) and repair (Roever, 2011; Schegloff, 2007) when necessary.</td>
</tr>
<tr>
<td>Language Use to Deliver the Intended Meanings</td>
<td>Being able to control varied linguistic resources and employ linguistic resources to deliver the intended meanings. The speaker’s language use can minimize the addressee’s effort to understand what the speaker tries to say.</td>
</tr>
<tr>
<td>Language Use for Mitigation</td>
<td>Being able to control varied linguistic resources to mitigate imposition (Searle, 1969; Youn, 2013, 2015).</td>
</tr>
<tr>
<td>Engagement in Interaction (Dialogue tasks only)</td>
<td>Being able to engage in interaction naturally by understanding the interlocutor’s turn (Youn, 2013, 2015) and responding with varied patterns of feedback well-tailored for the ongoing context, without noticeable scaffolding from the interlocutor (Al-Gahtani &amp; Roever, 2012).</td>
</tr>
<tr>
<td>Turn Organization (Dialogue tasks only)</td>
<td>Being able to take turns smoothly (Youn, 2013, 2015) throughout the task by naturally taking and releasing conversation turns (Ten Have, 2007) so that the interlocutor understands when to take turns (Young, 2008).</td>
</tr>
</tbody>
</table>

Each feature above classified the test takers at four band levels with score 4.
(maximum) and score 1 (minimum). Score 4 (maximum) was awarded to performance with strong and consistent evidence of positive features summarized in Table 3.8 above throughout the task. The lower scores 3, 2 and 1 accounted for the performances with inconsistent evidence of positive features, partially-to-largely flawed by negative features. Thus, these lower scores represented performances with inconsistent or weak evidence of positive features or those with consistent or stronger evidence of negative features as described in the rating criteria. The current rubrics were designed to award score 3 to performances with minor flaws, which, in other words, were not seen in performances equivalent to score 4. The remaining two lower scores were applied to test taker performances where negative features were more noticeable and outweigh positive features. The full descriptors that the raters utilized are provided in APPENDIX F.

The analytic rating rubrics were developed based on the insights provided by previous studies, which were refined by the empirical data obtained in the current study, including the test takers’ oral role play performances collected in the pilot and the main study. For the development of the rating criteria, it was considered necessary to integrate features and sub-features that discriminate the test takers and, to develop descriptors that describe features which discriminate the test takers, which can be operated by raters to facilitate their rating work and which can provide diagnostic information for test users. The development of the rating criteria was guided by the following approaches: (a) a review of the literature, (b) analyses of role-play data from the pilot and from the main studies, and (c) comments from two interlocutors.

The overall framework of the rating criteria was adapted and refined from Youn’s (2013, 2015) rating criteria, which included Content Delivery, Language Use, Sensitivity to Situation, Engaging with Interactions, and Turn Organization. These criteria captured L2 pragmatic and interactional abilities that L2 students utilized in
extended oral discourse situations at academic settings. For the procedure of developing the rating criteria for the current study, the researcher first reviewed Youn’s criteria and other studies on oral pragmatics and interaction. As done by Youn (2015), the researcher transcribed randomly selected test takers’ role play performances across different groups of test takers (for the grouping, see Table 3.7 above in this chapter), using the Conversation Analysis conventions (see APPENDIX G). The dialogue data were analyzed from a conversation-analytical point of view (Atkinson & Heritage, 1984; Schegloff, 2007) as modeled by empirical studies which identified features related to pragmatics (Al-Gahtani & Roever, 2012; Walters, 2007; Youn, 2013, 2015). The primary reason for employing Conversation Analysis was to reveal speakers’ interactional features, including pause, pitch, turn-taking, in addition to describing the speaker’s social actions and linguistic devices. Conversation Analysis allows for describing sequentially organized discourse performances, which were the target of assessment for the current study. Sequenced performance, in particular, occurs under dialogue conditions where the next or subsequent turns are led by the previous turns in a conversation. This sequential nature of dialogue performance allows us to investigate whether and how the hearer’s next utterance reflects his or her understandings of the previous turns and the discourse context. Thus, analyzing the test takers’ oral discourse data from a conversation analytic point of view describes the language use in the extended discourse as well as how the test takers’ performance affects the interlocutor’s ways of communication, which clearly account for what Crystal (1997) defined as pragmatics.

The monologue role play performances were also examined qualitatively. The monologue data did not contain the features that could be exclusively seen in dialogue performance (e.g., turn taking). As the monologue discourse performances could not be
described in the same way as for the dialogue performance, the features were described partially by applying what was described in the dialogue: pitch, speed, pause and initiation of a repair as well as how test takers recognize the given context manage the pragmatic demands for the given communicative goal. As demonstrated in Youn (2013, 2015), the development of the rating criteria was based on the researcher’s qualitative analyses of the test takers’ task performances.

Subsequently, the researcher listened to all of the audio-recoded task samples multiple times, from the perspectives of the criteria discussed above (e.g., interactional features, language use, and sound variations). The purpose of examining the test takers’ performance through the medium of audio materials was to identify whether the CA-analyzed features of the selected task performances were seen in the other test takers’ performances and had discriminated the test takers, as well as to confirm the degree of demands upon the raters in scoring the audio-recorded task samples. It was deemed important that the descriptors which the raters applied should not be too detailed (thus too demanding in scoring), while giving them guidance to facilitate the rating. Also, the descriptors were expected to provide diagnostic information to test users.

Additionally, the drafted rating descriptors were reviewed by one of the raters who was the ESL-certified teacher, serving as an interlocutor for pre-pilot study 1, the pilot study and the main data collection at the same time. It was confirmed that the rating was feasible using the drafted rating criteria and the descriptors. At the same time, the descriptors of two criteria (Engagement in Interaction and Turn Organization, see Table 3.8 above for the criteria of the current study) were partly refined to reflect test takers’ interactional behaviours more accurately, following the suggestion of the ESL teacher serving as an interlocutor for a number of test takers. The researcher re-examined the conversation analyses that he had conducted and used another
interlocutor’s voluntary comments in order to confirm the commented features. It was then deemed reasonable to conduct the partial revision of the descriptors.

For administration of the rating, a fully-crossed rating design, as illustrated in Knoch (2009), was employed to make the analysis more stable and thus to allow for a better interpretation of results (Myford & Wolfe, 2003). This rating design required all three raters to rate all 402 audio-recorded samples (67 test taker task performances times six tasks). To be more exact, each rater made 2010 scoring judgments in total, which were composed of six rating criteria multiplied by three dialogue tasks, multiplied by 67 test takers, plus four rating criteria, multiplied by three monologue tasks, multiplied by 67 test takers. Considering practicality and budget constraints (Bachman & Palmer, 1996) for rating administration (and accordingly, assessment as a whole), the raters were not asked to re-play the audio-files.

The rating was conducted in the following order in Table 3.9. The bracketed numbers, (1) to (12), in the table represent the order of rating.

<table>
<thead>
<tr>
<th>Addressee</th>
<th>33 test takers with set 1</th>
<th>34 test takers with set 2</th>
<th>Topic title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Professor</td>
<td>Dialogue (1)</td>
<td>Monologue (7)</td>
<td>Get a signature to change classes</td>
</tr>
<tr>
<td>2. Administrator</td>
<td>Monologue (8)</td>
<td>Dialogue (2)</td>
<td>Change the registration of the class</td>
</tr>
<tr>
<td>3. Peer student</td>
<td>Dialogue (3)</td>
<td>Monologue (9)</td>
<td>Make pairs for a presentation</td>
</tr>
<tr>
<td>4. Peer student</td>
<td>Monologue (10)</td>
<td>Dialogue (4)</td>
<td>Reorganize a presentation</td>
</tr>
<tr>
<td>5. Administrator</td>
<td>Dialogue (5)</td>
<td>Monologue (11)</td>
<td>Solve a projector problem</td>
</tr>
<tr>
<td>6. Professor</td>
<td>Monologue (12)</td>
<td>Dialogue (6)</td>
<td>Avoid losing marks for an assignment</td>
</tr>
</tbody>
</table>

The rating started with (1) All 34 samples of 1. Professor topic (Dialogue),
followed by (2) All 33 samples of 2. Administrator topic (Dialogue) and concluded with (12) All 34 samples for 6. Professor topic (Monologue), instead of rating one test taker’s six different task samples and moving to the next test taker’s six different task samples. The raters rated on an individual basis. They were given the audio files and an Excel rating worksheet to input the score of each criterion for each test taker.

The raters rated individually and returned their work every time they completed one rating batch from (1) to (12) respectively. After the raters completed rating batch (1), the researcher had discussions with the raters individually to confirm that the rating criteria were appropriately operated and to remind them of the intention of each band level of each criterion. In the discussions, the researcher asked the raters to rate some audio samples that the researcher chose and to provide reasons for the rating for the purposes of maintaining the shared understanding among the raters. The same procedure was taken after the rater returned rating batches (2) and (6) to the researcher and when the availability of the raters allowed for it.

3.6 Data Analysis

The current study collected multiple sources of data obtained from a performance-based speaking assessment, which includes the test takers’ task performances on the oral role plays, the time that the test takers spent for the task preparation, the test takers’ perspectives, and the test scores provided by the raters. This section provides an overview of the types of data and how these data were analyzed. The details of the analyses and their procedures are described in each section in Chapter 4 addressing each of the main research questions.
3.6.1 Qualitative Analyses of the Oral Discourse Data

The test takers in the main study provided their performances on the oral role play tasks, including both the dialogue and monologue data. As described in section 3.5.3 earlier, the speaking data were analyzed qualitatively to identify features of pragmatics frequently seen in the test takers’ performances and how the features discriminated the test takers.

3.6.2 Test Takers’ Questionnaire Responses and the Analyses

Descriptive statistics were generated based on the test takers’ questionnaire responses by the groups (see Table 3.7 above for the test taker groups). As described in section 3.3.6, the data included (a) the test takers’ perception of the two types of construct-irrelevance, (b) their perceived similarity between their task performances and those in reality, (c) their degree of awareness of the pragmatic features, and (d) their self-assessment according to the pragmatic features. The statistical results were summarized by the groups using SPSS version 23. However, no presumptions were made about the group differences for (a), (b), (c), as there were few theoretical assumptions and little empirical evidence on how the test takers in each group respond to these questions. (4) Self-assessed performances were analyzed by comparing the mean scores of each group. The self-assessment results were also integrated in the correlational analyses with the test results produced by the Rasch analysis, as described in the following section.

3.6.3 Test Scores and the Analyses

The test takers’ task performances were converted into test scores by the three raters including the researcher, according to the rating criteria. A fully-crossed rating
design was made possible as all three raters scored all 402 audio-recorded samples (67 test taker task performances times six tasks) according to the criteria (four criteria for both dialogue and monologue samples and two more criteria for dialogue samples). The test scores allowed for investigation of characteristics of performance assessment (McNamara, 1996) including test takers, raters, test tasks and rating criteria.

Using the test scores, firstly, quantitative analyses were carried out to explore test reliability and inter-rater reliability on Classic Test Theory using SPSS version 23. Secondly, four facets (test taker, rater, test task, and rating criterion) were specified in the multi-faceted Rasch measurement analyses employing the computer program, FACETS (Linacre, 2015). The FACETS program has been widely used for research on second language performance assessment (McNamara, 1996). The Rasch-informed results were summarized for descriptive statistics and displayed graphically where applicable.

The Rasch analysis was conducted based on the whole data set as described above. Separate Rasch analyses were also performed using parts of the data depending upon the research questions and sub-questions to address. For all analyses, the partial credit model (Masters, 1982) was adapted as the particular band score because one rating category (e.g., score 3 of Facility with the Language) did not always have the same level of difficulty as the same score in another category (e.g., score 3 of Facility with the Language) and because each rater may interpret the rating scale in a different manner. The outcome produced from this procedure allowed the researcher to examine how the rating scales were operated by the raters.

To perform the analyses above, the test takers assigned to task set 1 and task set 2, were group-anchored (M. Linacre, personal communication September 10, 2015) to achieve connectivity between the test taker groups on two different task sets, except on
a case where an analysis using the data on task set 1 and another analysis using the data on task set 2 were conducted separately to compare the results.

The test takers’ abilities, estimated by the Rasch analyses, allowed for further statistical analyses including group comparisons and correlational analyses. Group comparisons were conducted to investigate whether and the extent to which the groups differed as assumed by the literature of L2 pragmatics (see Table 3.7 for the test takers’ groups and the hypotheses about the group differences). The correlational analyses allowed the researcher to examine the relationship between the estimated abilities and the self-assessment results as well as the relationship between the test takers’ abilities on the dialogue tasks and those on the monologue tasks.

3.6.4 Task Preparation Time and the Analysis

The test takers’ engagement in the role play tasks also provided their task preparation time measured in seconds. The test takers’ task preparation time was recorded by the interlocutors as instructed by the researcher or measured by the researcher himself by replaying the recoded back-up audio files (a recording of each test taker’s engagement in the data collection from the general instruction to the end of the interview, without being stopped). The preparation time was measured and recorded in seconds. The task preparation time was summarized for descriptive statistics and analyzed for group comparison as per average time (in seconds) of each test taker group (see Table 3.7 above for the test taker groups).

Although not a central concern to the study, the time each that test taker participant spent for preparing for the assigned tasks was considered to be a piece of supplementary data. The preparation time was principally given to the test takers to situate themselves in the simulated university settings in each task, not to measure their
language abilities. The test takers were allowed to use the given time to prepare for the assigned tasks and were told that how much of the allocated time they used was up to them. It was difficult to predict the usage patterns of the preparation time across all of the test takers and how the results related to their language abilities or to the task design. However, it was deemed reasonable to assume that the Pre-entry students need more time to situate themselves in the simulated university settings where they are required to handle pragmatic demands and to prepare for their language output.

3.7 Chapter Summary

This chapter has described the methodology adapted in the current study. This study encompassed the comprehensive validation and the issue of instrument practicality, which involved a range of analytical procedures and multiple types of data. The role play tasks were drafted and revised through the multiple phases of the task development activities in order to collect oral extended discourse data from L2 students. They also provided their perceptions as well as their task performances. The test takers’ task performances were scored by the raters according to the rating criteria developed for the study. A range of analyses were used to address the research questions for the study.
Chapter 4 Results and Discussion: Validation and Addressing of Practicality

4.1 Introduction

This chapter presents and discusses multiple sources of evidence (empirical as well as methodological) for the purposes of validation addressing the inferences of Domain Description, Evaluation, Generalization, Explanation, Extrapolation and Utilization in the argument-based frameworks (Chapelle, 2008; Kane, 2006, 2013; Knoch & Elder, 2013). The chapter also discusses how the discursively-orientated assessment of L2 pragmatics can be implemented practically. In the subsequent sections which are addressing each validation inference and practicality issue, I will develop the discussions based on the overall results and more specific backings for each section.

Each of the sections 4.2 to 4.7 addresses each validity inference by discussing backings to support the inferences. In this chapter, the discussion of the Evaluation inference (section 4.3) will be lengthier than for the other sections since a dedicated qualitative study was done to ensure the viability of the rating scale, by providing descriptions of sequentially-organized discourse performances across the assessment band levels and across the rating criteria.

4.2 Domain Description

The Domain Description inference looks at the target domain relevance of the abilities. This section addresses the following research question:

To what extent does the assessment content reflect L2 students’ oral pragmatic abilities for use in university settings as the target domain?
The backings to address the Domain Description inference are listed in Table 4.1 below. The first was sought from the reported practices in the methodology section, including how the target language activity domain was specified and how the test tasks were refined. The findings of the test takers’ perception about their awareness of pragmatic demands were also integrated to further support the Domain Description inference.

Table 4.1
Backings and Sub-questions to Address the Domain Description Inference

<table>
<thead>
<tr>
<th>Backing</th>
<th>Source of supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ Research question</td>
<td></td>
</tr>
<tr>
<td>1. Language activity domain, sub-domain specified and perspectives of domain experts integrated in the task design</td>
<td>The practice of the task development in the current study</td>
</tr>
<tr>
<td>2. To what extent the test takers aware of pragmatic demands in language activities at university?</td>
<td>Descriptive statistics of the test takers’ questionnaire items on their awareness of pragmatic demands</td>
</tr>
</tbody>
</table>

4.2.1 Backing from the Methodological Design

In the current study, both the test tasks and the test takers were the most relevant to the higher education domain. The first source of evidence is concerned with the target language activity domain that the current study defined. Specifying the target domain, to which the observed performance is extrapolated (Kane, 2006, 2013), is crucial as the conclusions can be made based on the target domain about how the test scores are indicative of the traits relevant to the specified domain. A university domain, although narrowed down from every possible language use setting in real life, is still broad. The current study further specified the language activities domain, where pragmatic abilities are utilized at university to which test takers’ observed task performances are extrapolated. As there was no explicit guidance in the literature about the extent to
which the target domain needs to be specified, specification of the target domain would be left to the test developers’ purposes. Broadly, the current research targeted higher education as the domain. Thus, the task instruments designed in the current research situated target test takers in the university contexts. The current research further specified the domain in terms of when, where, to whom and for what purpose university students engage in the simulated language activities, but it does not do this in too narrow a focus.

Specifying the domain is an initial but crucial phase for argument-based approaches to validation because interpretation of the test takers’ performances underlying the test score is made based on the target domain. In the process of developing the tasks, considerations were given to simulate authentic practices of the university domain (e.g., the authentic addressees such as professors and administrators in addition to peer students, as well as situations that university students may possibly encounter). Also, the arrangement of the role play tasks assigned to each test taker followed the chronological flow from the beginning to the end of a semester which presumably is the period when university students are at their most active on campus. The role play tasks did not simulate language activities that exclusively occur in particular disciplines (e.g., activities such as those seen in a law school or in particular backgrounds of students). The tasks simulated activities occurring to students in general in the target domain. Also, the tasks were designed to include a range of possible university activities across a semester.

The second source of the evidence was obtained by the process of developing test tasks, which reflected the domain experts’ review (Chapelle, 2008). The perspectives of domain experts, a group of academic staff and students in a university domain, were also integrated into the work of drafting and refining the tasks simulating
activities, where students need to address others and deal with pragmatic demands. Experiences from these domain experts as well as the current researcher himself being active in the target domain were utilized to draft and revise the task situations. These perspectives served to confirm and revise the situations where pragmatic demands authentically occur in the target domain. The experiences of the researcher him/herself in the target domain provide useful source of information to draft role play situations as Timpe (2013) practiced. Multiple members of domain experts combined with the researcher’s view contributed to maintaining the quality of the test tasks.

4.2.2 To What Extent Are the Test Takers Aware of Pragmatic Demands in Language Activities at University?

The Domain Description inference can be further enhanced by the test takers’ perspectives about the extent to which the features relevant to pragmatics actually matter to them in language activities at English-medium university. Since the Pre-entry students had not necessarily experienced any activities in English-medium universities, they were asked to provide their view about the extent to which they would be conscious about these features in the university domain that they are going to enter. The pre-entry students as the prospective students were expected to be highly conscious of the target domain as they were studying at a university preparation course thus their perspectives were also deemed meaningful.

The test takers were asked to express their perceived degree of importance of the six features of pragmatics in a university domain on four-point Likert scales. The question items were the same as the six self-assessment items, as described in APPENDIX E. The Likert scale data were computed to provide descriptive statistics of each question indicating pragmatic features and of how the test takers’ responses were
distributed across the response types.

The central question to this analysis is to what extent the features were valued by the test takers as a whole. Its primary purpose was not to compare and distinguish each feature from the others but to confirm that there are no individual feature and the features as a whole which are considerably undervalued or are not recognized at all. The four scales indicating the test takers’ perspectives were “4: Very much”, “3: To some extent”, “2: Not so much”, and “1: Not at all.”

“4: Very much,” indicates the greatest degree of the test taker’s consciousness whereas “1: Not at all,” indicates the least. Any answers unclear or not falling within the four categories were classified as “Other.” Table 4.2 below reports the distribution of the responses in each question item and the number of valid responses (responses 4 to 1) to calculate the mean scores in Table 4.3 later.

Table 4.2
Response Distribution of Test Takers’ Awareness of Pragmatic Demands

<table>
<thead>
<tr>
<th>Feature</th>
<th>Response category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>(a) Language use depending on social status</td>
<td>43</td>
</tr>
<tr>
<td>(b) Language use depending on familiarity (social distance)</td>
<td>24</td>
</tr>
<tr>
<td>(c) Language use to deliver intended meaning</td>
<td>33</td>
</tr>
<tr>
<td>(d) Language use for mitigation</td>
<td>36</td>
</tr>
<tr>
<td>(d) Repair to stress and to mitigate the information</td>
<td>23</td>
</tr>
<tr>
<td>(f) Repair for language mistakes</td>
<td>24</td>
</tr>
</tbody>
</table>

Note. (a) to (f) correspond to the question items (a) and (f) in APPENDIX E.

One Pre-entry student chose “Other” for (c) Language use to deliver intended meaning and one Other university student also chose “Other” for (e) Repair to mitigate and to stress the information. The results suggest that overall the test takers were aware of the use of these features for communication in their current or future activity domain.
as their responses were largely distributed in the response categories 4 and 3 across the six features. The small number of test takers chose “2: Not so much” for each feature.

Table 4.3 below reports the mean scores of the test takers’ awareness about the features as per each test taker group. The means scores were above 3.00 for all three groups on all six features. The group of Advanced university students did not necessarily mark higher mean scores than the other two groups, which depended on the feature. The central concern here was to confirm whether or not there were any features particularly considered unimportant, and whether there were any test taker groups that expressed this unimportance.

Table 4.3
Mean Scores of Test Takers’ Awareness of Pragmatic Demands

<table>
<thead>
<tr>
<th>Feature</th>
<th>Advanced university student</th>
<th>Other university student</th>
<th>Pre-entry student</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Language use depending on social status</td>
<td>3.33 (0.87)</td>
<td>3.69 (0.53)</td>
<td>3.52 (0.59)</td>
<td>3.58 (0.61)</td>
</tr>
<tr>
<td>(b) Language use depending on familiarity (social distance)</td>
<td>3.11 (0.78)</td>
<td>3.17 (0.79)</td>
<td>3.09 (0.79)</td>
<td>3.13 (0.77)</td>
</tr>
<tr>
<td>(c) Language use to deliver intended meaning</td>
<td>3.44 (0.73)</td>
<td>3.34 (0.77)</td>
<td>3.36 (0.73)</td>
<td>3.36 (0.74)</td>
</tr>
<tr>
<td>(d) Language use for mitigation</td>
<td>3.78 (0.44)</td>
<td>3.51 (0.61)</td>
<td>3.26 (0.69)</td>
<td>3.46 (0.64)</td>
</tr>
<tr>
<td>(e) Repair to stress and to mitigate the information</td>
<td>3.44 (0.53)</td>
<td>3.06 (0.85)</td>
<td>3.13 (0.76)</td>
<td>3.14 (0.78)</td>
</tr>
<tr>
<td>(f) Repair for language mistakes</td>
<td>3.33 (0.87)</td>
<td>3.20 (0.68)</td>
<td>3.17 (0.65)</td>
<td>3.21 (0.69)</td>
</tr>
</tbody>
</table>

The results indicate the test takers’ sufficient awareness of pragmatic demands required to handle language activities at university. Therefore, the target of the assessment showed the close relevance to the target domain, as the test takers and the tasks reflected the domain.
4.3 Evaluation

The Evaluation inference connects test takers’ observed test performance to test scores (Kane, 2006) and rests on an understanding of the procedure for assigning test scores. Unlike an assessment providing a small number of test takers with descriptive feedback, the current study targeted a number of test takers and aimed to convert the test takers’ observed task performances into scores. Test scores allow us to indicate test takers’ ability levels, to conduct quantitative investigations of characteristics of the test, and to facilitate educational decisions by setting up a cut-off score point. However, the process of transforming the observed task performances into scores are subject to losing information of observed task performance. This section thus addresses the following question:

To what extent is the procedure for assigning test scores appropriate?

The procedure for assigning test scores involved the development of rating criteria and the practice of rating. It also includes the raters’ actual rating performances, which were quantitatively investigated. Specifically, the investigation for the Evaluation inference was guided by the following backings and the questions outlined in Table 4.4 below.
Table 4.4
Backings and Sub-questions to Address the Evaluation Inference

<table>
<thead>
<tr>
<th>Backings</th>
<th>Source of supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>Sub-questions</td>
<td></td>
</tr>
<tr>
<td>1. To facilitate rating, the raters were trained and, the raters were provided with the descriptors that capture features differentiating test takers</td>
<td>The practice of the rating in the current study</td>
</tr>
<tr>
<td>2. Do the rating criteria describe clearly distinguishable levels of discourse performance?</td>
<td>Qualitative analyses of the test takers’ task performances</td>
</tr>
<tr>
<td>3. Do the rating criteria provide a stable assessment of L2 pragmatics defined in the study?</td>
<td>Measurement reports produced by the multi-faceted Rasch analysis</td>
</tr>
<tr>
<td>4. To what extent were the raters’ rating performances reliable and consistent?</td>
<td>Raters’ rating performances computed based on the test takers’ test scores including: Inter-rater reliability (correlation between raters), Severity, fit statistics (by the Rasch analysis)</td>
</tr>
<tr>
<td>5. How appropriately were the rating rubrics operated by the raters?</td>
<td>Rating Scale functionality (probability curves)</td>
</tr>
</tbody>
</table>

The backings for the Evaluation inference could be sought from the methodology (the first backing) and the empirical evidence (the remaining backings).

The first backing was supported from the practice of the rating activities. The qualitative discourse analyses were integrated into the second backing. The third backing was sought from the findings of the FACETS output, which evaluated the rating criteria. The fourth backings were sought from the empirical evidence from the raters’ performances by investigating inter-rater reliability indices. The fifth backing was explored by the FACETS output which provided the probability curves indicating how rating scales were operated by the raters to assign the test scores to the test takers.

4.3.1 Backing from the Methodological Design

Evidence was first sought from the practice of the rating activities. For the raters
to facilitate the rating as intended by the researcher, the raters were provided with the analytical rating criteria with the descriptors, which captured the features of each level on each assessment criterion. These descriptors were to inform the trained raters of accounts as to why the test taker performance deserved a particular band level and how they were differentiated from those performances which deserved of the other levels. Providing these descriptors could help the raters “so that the raters will have a clear idea of what each scale category mean and will be better able to distinguish between the different levels of a trait” (Myford & Wolfe, 2003, p. 394). Prior to the rating, the raters and the researcher had the training session to come to an agreed view of the rating. In the process of the rating, the trained raters were reminded by the researcher (as the developer of the rating instrument and rating procedure) of the meaning of the rating descriptors and how to apply them to test takers’ performances to score.

A possible rating practice would be to entirely rely on trained raters’ intuition as to which pragmatic features to pay attention to, as to how they are assessed (number of level bands), and as to the judgment of test performance. This case however might result in unpredictable outcomes. An alternative could be that developing the rating criteria without any descriptors, relying on the raters’ intuitive judgments on the degree of appropriateness, as traditionally practiced in pragmatic assessments (Hudson et al., 1995). A strength of this rating design would be to reduce the work of developing the rubrics and to reduce financial cost for developing criteria since rating is left to raters’ intuition. It may also lead to a consequence that the insufficient guidance on how to differentiate test takers’ performance gives raters a greater difficulty in rating and that the freedom of the rating leads to a greater variation of rating performance. Given that the outcomes of raters’ performance rising from these issues remained unpredictable, it was deemed necessary to develop descriptors which define the reasons why the
performance is assessed positively and negatively and arrives at a certain band level for the rating.

4.3.2 Do the Rating Criteria Describe Clearly Distinguishable Levels of Discourse Performance?

Following the insights of the related literature, the rating criteria in the current study were developed to capture the pragmatic performances of the test takers in the current study. In this section, I will present how the findings of the qualitative discourse analyses as well as the insights of the literature informed the six assessment criteria and their descriptors characterizing each level. The six criteria were: Social Actions to Achieve the Communicative Goal, Facility with the Language, Language Use to Deliver the Intended Meanings, Language Use for Mitigation, Engagement in Interaction and Turn Organization. Each criterion has four band levels with the descriptors characterizing the levels (as shown in APPENDIX F). As such, this section is arranged by criterion and by band level. In this section, I will start out with the first criterion, Social Actions to Achieve the Communicative Goal and qualitatively describe how the four band levels differentiate the test takers’ performances. The transcription conventions used in the study are provided in APPENDIX G.

4.3.2.1 Social Actions to Achieve the Communicative Goal

In this section, I will show examples of the features characterizing the test takers’ Social Actions to Achieve the Communicative Goal at different assessment levels. This criterion looks at adequacy and quality of actions that test takers took to constitute their extended discourse performances from the opening through to the closing (Roever, 2011) to achieve the given communicative goals.
This criterion shares aspects with Youn’s (2013, 2015) empirically-developed criteria, *Sensitivity to Situation*, which measures whether test takers take expected actions for the given situations (e.g., providing explanations and so forth depending on the task situation), which would reflect the speakers’ awareness and sensitivity to a specific situation. Sensitivity to or understanding of a given context may account for receptive abilities and thus could not be directly evaluated. However, taking such actions requires speakers to be sensitive to the given context first.

*Social Actions to Achieve the Communicative Goal* for the current study assesses whether adequate actions are taken and whether the quality of the actions are adequate to achieve the given communicative context. The open role play tasks in the current study were designed to situate test takers in the contexts where they need to handle pragmatic demands to achieve the given communicative goals, as well as to elicit their extended discourse performances from the opening to the closing. The tasks were designed not to overly restrict test takers’ performances, which allows pragmatically competent test takers to take more diverse actions tailored for the given and the ongoing contexts.

The majority of the test takers in the current study were able to take at least minimum actions including pre-expansions (Al-Gahtani & Roever, 2012; Schegloff, 2007) before the main topic and some closing actions (Roever, 2011). The test takers’ abilities to produce social actions should not be surprising, given the test takers’ proficiency levels equivalent to university admission levels in Australia or university pre-entry levels, except the two pilot test takers with proficiency levels lower than the pre-entry levels. What was discriminating the test takers was whether the action was tailored for the context and adequacy of actions and not simply whether or not expected actions were taken.
It was found in the discourse data that the quality of the action was sometimes undermined by features such as being too simple, being redundant for the context, and being obvious copies of the task prompt. Examples were provided below to illustrate the test takers’ performance features at different score levels.

**Score 4 Level of Social Action to Achieve the Communicative Goal**

In all tasks, the test takers were expected to recognize the context and to take actions appropriate to achieve the communicative goals. The tasks were designed to give the test takers the goal of eliciting support from the intended addressee, in order for them (as a test taker) to avoid being in a problematic situation and to promote success for their academic work, although respective task situations were different. The performances deserving score 4 (the highest score of the four band levels) were characterized by strong evidence of taking solid actions to introduce the topic, to make the speaker himself or herself understood, to seek the intended addressee’s support and to close the extended discourse naturally. Performances equivalent to score 4 thus provided strong and consistent evidence of social actions throughout the task. The descriptor for score 4 states as follows:

The test taker’s performance shows understanding of the given or ongoing situation with clear, necessary and appropriate actions and information to prevent the problem and promote success in the situation (e.g., negotiate an alternative way with the administrator, proposing a meeting with the classmate, elaborate what the test taker says or tries to say) while the test taker handles the interlocutor’s unwillingness and resistance well.
Excerpt 1 below is an example of the opening part explicitly tailored to the situation. This task simulates a situation where a student (test taker) is addressing a professor at the beginning of a semester. In this task situation, the student needed to change the class of the professor (Professor Smith in Excerpt 1) to another class due to a timetable clash. This has the student ask Professor Smith to give the professor’s signature on a class registration form to be submitted to an administration office by the following day. The student has already made that request to the professor by email but have not yet received a reply. Consequently, the student has decided to visit the professor’s office. The student has run into the professor on campus on the way to the office, which initiates the opening of the conversation from line 1 below.

At the opening of an extended discourse performance, Al-Gahtani and Roever (2012, 2013) empirically argues that less competent speakers are likely to rush to the main topic and to the main act (e.g., request) without any introduction whereas more competent speakers are able to lay pre-expansion (Schegloff, 2007) to project an upcoming main act. As predicted by Al-Gahtani and Roever (2012, 2013), the test taker in Excerpt 1 does not directly rush to the main topic but lays the introduction (lines 3-5) after addressing the interlocutor (line 1). Subsequently, the test taker provides an explanation of why she is addressing the professor (lines 14, 16-18, 20, 21, 23 and 24) before delivering the main act (lines 25, 28 and 29) with an additional explanation about the student’s situation (line 29).

**Excerpt 1 (Task ID: D1, Professor topic, Dialogue)**  
S=Student (Test taker), P=Professor (Interlocutor)

1  S:  excuse me professor Smith? [(I) was won
2  P:  oh hi
3  S:  hi I was wondering whether you received my email (0.5)
4  P:  >that I sent you< regarding the:: ah:: timetable clash
5  that I:
6  (0.3)
7  P:  ah
To examine these actions more qualitatively, it must be noted that the test taker uses her own words to explain the situation rather than copying the wording as it was written in the task prompts. The task prompts provided the situations where a university student (test taker) is in a problematic situation. The current test administration procedure allowed the test takers to look at the task prompts before and while they engaged in the conversation. In other words, the test takers were allowed to use the information, sentences and words written in the task prompts as they wished. Even though the test takers were allowed to do so, competent test takers at score 4 level were much more likely to use their own words (as in lines 16-18, 20, 21 and 23-24) combining extra words such as, when (0.5) umm when I looked at my schedule the other day (lines 16-17), to elaborate the explanation. The test taker’s introduction of the topic, I was wondering whether you received my email (0.5) >that I sent you< regarding the:: ah::: timetable clash that I: I ha:::ve (lines 3-5 and 8) can serve to remind the professor of the problem that the student encountered. The test taker organizes the structure of the extended discourse
smoothly and naturally so as to make herself understood by her interlocutor.

The positive features of competent test takers are also seen in the middle and in the closing part of the performance as illustrated in Excerpt 2 below. This dialogue situation simulates a situation at the beginning of a semester and requires test takers (as a university student) to address an imaginary classmate (interlocutor), in order to propose making a pair for a class presentation. In this task, test takers are required to convince the classmate (interlocutor) to work together for a class presentation on a topic in which both the student and the classmate have a mutual interest. Excerpt 2 below demonstrates the test taker’s solid social actions after the test taker has reached the mutual agreement with the classmate (interlocutor) on working together for the class presentation and the clear action to close the performance.

**Excerpt 2 (Task ID: D3, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

1. S: okay am: (0.6) all right >let’s work< together
2. P: >okay< cool. so: yeah so we’ll start working together:
3. and you know (. ) you know with- (. ) I will get to know
4. each other >more (for)< [working =
5. P: [sure
6. S: and to get how we work and stuff >and if there is< any
7. problem and >we just have to be honest< and like >you
8. know< say hey >you have to work more< or your know stop
9. pressing on me stop like that >you know< okay
10. [so ((laughs)) —
11. P: [okay (((laughs)))
12. S: [you know okay so ((laughs)) we’ll [work it out=
13. P: [I understand
14. S: =for you ((laughs)) (. ) [okay? cool†
15. P: [great all right
16. S: so yeah‡ I’ll (0.5) so: I’ll I think I *ha:ve you:r
17. email* or I’ll: have a che:ck
18. P: yes
19. S: and I’ll drop you some line- a line about [this stuff
20. P: [mmhm
21. S: an[d arrange a meeting right?
22. P: [great=
23. S: =co[ol† yeah:
24. P: [that sounds good
25. S: sorry I have to run for another class
26. P: go for [it
27. S: [yeah Cheers‡ >so I’ll< see you so[on† thanks=
28. P: [okay see you

120
Soon=
S: yeah it was good running into you ((laughs))=

P: =Thank you for >approaching< me=

S: =oh cheers! see you soon >bye [bye<=

P: [see you soon (.).] bye

After the test taker (as the student) and the imaginary classmate (as the interlocutor) have agreed to work together for the class project as reflected in the interlocutor’s agreement in line 1, okay am: (0.6) all right >let’s work< together, the test taker, instead of closing the conversation, proposes to start working on the collaborative project.>okay< _cool._ so: yeah so we’ll start working together: , (line 2). Furthermore, in lines 6-9, 12 and 14, the test taker takes an action with the purpose of encouraging themselves with the words, >and if there is< any problem and >we just have to be honest< and like >you know< say _hey_ >you have to work _more_ or your know stop _pressing_ on me _stop_ like that >you know< okay, you know okay so ((laughs)) we’ll work it out for you ((laughs)). This is followed by the classmate being informed that the student (the test taker) is able to send an email to the classmate (the interlocutor). The classmate agrees and the student (the test taker) initiates a closing of the conversation (line 25) explaining that she has to go, an indication that the student (the test taker) initiates the closing. The boundary of the closing (sorry I have to run for another class in line 25) was clear and indicating that the test taker intended to close the conversation.

The diverse actions taken by test takers at the highest band levels were also reflected in their monologue performances. Excerpt 3 below is a part of a performance of leaving a voicemail message. In this situation, a student (test taker) is leaving a voicemail message to a classmate working on a collaborative project together for a class presentation in the middle of a semester. Although not stated in the task prompt, the
intended imaginary addressee’s relative power over the student and social distance between the two speakers are different from the simulated situation in Excerpt 1 in which a professor, not necessarily familiar to the student, was addressed in a situation which had taken place at the beginning of the semester.

In this situation, the project that the student (test taker) and the classmate are working on a collaborative project. The project is almost at a stage of completion but the student (test taker) has recognized a shortcoming, one which may affect its grade and thus their grade on the subject overall, an important consideration. The student found a way to overcome the shortcoming, but this work would entail major modification and an extra burden on both of them.

Excerpt 3 (Task ID: M4, Peer student topic, Monologue)

1 hey Sharon >it’s< me ((first name)) I have good news and I
2 have bad news(.) >remember< how we were talking about um
3 (0.6) the potential issue of only having ah two male (0.4) ah
4 respondents >to our< questionnaires? well I found ah: six
5 more:: who are willing to:: >participate in< our research
6 (1.6) however we also um:: (.): I’m also< very well aware of
7 the fact that umm (.): >we have to present< in four days (0.3)
8 a:m I would be able to meet them this afternoon >and I’m also
9 happy to< do the: (.) analysis
10 ((continues))

11 yeah I (.) I just thought it would be fantastic to have those
12 additional speakers because the way it looks now (0.6)
13 ((continues))

14 and I think um >it’s in interest of< our research to: make it
15 as balanced as we >possibly can< (.): so >anyway< let’s talk
16 about this in person so am (.): if you could call me back
17 that would be great< see ya bye

The opening and the closing are explicitly tailored to this situation (voicemail to a close friend). After the greeting, the test taker utters a native-like phrase, I have good news and I have bad news(.) (lines 1-2), to mitigate the test taker’s upcoming proposal, which could possibly cause inconvenience to the classmate.
Following this expression, the test taker takes an action of reminding the classmate of the shared concern by saying, *remember* how we were talking about um (0.6) the potential issue of only having ah two male (0.4) ah respondents *to our* questionnaires? (lines 2-4). In addition, as *in*, and I think um *it’s in interest of* our research to: make it as balanced as we *possibly can* (. ) (lines 14-15), the test taker attempts to seek the understanding and approval of the classmate by stressing the importance of increasing the quality of the collaborative work. The expression is made to show that the concern is shared between them which less competent test takers were very likely not to do and instead, stressed the importance of this change as his or her personal concern, by using such phrases as “I’m worried about..” and “the presentation is important for me.”

The boundary of initiating the closing (lines 15-17) is clear. Also, the closing part is tailored to the given situation and to the discourse context, so that the message receiver (the imaginary classmate) could clearly understand what happened to the student (the test taker was playing a role of), what the student wants to do, and what the student wants the classmate to do.

**Score 3 Level of Social Action to Achieve the Communicative Goal**

Performances for score 3 showed generally adequate actions to achieve the given communicative goals. However, the evidence was not as consistent or as strong as that for score 4. The descriptor for score 3 states as follows:

Actions (e.g., elaboration of explanations of the test taker’s situation) are taken to deal with the given situation although:

- introduction or closing is not fully tailored to the situation (e.g., redundant, too simple) and/or
some actions are redundant or not appropriate (e.g., initiating long self-introduction to the professor who may be in a hurry) for the situation.

The features seen in performances for score 3 were characterized by introduction or closing not fully tailored or actions in part redundant or not appropriate somewhere in a performance. Excerpt 4 below provides an example of redundant introduction before the main topic. The opening part is not fully tailored for the context. In this situation, a student (test taker) has met by chance with a professor on campus. Although not explicitly stated in the task prompt, the professor might be in a hurry, but this test taker kept talking and, in fact, started to talk about her personal story (lines 18-21 and 24-26). The given situation is the same as the one described for Excerpt 1 above.

**Excerpt 4 (Task ID: D4, Professor topic, Dialogue)**

S=Student (Test taker), P=Professor (Interlocutor)

1 S:  hi↑ ah: Professor Smith
2 P:  [hi
3 S:  [how are you:?  
4 P:  I’m very well how are you?
5 S:  ye:ah good (0.4) oh ah I’m I’m the student in your  
6 class
7 P:  oh [*right*  
8 S:  [and I regifu ((transcribed as pronounced)) ah  
9 wagistered ((transcribed as pronounced)) for the um  
10 (0.4) it’s (.) elective (0.4) a subject  
11 P:  yes [( ) okay  
12 S:  [yeah: you would be [the professor this semester  
13 right?=  
14 P:  =yes I am  
15 S:  yeah nice to meet you am I’m ((first name))  
16 P:  ((first name)) nice to meet you (((name)))(.  
17 S:  [yea:ah am (0.5) I  
18 >really<like your: actually I really like your subject  
19 because(.)you know this semester the reason (.why I  
20 choose (0.3) your subject because I think this subject  
21 (is) very practical and it’s [very professional  
22 P:  [oh↑ great  
23 I’m >happy< to [hear that  
24 S:  [yeah↑ yeah thank you; >am I< really  
25 want to learn something professional you know improve  
26 my skills

124
The main topic and the main request come much later than line 26 above. Before
the main topic is touched on, the test taker continues her personal story and future plan,
which do not matter to the professor. As expected in the literature (Al-Gahtani &
Roever, 2012; Youn, 2013), both of the test takers in Excerpt 1 earlier and Excerpt 4 (in
the same task situation) took the action to lay the introduction before they got onto the
main topic in the situations where they handle pragmatic demands. However, the long
introduction in Excerpt 4 is qualitatively different from the previous test taker’s in
Excerpt 1 above.

Score 2 Level of Social Action to Achieve the Communicative Goal

Performances equivalent to score 2 were differentiated from those for the higher
scores by negative features exemplified in the descriptors. The descriptor for score 2
states as follows:

- Actions are minimal in order to prevent the problem and promote success
  and/or,
- the test taker’s insufficient understanding of the situation is reflected in the test
taker’s performance and/or,
- some actions (including introduction and closing) are not tailored for the
  situation and/or missing.

Excerpt 5 below lacks an introduction, a flaw very rarely seen in the
performances for the higher scores. In this task situation, a student (test taker) taking
Professor Brown’s class has visited the professor’s office to ask for an extension of the
due date of an assignment. In this opening part, hi Brown I I can’t I can’t
submit my final assignment to you (lines 1-2), the test taker gets to the main topic without first laying an introduction.

**Excerpt 5 (Task ID: D6, Professor topic, Dialogue)**

$S=$Student (Test taker), $P=$Professor (Interlocutor)

1 S: hi Brown I can’ I can’t submit my final assignment to you so I know that tomorrow is deadline of the final assignment=
2 P: =mmhm
3 S: and there is the rule that marks ( ) (lost) (lost)
4 at the rate of the fifty percent per day for (0.4) for late late sub submission
5 P: yeah=
6 S: =but I’m sick I have no- (.) I really have no time
to finish it (.) [and
7 P: [okay am =

Performances at score 2 level also reflected test takers’ insufficient understanding of the task situations. Excerpt 6 below illustrates a part of the test taker’s minimal actions to solve the encountered problem along with his insufficient understanding of the task situation. In this given context, a student (test taker) is addressing an administrator in an IT technical support office seeking support for the encountered problem (malfunction of a device for his class presentation). The problem needs to be fixed within twenty minutes before the class starts.

**Excerpt 6 (Task ID: D5, Administrator topic, Dialogue)**

$S=$Student (Test taker), $A=$Administrator (Interlocutor)

1 S: ah: excuse me
2 A: hello how can I help you=
3 S: =ah so: today I’m going to make a presentation from now but ah: the data projector in the classroom doesn’t work ah: [so=
4 A: [oh dear
5 S: =and I’m not sure (.) how can I: how can I solve this problem so:
6 A: um:
7 S: ah:: (.) could you tell me how can I do
8 ((continues))
A: unfortunately I can’t (0.6) I’m >the only one< in the
office at the moment=
S: =ye[ah
A: [so I can’t leave
S: yeah
A: and (1.3) and my (.) the only other available
technician is is out at [a job at the moment
S: [(out) ah so:
A: but he said he should be back in about twenty minutes
or so
S: [twenty minutes] ah: okay twenty minutes yeah twenty
minutes i:s ah o- ok and (0.8) so (0.7) ah: so can I
borrow ah: another devices (.) ah to: present a visual
material I think ah: whatever is ok if I can use visual
materials so [other than projector
A: [ah okay

The conversation is initiated by the test taker’s addressing the administration staff, ah: excuse me (line 1), and the interlocutor’s response, hello how can I help you (line 2). The test taker then takes an action to explain his encountered situation, ah so: today I’m going to make a presentation from now but ah: the data projector in the classroom doesn’t work ah: so (lines 3-5). This explanation is followed by his expressing uncertainty as to what he should do, and I’m not sure (.). how can I: how can I solve this problem so: (lines 7-8) which implies a question to the administration receptionist. Soon after that, the test taker asks an explicit question, ah:: (.). could you tell me how can I do (line 10).

The explanation that the test taker has given was not sufficient to make himself understood by the staff; in fact, it makes it difficult for the administrator (the interlocutor) to provide a proper support or to give advice because the test taker (student)’s explanation gives no specific information regarding the problem. The task prompt, in fact, provided more information regarding this situation more than what the sentence, ah so: today I’m going to make a presentation from now but ah: the data projector in the classroom doesn’t work ah: so (lines 3-5)
implies.

The test taker’s insufficient understanding of the given situation is reflected in lines 22-23. As described in the task prompt (APPENDIX B), the presentation is about to start in twenty minutes. In the middle of the conversation in lines 12-13, 15, 17-18 and 20-21, the administrator (the interlocutor) explains what is not possible, as instructed in prompts for the interlocutors (APPENDIX D). The interlocutor explains to the test taker that the IT technician to fix the problem would be available for the student in twenty minutes. The test taker’s response, twenty minutes ah: okay twenty minutes yeah twenty minutes i:s ah o- ok (lines 22-23) would completely lead to the failure in alleviating the encountered problem. In this task context, test takers are supposed to negotiate with the administrator to seek any possibility to prevent the failure from occurring immediately rather than having to wait for twenty minutes. Because the class is about to start in twenty minutes, the problem needs to be fixed before that.

Score 1 Level of Social Action to Achieve the Communicative Goal

Score 1 (the lowest score) was applied to performances which contained more noticeably flawed features of social actions. Even though test takers at this level were able to produce language output, their abilities to produce social actions were found inadequate, making a marked difference to the performances presented in score 4 level. The descriptor for score 1 states as follows:

- The negative features characterizing score 2 are more noticeable and/or,
- insufficient actions are taken or insufficient information is provided to prevent the problem and promote success in the situation and/or,
copying (using the same wording) of part of the task prompt is noticeable.

Excerpt 7 is an example of a performance for score 1, exemplifying the insufficient information and the insufficient actions delivered to the addressee. It is noticeable that the test taker copies a substantial part of the task prompt. In addition, both the opening and the closing are not tailored to the context. The given task situation is the same as in Excerpt 3 in which the student (test taker) is leaving a voicemail message to the classmate working on the same project.

Excerpt 7 (Task ID: M4, Peer student topic, Monologue)

1 hi Sharon am: I’m ((first name)) (0.3) and (0.8) u::m I need
2 your help to: (1.0) analyze the data (1.0) u: you know we
3 have (0.5) we have (.) developed the questionnaire and given
4 it to (.5) fifteen students but only two (of) them (0.5) were:
5 male so: I (1.0) ah: so I’m worried about that (0.8) the data
6 may not present (.) may not represent all students (0.5) so::
7 I sh- I found ano- another six (0.8) male students
8 participants(0.8)and I want to (0.8) ah: collect the
9 information from them (1.0) so: (.) u::m I need your help to
10 analyze those data (2.4) u::m (6.2) so can you can you give
11 me ah: text message after you:: (0.5) ah: after you: (1.8)
12 read this (1.0) voice message >because< I: I will ah because
13 u:m I think it- it is not good way to call me because I: need
14 to meet the: (1.0) participants so (1.0) thank you

The test taker opens the speech with a greeting addressing the classmate, hi Sharon (line 1), followed by his self-introduction to a close friend, am: I’m ((first name)) (line 1). This self-introduction is not tailored for this context. The equivalents to this part of many of the test takers are, hey Sharon >it’s< me ((first name)), as in Excerpt 3, “it’s ((first name))” or “this is ((first name)).” Immediately after the greeting and introducing himself, the test taker rushes to the main topic to tell the interlocutor without any introduction regarding the upcoming main actions or main
topic, u::m I need your help to: (1.0) analyze the data (lines 1-2). The subsequent action, we have (0.5) we have (.). developed the questionnaire and given it to (.). fifteen students but only two (of) them (0.5) were: male so: I (1.0) ah: so I’m worried about that (0.8) the data may not present (.). may not represent all students, (lines 2-7) could serve as the reminder to the classmate, which was identified by the most advanced test taker in Excerpt 3 above. However, this part is almost the exact copy of the word and the sentences provided in the task prompt. In the closing part, the test taker abruptly and awkwardly concludes the speech and the closing is not fully tailored for the context. After requesting the listener to give a text message back, the test taker seems to want to continue to say something, so (1.0) (line 14), but terminates it. It is not clear where in the performance the test taker initiated the closing action.

To evaluate the performance in Excerpt 7 more qualitatively, overall, this test taker’s output was almost a repetition of the information provided in the task prompt including the main message to deliver to the interlocutor. The test taker’s own work was the greeting and self-introduction in the opening and the requesting of the listener to return his call. These features were particularly highlighted by comparing it with the performance equivalent to score 4 (in Excerpt 3). The test taker in Excerpt 3 took actions including the opening and the closing sufficiently tailored for the context, a reminder of the shared concern, the current limitation of the corroborative research with the classmate, and the explanation and encouragement for the extra work. This advanced test taker’s way of reminding and encouraging the intended listener (the classmate) also highlighted the fact that the project was the shared work between them. In contrast, the current test taker in Excerpt 7 simply employed the words and the sentences in the task prompt and delivered to the listener what happened and what the
student (the test taker) wants to do.

The discourse data of the two pilot test takers also provided further evidence of lack of actions to support the descriptor of score 1. As described in section 3.3.4.5 in Chapter 3, the proficiency levels of these two pilot test takers were far below university-entry level. They had no experience of language activities at English-medium university. They had more limited ability to understand the task situation simulating university setting and to produce social actions.

Excerpt 8 is a dialogue performance of one of the pilot test takers on a task simulating a situation where a student is addressing a classmate to make a pair for a class presentation, as described in Excerpt 2. Insufficient actions, an action to rush to the main topic, an abruptly concluded closing, and a serious misunderstanding of the task instruction are all reflected in this performance.

**Excerpt 8 (Task ID: D3, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

1  S:  hi Sharon
2  P:  oh hello
3  S:  ah:: so (0.7) I (. ) want to: presentation with you
4     beca:use (. ) so I know you a- >no I’m sorry< I’m
5     interested in cooking and I (heard) you’re also
6     interested in cooking so: (2.0) so if you (0.8) if you
7     don’t (. ) decide any- o (. ) person (1.0) can you (1.0)
8     can you make pair (with me)=
9 10 P:  =oh possibly yeah am I was thinking of talking to Peter
11 ((continues))
12 13 P:  ok yeah yeah that could be possible
14  S:  mmhm
15  P:  (. )
16  S:  "yeah" ha have you asked anyone else in the class?
17  P:  [yet
18  S:  no not [yet
19  P:  [no?
20  S:  yes
21  P:  ye[ah?
22  S:  [((laughs))
23  P:  oh okay
24 S: so please consider about that
25  (0.5)
26 P: yeah okay I think I think that would work
27 S: thank [you ((laughs))
28 P: [yeah you are welcome ((laughs))

After a greeting in line 1, the test taker directly rushes to the main action in line 4 without any pre-expansions (Schegloff, 2007; Al-Gahtani & Roever, 2012, 2013). The misunderstanding of this test taker about the task instruction is reflected in the utterance, I’m interested in cooking and I (heard) you’re also interested in cooking (lines 5-7). The classmate’s interest is not cooking as described in the task prompt. The task prompt states that the classmate is interested in health-related topics and wishes to do a presentation with someone who has similar interests. The task prompt also requires the student (test taker) to convince the classmate to do a class presentation together by sharing interests of both students.

The closing part which includes the sentences, so please consider about that (line 24) and, thank you ((laughs)) (line 27) is markedly different from that in Excerpt 2 (representing score 4). In Excerpt 2, the highly competent test taker initiated the action of proposing a meeting with the classmate for their collaborative project after the test taker reached the mutual agreement with the interlocutor. Furthermore, the highly competent test taker’s attempt to initiate the closing was clear, followed by several actions to express her feelings and gratitude in the test taker in Excerpt 2. On the contrary, the current test taker’s insufficient and unclear actions makes the boundary of or even existence of the closing part unclear. Although these pilot test takers’ task samples were not the target for the rating, the performance in Excerpt 8 explains a part of the descriptor for score 1 indicating noticeable negative features of Social Actions to Achieve the Communicative Goal. In the monologue situations, the pilot test takers produced extremely limited outputs, some of which were
not even enough to be assessed in the other criteria, *Facility with the Language, Language Use to Deliver the Intended Meanings* and *Language Use for Mitigation*.

### 4.3.2.2 Facility with the Language

In this section, I will discuss the features of the next assessment criterion, *Facility with the Language*, according to the four levels. This criterion looks at how test takers deliver the speech contents to the intended addressees for the given communicative goal. The sub-features underlying this criterion are smoothness, clarity, fluency, control of sound including intonation, stress (Youn, 2013) and management of repairs (Kasper, 2006; Schegloff, 2007; Roever, 2011). This criterion does not simply assess whether language content is smoothly, clearly delivered and/or delivered at an appropriate speech speed, but it targets how the speech or utterance is delivered to the intended imaginary listeners, maintaining the clarity of the speech for the addressee. The tasks developed in the current study required the test takers to be aware of the given communicative situation (to whom, for what purpose, and in what situation) in which they delivered their performances.

*Social Actions to Achieve the Communicative Goal* (discussed above) targets adequacy and quality of social actions (e.g., initiating proposal and reminding) while the current criterion, *Facility with the Language* looks at how that content is delivered. Sound variation is not evaluated in the former criterion but in the latter. The current criterion, *Facility with Language*, also integrated management of repair as a sub-feature accounting for speech delivery. Repair is a mechanism used to manage speech when a problem is identified (Kasper, 2006). This is also recognized as an assessment component for L2 pragmatics (Roever, 2011). It is entirely possible that speakers achieve the given communicative goal without conducting repair regardless of the test
taker’s language abilities. Frequent use of repair may lead to successful communication for the goal or it may undermine smoothness of the speech if insufficiently controlled. Under the current rating rule, use of repair itself could not be assessed as a positive or a negative feature. However, it was found in the discourse data that insufficiently controlled repairs undermined *Facility with the Language*. Thus, repairs were negatively assessed when they were initiated but insufficiently controlled affecting the delivery of the speech.

*Score 4 Level of Facility with the Language*

The discourse data in the current study confirmed that pragmatically highly competent test takers as reflected in performances for score 4 were able to control their speech speed and sound variation, not simply being able to speak fast. They were also able to control repairs explicitly when necessary, while maintaining smoothness and clarity of the speech from the opening to the closing of a performance. The descriptor for score 4 states as follows:

Clear and fluent throughout the speech or conversation, and sound variation (e.g., intonation, stress) and speed control are explicitly seen. Repairs, restatement, insertions may be used to elaborate and/or to stress or to mitigate the content, but these actions are well managed and fluency is not seriously flawed. Short pauses may be seen but not noticeably affect clarity and fluency.

Excerpts 1, 2 and 3 reviewed for *Social Actions to Achieve the Communicative Goal* earlier also provided positive evidence of the highly competent test takers’ abilities to facilitate the speech. The speech is very clearly articulated. Also, test takers for this
level were able to control the sound variation as explicitly identified in the following excerpt. Despite the fact that speech occasionally slows down with short pauses and repetition, the smooth delivery is maintained throughout the tasks. The speech speed is explicitly increased (as in, and to get how we work and stuff >and if there is< any problem and >we just have to be honest< and like >you know< say hey >you have to work more< or your know stop pressing on me stop like that >you know< okay so ((laughs)) in lines 6-10 in Excerpt 2), and decreased (as in, so yeah↑ I’ll (0.5) so: I’ll I think I "ha:ve you:re email" or I’ll: have a check in lines 16-17 in Excerpt 2). The pitch is raised as in =cool↑ as in line 23 in Excerpt 2), and the parts are explicitly stressed by sound as underlined in I have good news and I have bad news in Excerpt 3. Test takers could maintain smoothness of the speech from the opening to the end of a performance. Put differently, the most performances deserving score 4 for this criterion were almost free from the negative features (for the lower scores), as discussed later.

As argued above, conducting repair was up to the test takers. In performances of pragmatically less competent test takers (as discussed later), purposes of repair or initiation of repairs were unclear because of their insufficiently controlled content delivery with little sound variation. This was the least likely flaw to be seen in performances for score 4. The purposes of repair and the boundary between the replaced part and the original part were much clearer when they initiated a repair to mitigate imposition, as in yeah am: I actually have (0.5) I don’t know if it’s good news or not in lines 12-13 in Excerpt 9 below. As well, speech was not noticeably inhibited even though a repair was conducted. Repair could be understood as an interactional behaviour when a problem is identified (Kasper, 2006). In the case of those at score 4 level, repairs were conducted for adjusting pragmatic meanings,
mitigating, or repairing information, not for correcting language errors or mistakes. The current test taker’s attempt to initiate the repair is clear because of the short pause laid before the replacing of the errant parts. Excerpt 9 below is a part of Excerpt 19 reviewed later.

Excerpt 9 (Task ID: D4, Peer student topic, Dialogue)
S=Student (Test taker), P=Peer student (Interlocutor)

1 S: hi Sharon ((laughs))
2 P: [hi ((first name)) how are you? 
3 S: good how are you? 
4 (.)
5 P: great
6 (.)
7 S: great?=
8 P: =we’ve= 
9 S: =that’s good to hear ((laughs))
10 P: finished data analysis so: (0.4) I’m 
11 really (0.7) I’m so happy (((laughs)))
12 S: [yeah am: I actually have 
13 (0.5) I don’t know if it’s good news or not (.D) but=  
14 P: =[“oh” 
15 S: =[I know we did all data analysis and (0.3) kind of  
16 started with our conclusion 
17 P: um:

Excerpt 9 is part of a dialogue performance of a test taker in a situation where a student (played by the test taker) is addressing a classmate regarding the collaborative project for a class (the dialogue version of task 4 described in Excerpts 3 and 7 above for the criterion, Social Actions to Achieve the Communicative Goal). In this situation, a student (test taker) is proposing an action to re-organize the corroborative project, which may possibly be rejected by the classmate (as seen later in this conversation but not provided in the above excerpt). As described in the task prompt, the student (test taker) has accidentally found additional research participants and that could remedy the current problem of the limitation of the collaborative project. However, including those additional participants would pose an extra work load burden on both the student and
the classmate. Therefore, repairing the original part, I actually have, to replace it with I don’t know if it’s good news or not (lines 12-13) could be considered to be functioning as mitigation.

In the above excerpts, the repair seems to adjust the speech and the performed repairs does not undermine the smooth deliver of the speech. The clearly delivered speeches with the words employed for the repaired and replacing part, makes it clear that the test taker attempted the repairs and why the repairs were initiated. Test takers whose performances were equivalent to score 4 provided evidence of maintaining the clarity and smoothness more consistently than those given score 3 as discussed next.

**Score 3 Level of Facility with the Language**

Performances at score 3 level contained positive features but did so inconsistently, because of the features as described in the descriptor for score 3:

- Generally clear and fluent although:
  - sound variation or speed is generally flat or unnatural and/or,
  - occasionally speech is paused (but resumed going).

Their performances were partially undermined by these features, which was why they were discriminated from the highest level. Excerpt 10 below demonstrates a performance for score 3, where speech is flawed by pauses, repairs and repetition while smoothness and clarity are generally maintained. Sound variation is mostly flat. The given task situation is the same as Excerpts 3 and 7 provide for *Social Actions to Achieve the Communicative Goal* in which a student (test taker) is leaving a voicemail message to the classmate working on the same class project.
hello Sharon. this is (first name) am from the presentation (of) the other class (.) a:m I’d like to talk about our presentation task (1.0) a:m because: (0.6) I just found out there is only two males out of fifteen people (0.5) in our survey (0.5) and then (.) I was thinking if (0.6) the there is a- like big limitation for this subject (.) and the search now (.) so (0.7) yeah so I was like wondering if there (0.4) is some other people which we can (. ) do the questionnaire (0.5) and then (.) I know that we’ve done like pretty much everything (. ) a:m (0.8) but am there’s a opportunity (that’s) I could mee:t six more m- am: (.) males for the survey (.) like today like this af- just (.) like now soon. a:m since I decided to do (1.2) a::m this questionnaire more (.) I have to (.) reorganize analyze analys ((transcribed as pronounced)) >sorry< analysis but (0.4) am I don’t have the skill to do it so I really need your help (.) to (0.7) do the conclusion and the analysis and "stuff" (0.4) because (1.0) yeah I don’t know how to do it (.) how to do it (.) so (.) I’m sure I’ll help you (.) to do it and I make a conclusion (0.5) because this is my fault that’s (.) we’re gonna do more but it’s ah I actually think it’s much better if we have like more female ↑ no more males because we only got lots of females and not many (0.8) am males now (0.6) u:m I really appreciate when you: (.) am listen this (0.8) ah this message (.) please call me back (.) but I would go: (.) to the survey with (.) I’m >going to< the questionnaire with other (1.2) six males student so (0.8) I’ll: can answer you (0.6) maybe in (.) in hour (.) so (0.6) I would appreciate it if you call me back (.) Thank you↓

The speech is in part paused in multiple places such as in, a:m I’d like to talk about our presentation task (1.0) a:m because:, (lines 2-3) and in since I decided to do (1.2) a::m this questionnaire more (. ) I have to (lines 13-14). The speech is resumed without any major pauses. False starts (as in the there in line 5) and repetitions (as in how to do it (.) how to do it in line 18) are seen but not as noticeably frequent as in the performances evaluated as score 2 below. Repair is also conducted in part (as in lines 14-15, I have to (. ) reorganize analyze analys ((transcribed as pronounced)) >sorry< analysis, and in line 21 (but it’s ah I actually think it’s), but it is clear that the test taker has intentionally conducted the repairs and how it was repaired. Although some grammatical errors (as in lines 7-8, I was like wondering if
there (0.4) is some other people which we can (.) do the
questionnaire are left unrepaired, generally intentions of repairs are clear and
successful once attempted. Leaving grammatical errors and mistakes without being
repaired is beyond the assessment target of Facility with the Language, which evaluates
insufficiently controlled repairs once attempted.

Score 2 Level of Facility with the Language

The difference between the higher and the lower score levels was marked by
whether smoothness was overall maintained or noticeably undermined. The descriptor
for score 2 states as follows:

Fluency is lost in some parts of the speech or conversation because of features
such as pauses between words, pause between sentences, repairs, repetition,
restatement, and false starts, although clarity is maintained generally.

Score 2 represents performances whose fluency was substantially lost in some
parts of the speech (because of pause, repair, repetition, restatement and false starts)
more than those for score 3. Excerpts 11 and 12 below exemplify the performances with
these negative features with its clarity being generally maintained.

Excerpt 11 describes the same task situation as Excerpt 6, where a student (test
taker) is asking the administrator for support for the encountered technical problem. As
illustrated in Excerpt 11 below, repetitions, and false starts are noticeably seen in some
parts of the speeches. A typical feature of the lower scores (score 2 and score 1) was
repetition of the same word several times, as in could you (0.4) could you- ah:
sorry ah could you help me right now? (lines 3 and 5).
**Excerpt 11 (Task ID: D5, Administrator topic, Dialogue)**

S=Student (Test taker), A=Administrator (Interlocutor)

1  S: am excuse me (. ) ah:
2  A: yes how can I help you=
3  S: =could you (0.4) could you- ah: sorry=
4  A: =mmhm
5  S: ah could you help me right now?
6  A: [um:
7  S: [now I’m connecting to my (.) a:m I’m connecting my
8  laptop to the data projector
9  A: yes
10 S: i in (0.8) but nothing (0.7) a:m (. ) nothing happened
11  (0.4)
12 A: oh
13 S: yeah nothing displayed on the compu- ah: on the
14  projector
15 A: oh really?

Repair to replace the original part, I’d connecting to my, with a:m I’m connecting my laptop to the data projector (lines 7-8), is successful. This repair shows the evidence of the test taker’s awareness of the language error in the original utterance. What makes it distinct from the advanced speakers is that the advanced speakers did not conduct repair for language errors and that the speech flow was not so obviously undermined as by a pause, (.), and a:m (line 7) laid between the original part and the replacing part in the excerpt above.

Although not always seen in every performance, another feature of performances equivalent to score 2 was unclear or insufficiently controlled repairs as in okay so ah: okay I (1.0) I I will explain - I explain this problem to my professor (lines 38-39) in Excerpt 12 below. As confirmed in these two excerpts, the same word (in the case of Excerpt 12, I, in lines 38-39), is repeated several times. In addition, a repair is conducted being combined with the repetition of I and explain in the same line. The original utterance, I I will explain seems to be replaced by I explain. In this context, the former referring to the action in the near future thus would be considered to be more appropriate and natural although the latter (merely a
statement) does not contain any grammatical errors or mistakes. Excerpt 12 below is a part of Excerpt 16 to be reviewed later.

**Excerpt 12 (Task ID: D5, Administrator topic, Dialogue)**

S=Student (Test taker), A=Administrator (Interlocutor)

29 S: so if I (0.4) can’t borrow
30 A: um
31 S: so could you (0.4) >could you help< me (of) this problem?
32 A: sure: okay
33 S: in the so (0.5) ah: the(.)technical officer
35 A: um:
36 S: will be back in a twenty minutes?
37 A: that’s ri[ght
38 S: [ah: okay so ah: okay I (1.0) I I will
39 A: ok[ay
40 S: I explain this problem to my professor
41 S: [so I think I can make a (0.3) time

It is unclear why the test taker replaced the original part by a less natural alternative, whether, in fact, he replaced the former with the latter intentionally.

Performances equivalent to score 2 provided considerably insufficient evidence of the speaker’s ability to control repairs efficiently. Sound variation was in the criteria as a sub-feature to distinguish score 4 from score 3 and it was not explicitly integrated in the descriptors for score 2 and score 1. However, as described in the three excerpts above, test takers whose performances were equivalent to these two lower scores showed insufficient ability to facilitate the language before evaluating their sound control. Thus, score 2 and score 1 did not necessarily reflect the test takers’ explicit control of sound variation. However, speeches of most of the test takers’ speaking performances at these band levels were generally flat as reflected in the two excerpts above showing the test takers’ struggle with delivering the utterances.

**Score 1 Level of Facility with the Language**

As for the other criterion, performances equivalent for score 1 were highly
undermined by negative features. The descriptor for score 1 states as follows:

Clarity and fluency are noticeably flawed because of features such as:
- unclear articulation and/or,
- insufficient control of repair (e.g., repair or restatement is attempted but given up leaving parts or sentence(s) incomplete and/or the original part is replaced by another mistake) and/or,
- delivery being choppy, fragmented and minimal and/or,
- combined and/or more frequent features characterizing score 2.

In the case of performance for score 1 level, clarity and smoothness were noticeably flawed in the speech, regarding the features discussed for score 2 and the utterances were unclearly articulated and fragmented. Excerpt 13 below is an example of a performance in score 1. Unlike the performances for score 2 above, which maintained clarity, in performances at score 1 level (as in Excerpt 13), unintelligible parts, combined with repetitions, repairs, and false starts were noticeable. In this task situation, a student (test taker) is leaving a voicemail message to the classmate working on the same class project, as described for Excerpts 3 and 7 (discussing Social Actions to Achieve the Communicative Goal) and Excerpt 10 above.

Excerpt 13 (Task ID: M4, Peer student topic, Monologue)

1 hi Paul (. ) I hav- ah: I think we have done ah ah: better
2 (1.0) better ( ) the presentation but I fo- a:m but I found
3 that we have (among among the) fifty we have do a
4 questionnaire with a fifty (. ) students but only two of them
5 are ah: are males a males a males so I thin- a:m I’m not sure
6 ( ) data is very is is represent the the: the: truth (. ) so:
7 I fo- I found another six more- ah male students in:: in our
8 class (. ) a: a in our- (. ) yeah so I: I need to u:m I need to
9 meet them ah: ah I (have) give them the questionnaire to
answer and also questions and a:m collect the a:m information
then and (I’ve rewrite) ( ) the the I’ve rewrite the
conclusions to make our ques- ah presentation is more (.) ah:
truthful and so (. ) um I don’t- (. ) but I ha- a:m I don’t
know how to use the analyzisthe ((transcribed as pronounced))
software so can you a:m a:m (0.6) if you see the message
please contact me >okay<? and I (have meet) I I need to go
see the ah: ah: participations soon (0.6) u:m okay.

The bracketed parts indicate the utterances transcribed were the best guess of the
transcriber (the researcher) with uncertainties such as, (among among the) (line 3),
and with utterances that were almost unintelligible for the transcriber as indicated by the
empty parentheses. These negative features combined with unintelligible utterances
could pose a greater challenge to the listener.

4.3.2.3 Language Use to Deliver the Intended Meanings

The next criterion to discuss is Language Use to Deliver the Intended Meanings.
It assesses the test takers’ ability to employ linguistic resources to deliver the intended
meanings when describing given situations by expressing emotions, providing opinions,
asking questions and by taking other actions to serve the intended communicative goal.
Ability to control linguistic resources is fundamentally crucial in the simulated
situation. Here, test takers need to make themselves understood by the addresses who
are tasked to provide support for the student (played by test takers in the role play
tasks). In real circumstances, undermined clarity in an utterance could distort the
speaker’s intention and lead to a misunderstanding of the hearers.

Some might argue that this assessment criterion simply refers to the
measurement of language accuracy in oral discourse measurements. Accuracy may be a
primary requisite for the speaker to deliver the intended meaning and to handle
pragmatic demands. However, beyond language accuracy measured in a
decontextualized task, the criterion, Language Use to Deliver the Intended Meanings is
intended to measure language use in a context where test takers are addressing the intended addressees. It is entirely possible to construct a sentence syntactically accurately, which does not reflect the speaker’s intention as exemplified later in this section.

**Score 4 Level of Language Use to Deliver the Intended Meanings**

Performances deserving Score 4 were characterized by showing sufficient control of linguistic resources (lexical choices and sentence structures) to deliver the intended meaning. The descriptor for score 4 states as follows:

The test taker’s performance shows explicit control of linguistic resources (lexical choices and sentence structures) with few unclear parts to deliver the intended meanings throughout the speech. Employment of linguistics resources is natural.

What made score 4 level distinct from the other score levels was that the clarity of the intended meaning was maintained from the opening to the closing of a performance as seen in the examples following. The examples provided here were extracted from the excerpts demonstrating score 4 in the previous criteria (*Social Actions to Achieve the Communicative Goal and Facility with the Language*). Word choice was natural as was the sentence structure as exemplified in the following sentences:

- I wrote you about um (0.7) a >problem that< I discovered when (0.5) umm when I looked at my schedule the other day *(lines 14*
and 16-17 in Excerpt 1)

- I’m also very well aware of the fact that umm (.) we have to present in four days (lines 6-7 in Excerpt 3)

- I think um it’s in interest of our research to: make it as balanced as we possibly can (lines 14-15 in Excerpt 3)

Receiving score 4 did not necessary require perfect accuracy in their language use nor the sophistication of sentence structure used by those above, so long as very few unclear parts were seen and the employment of linguistic resources was natural. In other words, the performances equivalent to score 4 can also be accounted for by being free from negative features seen in scores 3, 2, and 1 discussed below.

Score 3 Level of Language Use to Deliver the Intended Meanings

In the case of performances equivalent to score 3, intended meanings were partially flawed by minor negative features. The descriptor for score 3 states as follows:

The performance contains the following features although they do not seriously obscure the intended meaning:

- unnatural word combination or lexical choice and/or,

- missing or unnecessary components to structure the sentence or incorrect word order (esp. in a wh-clause) or a different part of speech word is used (e.g., “data analyze” instead of “data analysis,” “consideration for healthy” instead of “consideration for health”).

Score 3 was awarded because the utterance partially contained either one of
these negative features above. Utterances containing these features would not seriously undermine the speaker’s intended meaning although they might have required interlocutors to guess the speakers’ intended meaning(s). Also, even with these features, sentence structures in performances for score 3 were clear enough to serve the intended actions.

Excerpt 14 provides an example of a different part of speech used to construct a sentence (data analyze in lines 2 and 6 instead of “data analysis”), which is seen twice. The task situation is the same as the one in Excerpts 3, 7, 10, and 13 where a student (test taker) is leaving a voicemail message to the classmate working on the same project.

**Excerpt 14 (Task ID: M4, Peer student topic, Monologue)**

1 am hi Paul this is (.) (first name) am I just finished
2 reading(0.4) about the conclusion and the data analyze that
3 you have been done- that we have been done and I think it’s
4 great
5 (continues)
6 but I need your help to do the ah: data analyze because I’m
7 not quite sure how to use this software (0.3) am: so I’m
8 thinking

Language use of test takers whose performances represented score 3 was generally controlled well enough to maintain the clarity of what they wanted to say. The choice of the linguistic resource (“analyze” as in lines 2-3, the data analyze that you have been done) changes what the test taker should have wanted to say, “data analysis.” As such, the positive evidence seen in performance for score 3 was not as consistent as that for score 4. However, performance equivalent to score 3 was distinct from score 2 (as discussed next) in that the intended meanings were not seriously undermined by the features as discussed below.
Score 2 Level of Language Use to Deliver the Intended Meanings

Performances given score 2 contained linguistic resources which were insufficiently and/or inconsistently controlled to deliver the intended meanings. The descriptor for score 2 states as follows:

Linguistics resources are insufficiently or inconsistently controlled with:
- unnatural part(s) or word combination(s) that change or obscure the intended meaning (the intended meaning should be different) and/or,
- non-native like choice of linguistic resources being more noticeable and/or,
- the negative features characterizing score 3 being more noticeable

Performances for score 2 were undermined more than those for score 3, as evidenced by the negative features seen in the performances. These included linguistic resources wrongly employed, or a non-native combination of linguistic resources, and/or parts of speech changing or obscuring the meaning and/or where faulty features of score 3 being more noticeable.

One of the features being more noticeable (as shown in Excerpt 15 below) was a non-native choice of linguistic resources. In this situation, a student (test taker) is trying to convince the classmate (working on the same project) to re-organize the presentation. This test taker is on the same task as the test taker in Excerpt 9. The word choices are unnatural with one part in particular where the intended meaning is hardly interpretable.

Excerpt 15 (Task ID: D4, Peer student topic, Dialogue)

S=Student (Test taker), P=Peer student (Interlocutor)

1  S:  yeah it’s[good]
2  P:  [six new] participants we have to we have to
3  interview them that’s gonna- we’ve already done the
4  analysis on the data you are looking actually including
In line 6, the test taker said, our destination is, to stress the importance of the collaborative project for the final evaluation of the class. However, the noun, destination, should have been “our goal”, or other similar alternatives. Lines 11-13, maybe I can I can ah: (0.5) introduce why the it is really important for ah: me to ah: put more people in the ah: in the:: the questionnaire, also demonstrate noticeable non-native word choice (introduce why the it is).

What is more, the intended meaning of introduce why.... in the ah: in the:: the questionnaire itself is difficult to interpret as the test taker’s intention of the utterance is also very unclear as to who is to be introduced and for what purposes. In this performance, the test taker and the interlocutor had reached a mutual agreement on re-organizing the presentation as mentioned in the title of this task (see APPENDIX C). Thus, this unclear utterance may have made sense if it was put thus: “maybe I can write in the final report why it is important for our research to include more people in the data.”

The subsequent interlocutor’s utterance (lines 14-17), y::eah yeah yeah we we’ve discussed that now it’s great if you’ve gone out and find a more people ‘cos it was a weakness in the study, seems to be a general
comment, summarizing the conversation, which could be provided without understanding the test taker’s previous concern accurately. This interlocutor’s utterance appears to respond to “important” and “to ah: put more people in the ah: in the:: the questionnaire” in the test takers’ previous utterance. Even so, the test taker’s whole utterance remains unclear. Put differently, it would be possible that the interlocutor’s general comment reflects unclear meanings and intention of the test taker’s utterance.

Excerpt 16 below demonstrates a performance with the negative features for score 3 being more noticeable and includes a section the intended meaning of which was unclear. An incorrect word order to structure an utterance using a wh-clause, one of the features seen in performances of score 3 is identified twice in lines 7-8 and 10, and

I’m not sure (.) how can I: how can I solve this problem so: and could you tell me how can I do. The structure of the sentence is grammatically incorrect, making it unclear whether or not the test taker is forming a question. In another part, components necessary to the structures are missing, I go: a facility office (line 44), and (1.0) and so- (0.7) and so and so (1.0) from now I go: a facility office and I soon go- come back (lines 43-44 and 46). Again, the test taker’s intention of these utterances and whether he is talking to himself or is asking the interlocutor are very unclear. In addition, unnecessary components are added, in a twenty minutes (line 36) and I can make a (0.3) time (line 41). Unnatural word combination is seen in line 24 another devices, which the test taker supposedly attempts to say, “some kind of device” as written in the task prompt (see APPENDIX B), “other devices,” or “another device.” Furthermore, although not completely inappropriate, the intended meaning of whatever is ok if I can use visual materials so other than projector (lines 25-26) in this context does
Excerpt 16 (Task ID: D5, Administrator topic, Dialogue)

S=Student (Test taker), A=Administrator (Interlocutor)

1 S: ah: excuse me
2 A: hello how can I help you=
3 S: =ah so: today I’m going to make a presentation from
4 now but ah: the data projector in the classroom doesn’t
5 work ah: [so=
6 A: "oh dear
7 S: =and I’m not sure (. ) how can I: how can I solve this
8 problem so:
9 A: um:
10 S: ah:: ( . ) could you tell me how can I do
11 ((continues))

12 A: unfortunately I can’t (0.6) I’m >the only one< in the
13 office at the moment=
14 S: =ye[ah
15 A: "[so I can’t leave
16 S: yeah
17 A: and (1.3) and my (. ) the only other available
18 technician is is out at [a job at the moment
19 S: [(out) ah so:
20 A: but he said he should be back in about twenty minutes
21 [or so
22 S: [twenty minutes] ah: okay twenty minutes yeah twenty
23 minutes i:s ah o- ok and (0.8) so (0.7) ah: so can I
24 borrow ah: another devices (. ) ah to: present a visual
25 material I think ah: whatever is ok if I can use visual
26 materials so [other than projector
27 A: [ah okay
28 ((continues))

29 S: so if I (0.4) can’t borrow
30 A: u:m
31 S: so could you (0.4) >could you help< me (of) this
32 problem?
33 A: sure: okay
34 S: in the so (0.5) ah: the (. ) technical officer
35 A: um:
36 S: will be back in a twenty minutes?
37 A: that’s ri[ght
38 S: [ah: okay so ah: okay I (1.0) I I will
39 explain- I explain this problem to my professor
40 A: ok[ay
41 S: [so I think I can make a (0.3) time
42 A: okay [all right
43 S: [Yeah thank you and (1.0) and so- (0.7) and so
44 and so (1.0) from now I go: a facility office and
45 [I soon
46 A: [sure
47 S: go- come back

not make much sense.
48 A: okay all right ((laughs)) [good luck
49 S: [((laughs)) thank you very
50 much
51 A: okay

Each of these negative features seen in parts of a performance may be minor but
the noticeable combination of these negative features which also contained incompletely
structured utterances, thus obscuring the meanings was distinct from performances
given score 3 level.

Another feature characterizing score 2 is exemplified in Excerpt 17 below,
where a performance containing unnatural part(s) and/or word combination(s) that
changed the intended meaning. This feature was minimally identified in the
performances for the higher scores, making it a distinctive feature for score 2. This
performance is the middle to the closing part of a monologue performance (voicemail
message to a professor, the monologue version of the professor task described for
Excerpts 1 and 4).

**Excerpt ID: 17 (Task M1, Professor topic, Monologue)**

1 I want (0.5) get your: (0.5) get you ge get your signature
2 for the(0.5) ah:(0.4) for the application (0.6) and the send
3 it (0.5) to send it to the adminst- admin administration
4 office before five pm tomorrow (0.7) so I need- ah I need ah
5 if you have ss- if you are convenient (. ) please call me back
6 (0.5) ah:: (0.7) please call me back thank you

The test taker (playing a role of a student) was asking a professor as the intended
listener of the voicemail message to return his call by saying, if you are
convenient (. ) please call me back (0.5) ah:: (0.7) please call me
back (lines 5-6). The test taker’s intention should be different in a given context where
the student (test taker) is asking about the professor’s convenience. The intended
meaning is supposed to be “if or when it is convenient for you” as the test taker
probably attempted to express her consideration for the professor’s convenience.

**Score 1 Level of Language Use to Deliver the Intended Meanings**

In performances for score 1, the speaker’s intended meanings were highly undermined because, as stated in the descriptor, performances equivalent to score 1 contained noticeable negative features. The descriptor for score 1 states as follows:

The performance contains:

- errors and/or unintelligible parts that obscure the meanings and/or,
- part(s) where sentence structure is collapsed (clause boundary is unclear and/or jagged, most likely due to repairs, cancellation, and/or combined language errors or mistakes).

In the case of performances at score 1, meanings were severely undermined by flawed features which obscured the meaning, such as lexical errors and/or unintelligibility or collapsed sentence structures. Collapsed sentence structures were likely to occur due to multiple repairs, cancellations, combined lexical errors, or mistakes. Excerpt 18 below demonstrates features accounting for score 1. Because of the unclearly articulated utterances combined with the unclearly pronounced words and the partially collapsed structure, it is difficult to discern what the test taker intends to express or how he intends to structure a sentence. The discussion below relies fully on the best of the researcher’s interpretation.
Excerpt 18 (Task ID: D2, Administrator topic, Dialogue)
S=Student (Test taker), A=Administrator (Interlocutor)

1 S:  excuse me ah excuse me ah can
2 I have you name?
3 A:  oh oh my name is ((first name)) >what can I do<
4 for you =
5 S:  yeah my name is ((first name)) actually a: (0.8) I
6 have a timetable ah timetable clash because am: (0.6)
7 ah: the e- elective a is ah: um um: has a the taime ((
8 transcribed as pronounced)) ah same time to
9 compulsory so I have to change ah: ah elective a to
10 elective b=
11 A:  you are leaving late aren’t you we’ve got fifteen
12 minutes (we’re) closing
13 S:  yeah: yeah: I know so m- I: (will) speak the con-
14 situation shortly beca[use a: I got am (0.7)
15 A:  yeah
16 S:  ah I got I got a elect signature from ah: the pro
17 professor Smith=
18 A:  that’s right you need to get two signatures
19 [is that all right? yeah
20 S:  yeah yeah yeah
21 (0.5)
22 S:  so ah: I don’t know can I hap ((transcribed as
23 pronounced)) you ah: ah: so (0.3) I don’t know u: ah:
24 (if): you can a: send a email or ah ta- a call the:
25 teacher from ah elective b maybe it’s "ah: okay"

The meaning of the part, I have a timetable ah timetable clash
because am: (0.6) ah: the e- elective a is ah: um um: has a the
.taime ((transcribed as pronounced)) ah same time to compulsory so I
have to change ah: ah elective a to elective b (lines 5-10), may be
interpretable even though it contains an unclearly pronounced word, taime and
unnatural word employment, the e- elective a is ah: um um: has a the
taime ((transcribed as pronounced)) ah same time to compulsory.

Lines 13-14 and 16, yeah: yeah: I know so m- I: (will) speak the
con- situation shortly beca[use a: I got am (0.7) ah I got I got a
elect signature, contains a noticeably non-native and unnatural word combination,
speak the con- situation shortly (lines 13-14), with a part difficult to
transcribe because it is unclearly pronounced, will, bracketed with double parentheses.
The intended meaning is supposed to be “I’ll explain the situation.” This part is followed by because (line 14), the purpose of which is to describe a reason for particular outcome. However, there are no parts that could be connected with the word, because (line 14). It could be interpreted that the test taker might have cancelled a subsequent line, although this is not certain. In the same utterance, a elect signature (line 16), should be “the signature from the Elective A teacher” as written in the task prompt (see APPENDIX C). The test taker’s incomplete utterance, a elect, may be a partly cancelled utterance of “an elective” or the test taker might have intended to describe the signature by his actual output, a elect signature.

The language use is very insufficiently controlled to deliver the meaning. The last lines, so ah: I don’t know can I hap ((transcribed as pronounced)) you ah: ah: so (0.3) I don’t know u: ah: (if): you can a: send a email or ah ta- a call the: teacher from ah elective b (lines 22-25), creates even greater difficulty in interpreting the meaning. In particular, the part, so ah: I don’t know can I hap ((transcribed as pronounced)) you ah: ah: (lines 22-23), contains an unclearly pronounced part, which could be “have” or “help.” Whichever is the case, it still does not make much sense. Moreover, it could not be judged by the test taker’s tone of voice whether the part, can I hap ((transcribed as pronounced)) you, is a question or a statement. The subsequent part, I don’t know u: ah: (if): you can a: send a email or ah ta- a call the: teacher from ah elective b maybe it’s "ah: okay", contains another part unclearly pronounced (described by the empty brackets indicating the transcriber’s best guessing of what was pronounced), which could be “if.” If we are to suppose that this test taker attempted to seek the administrator’s (the interlocutor’s) help, the test taker’s intention should be “I don’t know if you can.” However, this is not connected naturally
with the subsequent part, maybe it’s “ah: okay” (line 25). If it is the case that the word “if” referred to a condition, the sentence should have been, “if you can a: send a email or ah ta- a call the: teacher from ah elective b maybe it’s “ah: okay”,” which at least maintains the sentence structure. However, the part, maybe it’s “ah: okay”, is extremely unclear as to what the test taker found to be okay. In this task situation, the test taker (playing a role of a student) was expected to explain the encountered situation and problem to the administrator (the interlocutor) to make himself clearly understood. However, the test taker did not do so adequately for the interlocutor to understand what was concerned with being “okay.”

In the above example in lines 22-25, the unclearly pronounced parts makes it difficult to discern what are the intended words or whether the words originally existed or not. Even if the words were found to be significant, the meaning and the test taker’s intention remains unclear. In addition, boundaries of the parts of the sentence are collapsed. These unclearly constructed utterances together with the ambiguity of the meaning of each of the employed words can conceivably allow for an alternative and mistaken explanation about what was meant. The ambiguity and collapsed utterance in turn exemplifies the evidence of the test taker’s insufficient ability to handle language use to express the meaning clearly and accurately.

4.3.2.4 Language Use for Mitigation

This section will discuss the features of four different levels for the next criterion, Language Use for Mitigation. Each task simulated university situations where students are required to recognize and to handle pragmatic demands in order to achieve the given communicative goal, which is crucial for preventing failure or for promoting success for the student’s academic work. In the given contexts, test takers need to
provoke the addressee to take actions for themselves. Their actions are expected to pose imposition upon the hearer, which needs to be recognized by the speaker and be appropriately dealt with. *Language Use for Mitigation* evaluates test takers’ varied linguistic resources for mitigating imposition on the addressees and their ability to employ the linguistic resources to be connected in extended discourse for mitigation.

**Score 4 Level of Language Use for Mitigation**

The test takers’ discourse data provided the features discriminating the test takers’ language use for mitigation and pragmatically appropriate linguistics expressions (Youn, 2013). More competent test takers’ language use for mitigation was characterized by the features identified in the literature including bi-clausal structure “wonder if” or “wonder whether” (Youn, 2013) and modal verbs, “would,” “could” (Youn, 2013; Bardovi-Harling, 2013) in statement forms for mitigation. In the discourse data obtained in the current study, pragmatically competent test takers were able to introduce these linguistic resources, although their linguistic resources were not restricted to them alone. Pragmatically less competent test takers were more likely to rely on a limited range of expressions such as “can you,” “could you,” “please” and so forth structured as a simple sentence. Excerpt 19 below and other excerpts reviewed above such as Excerpt 1, illustrate performances deserving the highest score in which the advanced features for mitigation were used as predicted by the related literature and employed naturally in the flow of the speech or the conversation. The descriptor for score 4 states as follows:

Lexical choices and/or combinations well mitigate imposition with devices including (bi-clausal, conditional: I wonder if/whether and/or modal verbs such
as “would,” “could,” “might” in a statement form). Even if these devices are not used and/or more direct expressions are combined, the test taker does not simply rely on direct expressions. The linguistic resources used for mitigation are naturally connected in the speech or in the ongoing flow of the conversation.

Excerpt 1 (reviewed for Social Action to Achieve the Communicative Goal) above illustrates a part of a performance representing score 4 level with pragmatically appropriate expressions as indicators of advanced speakers (Youn, 2013). In the excerpt, the student (test taker) addressed Professor Smith. After explaining the encountered issue to the professor, the test taker was asking the professor to take an action for the test taker, I was wondering whether you could (1.0) you could sign that form for me (lines 25 and 28-29) using bi-clausal structure “wonder if” or “wonder whether” and modal verbs, “could” in a statement form as discussed in the literature as the advanced features of pragmatics.

As reported in the literature, competent test takers were able to use either (a) bi-clausal structure, “wonder if/whether” or (b) modal verbs, “would,” “could,” “might” in statement forms or a combination of both. Even if these linguistic tools for mitigation were not utilized, competent test takers’ language use for mitigation was not limited to simple expressions (“can you?,” “could you?,” “please do,” “I need your help” and so forth) which less competent test takers (as discussed later) were the most likely to rely on. Excerpt 19 below (an extension of Excerpt 9) illustrates the equivalent performance of score 4 level. The task situation is the same as described for Excerpt 9, where a student (test taker) is addressing the classmate to re-organize the collaborative work, which may cause extra burden on both students.
Excerpt 19 (Task ID: D4, Peer student topic, Dialogue)

S=Student (Test taker), P=Peer student (Interlocutor)

1  S:  hi Sharon((laughs))
2  P:  [hi {{first name}}) how are you?
3  S:  good how are you?
4  (.)
5  P:  great
6  (.)
7  S:  great?=
8  P:  =we’ve=
9  S:  =that’s good to hear {{laughs}}
10 P:  [finished data analysis so: (0.4) I’m really (0.7) I’m so happy {{(laughs)}}
11 S:  (0.5) I don’t know if it’s good news or not (.) but=
12 P:  =["oh"
13 S:  =[I know we did all data analysis and (0.3) kind of started with our conclusion
14 P:  um:
15 S:  but I just found out that we could get six more;
16 actually the male students in survey (0.7) a:nd
17 P:  =[oh:]
18 S:  I know it’s gonna be additional data but I see that we only have (.) two male and ah=
19 P:  =yeah: that’s true:
20 S:  there is a >bit of< bias in that (.) that was unrelated
to our "conclusions"
21 P:  (0.5)
22 I:  that’s "true"
23 S:  and: I know >it’s gonna< take more time but I think by getting six more male a:m students results (.) I think (.) we could do more better a:m presentation and I >needed to< ask help (.) because am: I don’t know how to use the actual analysis software (.)
24 P:  oh: right oh my gosh
25 S:  "{yeah": and I’ll collect the data as soon as possible from those male participants [but-
26 P:  =[really?
27 S:  yeah but you know but I >just< wanted to know if you are happy (.) with (.) [getting additional data
28 P:  =[oh: {{(laughs)}}
29 S:  thanks for asking=
30 P:  =(laughs)) yeah ah
31 S:  =((first name)) yeah: I’m really swamped this week with my part time job and=
32 P:  ="yeah"
33 S:  I have other assignments due next week so and I feel really (.). stressed with tho[se but
34 P:  ="yeah": sorry about the last minute (.). yeah notice but am:
35 S:  I’ll [try and do as much as I can as well=
36 P:  [( )
37 S:  =and if you could just guide me though the process of using the system
The test taker is implying a proposal of her idea. The test taker’s utterance in lines 39-40, yeah but you know but I >just< wanted to know if you are happy (.) with (.) getting additional data, which implies the test taker’s opinion is clearly different from those such as “I want you to do,” “you need to,” “please do,” and so forth. A bi-clausal structure used in this action, I >just< wanted to know if you are ...(lines 39-40) is a linguistics form used as a feature of L2 pragmatics identified in advanced speakers (Youn, 2013). Apart from the bi-clausal structure using if, this test taker used bi-clausal structure also using another linguistic resource, namely, I know we did...(line 15), I know it’s gonna be ... (line 21), I know >it’s gonna< take...(line 29), I think (.) we could...(lines 30-31). As the use of bi-clausal has been discussed in the literature as a trait of development of L2 learners’ pragmatic abilities (Youn, 2013), these utterances are indicators of the test taker’s ability to recognize and to handle pragmatic demands by mitigating imposition and/or stressing the test taker’s intention.

What is also noticeable here is that the test taker’s utterances are tailored for the context, serving to deliver the test taker’s understanding of the classmate’s situation, as well as to mitigate the imposition of what the test taker wants to do and to stress the point that needs to be understood by the classmate. These linguistic resources are constructed to mitigate the imposition were all naturally integrated in the flow of the extended discourse performance. Use of modal verbs, “could,” “would” in a statement
form for mitigation (not to refer to events in the past) is another developed feature to handle pragmatic demands, which is seen in I just found out that we could get six more: ... (line 18), I think (. ) we could do more... (lines 30-31), and if you could just guide me though the process... (line 53).

**Score 3 Level of Language Use for Mitigation**

In the case of performances deserving score 3, imposition was generally mitigated with more or less varied linguistic resources but the evidence was not as strong as for score 4. The descriptor for score 3 states as follows:

Imposition is mitigated generally although the test taker’s speech contains:

- linguistic resources used for mitigation abruptly inserted or not always connected in the speech or in the ongoing flow of the conversation naturally and/or,

- lexical choices/combinations not effectively mitigating or face-threatening.

Performances deserving score 3 contained features such as linguistic resources for mitigation in part abruptly inserted or not naturally connected in the extended discourse or not effectively mitigating and/or face-threatening. Excerpt 20 below provides an example of a somewhat abruptly inserted act (would you like to work with me to make this presentation?) in line 6 after 1.4 seconds of silence (line 5). In this task situation, a student (test taker) is leaving a voicemail message to the classmate to ask him or her to make a pair for the class presentation (the monologue version of the task described for Excerpts 2 and 8). Soon after the test taker briefly has explained the situation to the unfamiliar student (classmate) (lines 3-5), she requests that
they work on the project together, in a somewhat abrupt and direct way (line 6). It does not necessarily mean that this way of delivering the act is inappropriate. However, test takers whose ways of constructing mitigation deserved score 4 performed distinctively from that seen in Excerpt 20.

**Excerpt 20 (Task ID: M3, Peer student topic, Monologue)**

1. hi Sharon I’m ((first name)) (. remember we met in the first
2. class societal issues today? (3.0) I don’t think we met (0.6)
3. >at any< other classes (0.6) now we are supposed to do
4. presentation (.) I tried to reach you before (0.6) I tried to
5. reach you couple of days ago but (0.7) failed so:: (1.4)
6. would you like to work with me to make this presentation?
7. (1.0) ah actually I think that your interests are very close
8. to mine >that’s why we’d make< (0.3) the best couple here
9. (0.6) a:m (2.0) as far as I know >you’re interested< in
10. people’s perception of health and what people to do stay
11. healthy (0.5) well that’s exactly what I’m interested in
12. (0.3) but from the point of cooking

Despite the insufficient evidence to deserve score 4, their range of language use for mitigation represented in performances of score 3 was not as limited as and thus was distinguished from performances for score 2, which I will discuss next.

**Score 2 Level of Language Use for Mitigation**

Less competent test takers were more likely to rely on a limited range of relatively simple expressions, such as “I need your help,” “can/could/would you help me?” “can/could/would you…,” “can I,” “could I,” “please help me,” and/or “Please do,” and so forth. The use of these expressions should not of itself be viewed as inappropriate. In fact, in Excerpt 19 above representing score 4, the test taker uttered, I >needed to< ask help (lines 31-32), which is relatively simply structured. However, test takers at score 4 level used different linguistic resources to formulate other expressions as well. In performances of competent test takers, these simply structured
expressions illustrated above were one of the linguistic devices that they could utilize depending on the context. Mitigation device of less competent test takers was characterized by limited and relatively simple linguistic resources, direct expressions for mitigation and inconsistent use of complex structures for pragmatic meanings. The descriptor for score 2 states as follows:

The test taker is more likely to rely on limited and/or direct expressions (e.g., I need, I want) and “can you (please)?” “could you (please)?” or simple structures, considering the amount of speech and/or the test taker inconsistently uses complex structures for pragmatic meaning.

Excerpt 21 below illustrates a performance of score 2, which demonstrated a limited range of simple expressions for mitigation and never showed the advanced features identified in the literature. This performance makes a clear contrast with the test takers in Excerpts 19 and 20 above. The task situation was the same as for Excerpts 3, 7, 10 and 13 where a student (test taker) is leaving a voicemail to the classmate working on the same project. The voicemail is to propose an action to re-organize their work, which may generate additional burden on both students, thus may be rejected by the classmate.

**Excerpt 21 (Task ID: M4, Peer student topic, Monologue)**

1 hi Sharon ah: still questionary survey so: (.) today I found
2 six ah: male who (0.3) who can (1.3) who can join my- ( )
3 who can answer the (.) our questionnaire survey >and< I wanna
4 do additional ah: additional survey and I wanna do- I wanna
5 add these (.) additional data ah: so >is that< ok? I th: nk
6 the data I hav- we have now i: s not enough and (0.6) and also
7 I’m not sure how can I use the data anal- analysis so (.) I
8 need your help ah: (0.8) so: (0.6) please return (.) and
9 please give return call soon and (0.5) I think we have ah we
The test taker expresses his wish (to do the additional work) by repeating the words, I wanna as in, I wanna do additional ah: additional survey and I wanna do- I wanna add these (.) additional data (lines 3-5), followed by somewhat abruptly inserting the action to seek an agreement from the classmate, >is that< ok? (line 5). The subsequent expression for request, I need your help (line 7-8), is an expression that less competent test takers were more likely to rely on. The concluding part is initiated with a request for a call back to the test taker, so: (0.6) please return (.) and please give return call soon (lines 8-9). The mitigation devices in this performance are mostly simply structured from the opening and the closing (request for a call back).

The difference in the language use for mitigation was made clear when compared with the performance in Excerpt 3 representing score 4 (in the same task situation although the focus was on Social Actions to Achieve the Communicative Goal). In Excerpt 3, an attempt to convince the classmate was made with clearly different linguistic resources in addition to more diverse actions in mitigated ways. For example, the test taker in Excerpt 3 expressed her desire with the words, I’m also happy to< do the: (.) analysis, and subsequently, I (.) I just thought it would be fantastic to have those additional speakers because, which were noticeably different from that of the less competent test taker in sentence structure and quality of word choice. The pragmatically highly competent test taker’s requesting the classmate for a call back, >if you could call me back that would be great<, was also markedly different from the corresponding remark that the less competent produced in the current excerpt, please return (.) and please give return
call soon. Using simple expressions itself is not necessarily inappropriate as they are one of the linguistic tools for communication. However, mitigation devices representing score 2 such as the test taker’s in the excerpt above were limited to simple expressions.

Score Level of Language Use for Mitigation

Language Use for Mitigation representing score 1 was further undermined by features such as mitigation devices and/or ways of mitigation being noticeably limited, or by the use of simple sentence structures and limited expressions. The descriptor for score 1 states as follows:

Mitigation devices and ways of mitigation are noticeably limited.
The test taker is most likely to rely on simple structures and to use limited expressions.

Excerpt 22 below demonstrates a performance representing score 1 in which the ways of mitigation were noticeably limited. In this situation, a student (test taker) is addressing the classmate asking him to make a pair for a class presentation (the same task described for Excerpts 2 and 8). This performance reflects the test taker’s insufficient understanding of the task instruction, which is assessed by another criterion, Social Actions to Achieve the Communicative Goal. The performance in Excerpt 22 also provides solid evidence that the test taker’s linguistic resources for mitigation are very limited and thus ought to be assessed accordingly.
Excerpt 22 (Task ID: D3, Peer student topic, Dialogue)

S=Student (Test taker), P=Peer student (Interlocutor)

1  S:   hi Sharon
2  (0.5)
3  P:   hi:
4  S:   my name is ((first name)) (. ) do [you remember me?
5  P:   (((first name)) yeah I remember your face definitely
6  (. )
7  S:   um yeah (1.0) um: so do- um: (0.5) now I have- I have
8  to work- I have ah:: a presentation to do and (0.5) I want a (. ) partner
9  (0.8)
10 P:   ah:: okay yep
11  (. )
12 S:   yeah so: can you be- ah can work with me?
13  (0.6)
14 P:   oh right: okay mm: (2.0) can you tell me little bit
15 more about yo:ur interests?
16  (1.0)
17 S:   [oh]
18  P:   [because I’m actually ah yeah I was thinking about
19 contacting Lucy another classmate ['cause we have=
20  S:    [oh]
21 P:   =very similar interests but am: (0.4) I haven’t (. ) I
22 haven’t contacted her yet so (. ) a:m
23 S:   yeah:
24 P:    yeah
25  S:   I [I’m inter-
26 P:    [tell me bit more=
27 S:   =I’m I’m interested in sports oh sports (1.3) and
28 some (you know the) (0.6) cooking (0.6) yeah I think
29 P:   um: [Okay
30 S:   [u:mm I think people can (1.0) improve their health
31  (0.5) ff- ah through (. ) doing sports (or) having a
32 good habi- ah good diet
33  (0.7)
34 P:   that’s true yes yes okay (. ) so you have an interest
35  (. ) in how people keep healthy as [well?
36 S:   [yeah yeah
37 P:   oh that’s good (. ) okay (1.0) sure (. ) a:m have have
38 [you contacted anybody else?
39  S:   [so
40  S:   no
41 P:   no oh [okay
42 S:   [( )] just you yeah
43 P:   ah: thank you for approaching me
44  (0.7)
45  S:   so will (you) join me?
46  P:   am:: (. ) that sounds pretty good (0.5) a:m (1.8)
47 yeah I think you ha- you share similar interests to me=
48  S:    =a:m
49 P:   so (. ) am: we should be able to work well together=
50 S:   =so: (0.4) do you know something about the: do you know
51 some (. ) detail about the: the presentation?
52 P:   am:: (1.0) yes yes sure a:: I think we got some
53 notes in the last lecture [so

165
After explaining the student’s (test taker’s) situation that he needs to do a presentation (lines 8-9) with someone in the class, the test taker initiates an action to express his desire, I want a (.) partner (lines 9-10), followed by a requesting act, so: can you be- ah can work with me? (line 14), which is incompletely structured. The other attempt to convince the classmate is seen in line 47 (so will (you) join me?). The test taker inserts this line without responding to the interlocutor’s thanks, thank you for approaching me (line 45). Toward the closing of the conversation, the test taker takes few actions to convince the classmate. The test taker’s intention to question the interlocutor (so: (0.4) do you know something about the: do you know some (.) detail about the: the presentation?) is unclear in lines 52-53, although this part is not related to his language use for mitigation. This performance provides little evidence of the test taker’s ability to exercise a range of mitigation devices and clear evidence of his insufficient ability to handle the linguistic tools in extended discourse.

4.3.2.5 Engagement in Interaction

Engagement in Interaction is one of the two criteria exclusively employed for
dialogue tasks. This criterion and the sub-features were partly informed by and
developed from Youn’s (2013, 2015) Engaging with Interaction, which scores test
takers’ interactional performance by evaluating the extent to which test takers
understand the ongoing context constructed by two speakers. As reported in Youn
(2013, 2015), in dialogue contexts in the current study, the test takers’ recognition of the
ongoing context and of the shared understanding with the interlocutor were reflected in
the test takers’ responses including clarification questions and acknowledgement tokens.
Viewing these features as evidence of developed competence for engagement in
interaction is also in line with the previous empirical studies, which showed that more
competent speakers are more likely to be able to recognize the ongoing context (Ishida,

Speakers’ responses would provide evidence of recognition of the discourse
context and more active engagement in interaction to evaluate. Integrating these
features, the current study re-operationalized the competence for interaction by adding
two features to separate the current test takers in the current criterion (Engagement in
Interaction). These features are: quality of responses to the interlocutor to distinguish
pragmatically highly competent test takers from others, and effect on interlocutor (Al-
Gahtani & Roever, 2012, 2013) to distinguish the least competent test takers from
others. It was found in the discourse data in the current study that understanding the
interlocutor’s responses and engaging in conversation by responding to the
interlocutor’s part is not always strongly discriminating, as the majority of the test
takers were able to do this. What was found to be more discriminating the highly
competent test takers from others is the quality of responses including their personal
comments and evaluating feedback to the interlocutor. The discourse data in the current
study provided evidence of the speaker’s shared understanding with the interlocutor, as
identified by Youn (2013). In addition, competent test takers’ response patterns observed in the current study offered even stronger evidence of their active engagement in interactions. At the same time, although not very often observed, effect on the interlocutor (Al-Gahtani & Roever, 2012, 2013) or scaffolding from the interlocutor discriminated the least competent test takers from the others.

Effect on the interlocutor is seen in performances where a speaker’s lack of competence or the speaker’s insufficient understanding of the given task situation forces the interlocutor to do what the speaker is supposed to do (e.g., explanation of the speaker’s situation). As to the performances of the least competent test takers, the engagement in interaction could not be realized without scaffolding from the interlocutor.

**Score 4 Level of Engagement in Interaction**

Performances for score 4 showed strong evidence of their ability to engage in interaction as described above. The descriptor for score 4 states as follows:

A next turn clearly shows understanding of a previous turn response to the interlocutor. Responses to the interlocutor such as clarification backchannel, acknowledgement tokens and questions are well-tailored for the ongoing context throughout the interaction.

The following Excerpt 23 (a part of Excerpt 19) provides a task performance which shows the test taker’s clear understanding of the interlocutor’s literal and implied meanings. Test takers whose performances deserve score 4 were able to sustain these in every single instance throughout from the opening to the closing of the performance
along with responses explicitly tailored for the ongoing context. The performance in this excerpt is on a dialogue task in which a student (test taker) is addressing the classmate working on the same project for a class presentation. The test taker understands the two speakers’ shared meaning in a conversation from the opening to the closing, as well as the test taker’s understanding of the previous turn in each line. This in turn, implies that the test taker understands the ongoing context and the classmate’s feelings and responds appropriately. Excerpts 23-25 provide a whole performance of a test taker, which are divided into the three excerpts for readability of the thesis. Excerpt 23 and 24 are parts of Excerpt 19 reviewed for Language Use for Mitigation in the previous section.

Excerpt 23 (Task ID: D4, Peer student topic, Dialogue)

S=Student (Test taker), P=Peer student (Interlocutor)

1 S: hi Sharon((la[ugh]))
2 P: [hi ((first name)) how are you?]
3 S: good how are you?
4 (.)
5 P: great
6 (.)
7 S: great?=
8 P: =we’ve=
9 S: =that’s good to he[ar ((laughs))]
10 P: [finished data analysis so: (0.4) I’m really (0.7) I’m so happy [((laughs))]
11 S: [yeah am: I actually have (0.5) I don’t know if it’s good news or not (.) but=]
12 P: =[“oh”]
13 S: =[I know we did all data analysis and (0.3) kind of started with our conclusion]
14 P: um:
15 S: but I just found out that we could get six more:
16 actually the male students in survey (0.7) a:[nd]
17 P: [oh:]
18 S: I know it’s gonna be additional data but I see that we only have (.). two male and ah=
19 P: =yeah: that’s true:
20 S: there is a >bit of< bias in that (.). that was unrelated to our "conclusions" (0.5)
21 I: that’s “true”
22 S: and: I know >it’s gonna< take more time but I think by getting six more male a:m students results (.). I think (.). we could do more better am: (.). presentation and I
In the opening (line 1), the test taker initiates the conversation by exchanging greetings. After the interlocutor has responded, great (line 5), to the test taker’s question, the test taker responds with evaluative feedback, great? (line 7), followed by that’s good to hear ((laughs)) (line 9), a response which is more complete than simply saying “good” or “great.”

In the subsequent exchanges and in lines 8 and 10-11, in particular, we’ve finished data analysis so: (0.4) I’m really (0.7) I’m so happy ((laughs)), the interlocutor as the classmate expresses the view that the collaborative research is almost finished, even if the quality of the research is not altogether satisfactory due to the gender imbalance of the participants. This relief serves also to reveal the classmate’s unwillingness to comply with the test taker’s proposal. (see APPENDIX C for the detailed instructions to test takers and to the interlocutor respectively). The purpose of the test taker’s next turns (lines 12-13, yeah am: I actually have (0.5) I don’t know if it’s good news or not) seems to project the test taker’s upcoming proposal of revising the almost completed research.

This test taker’s utterance, yeah am: I actually have (0.5) I don’t know if it’s good news or not, shows her understanding of the classmate’s relief that the collaborative project is progressing well (as reflected in the test taker’s utterance in lines 15-16, I know we did all data analysis and (0.3) kind of started with our conclusion). At the same time, it shows the test taker’s understanding that the test taker’s upcoming proposal and request may cause extra and undesirable workload for the classmate, who is satisfied with the current state of the
project. The test taker’s acknowledgement of the workload for the classmate is in fact implied in line 21, I know it’s gonna be additional data, and in line 29, and:

I know >it’s gonna< take more time.

In the middle part of this conversation (Excerpt 24 below), the test taker understands the interlocutor’s (the classmate’s) feeling of unwillingness to take an action to revise the project. This is clearly reflected in the test takers’ response, “Yeah”: sorry about the last minute (.). yeah notice, in lines 49-50, and in the subsequent part, but am: I’ll try and do as much as I can as well and if you could just guide me though the process of using the system (lines 50-51 and 53-54), which expresses the test takers’ motivation to minimize the workload on the interlocutor (the classmate).

**Excerpt 24 (Task ID: D4, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

44 P:    (((first name)) yeah: I’m really swamped
45 this week with my part time job and=
46 S:    =°yeah°
47 P    I have other assignments due next week so and I
48 feel really (.). stressed with tho[se but
49 S:    [°yeah°: sorry about
50 the last minute (.). yeah notice but am:
51 I’ll [try and do as much as I can as well=
52 P:         [()
53 S:    =and if you could just guide me though the process of
54 using the system
55 P:    oka[y
56 S:    [um I’m sure there’s (.). won’t be too much (.)
57 changes in the conclusion or data analysis itself
58 because the [software is gonna do it [so
59 I:    [mmmhm [right that’s true
60 S:    and we have four more days as well to
61 fin[sh the project
62 I:    [four more days yes
63 ((continues))

After the student (the test taker) and the classmate (the interlocutor) have reached an agreement on the test taker’s proposal (lines 64-65) in Excerpt 25 below, the
test taker continues to interact with the classmate instead of concluding the conversation with simple greetings. These actions include expressions of understanding of the interlocutor’s feeling, I totally understand (line 78), a gratitude for the interlocutor’s understanding and support, oh: thanks Sharon (line 81) and thanks for your help (line 86), further encouragement for each other as a group, ((laughs)) yeah it’ll be interesting I think yeah (lines 83-84), something which could also serve to make their decision and agreement solid.

Excerpt 25 (Task ID: D4, Peer student topic, Dialogue)

S=Student (Test taker), P=Peer student (Interlocutor)

64 P: in that case I can help you with the software
65 S: great (((laughs))))
66 P: (((laughs))))
67 S: yeah I’m sure it’s gonna make our presentation better
68 P: [m:]
69 S: [I know it’s]=
70 P: =ye[ah
71 S: [extra work but]=
72 P: =ye[ah
73 S: [I think it’ll get paid off *afterwards*
74 P: I think so=
75 S: =ye[ah
76 P: [you’re right (.) I know I shouldn’t be complaining
77 S: so much=
78 P: =no ((laughs)) I I totally under[stand]
79 P: (((laughs))) all right
80 S: thanks ((first name)) for talking with me
81 P: oh:thanks Sha[ron
82 S: [finding those people (.) [that’s’s amazing
83 S: (((laughs))) yeah
84 P: it’ll be interesting I think [yeah
85 S: [u:m I think so yeah
86 P: thanks for your help
87 P: all right thank you: we’ll talk we’ll talk [soon
88 S: [yeah
89 S: I’ll get back when I get the data
90 P: cool thank you
91 (.)
92 S: thank [you
93 P: [all right see ya

In addition, as a response to the interlocutor’s turn, thank you: we’ll talk
we’ll talk soon (line 87), the test taker provides the concrete action to take next for their mutually decided work, Yeah I’ll get back when I get the data (lines 88-89), instead of simple and general answers, such as “yes” or “talk soon.”

**Score 3 Level of Engagement in Interaction**

Score 3 was awarded to performances that showed adequate understanding of the previous exchanges although the test takers’ feedback was relatively simple, compared to those identified in the highest level of interaction. The descriptor for score 3 states as follows:

A next turn shows understanding of a previous turn in most parts of the interaction although the responses to the interlocutor (e.g., clarification backchannel, acknowledgement tokens, questions) are generally simple in most part of the interaction.

It was found in the discourse data that the majority of test takers were able to recognize the context described in the task prompt as well as the ongoing context in the dialogue tasks. However, not all of the test takers engaged very actively in interactions using only simple responses such as “thanks,” “yeah,” and/or by a simple repetition of what the interlocutor said in the previous turn. Although the positive evidence was not as strong as that seen in the highest level of performance (as reviewed in Excerpts 23-25 above), the performance deserving score 3 maintained the quality of interaction in the view of the speakers’ understanding of the meaning and intention of each other’s utterances.

Test takers at score level 3 were able to understand the discourse context,
including the literal meaning and implied inconvenience of the interlocutor, although their responses were relatively simpler than those at score 4. Excerpt 26 below shows simple responses to the interlocutor’s utterances. The task situation is the same for Excerpts 23-25 above.

**Excerpt 26 (Task ID: D4, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

1 S: hi Sharon
2 P: hi (((first name))
3 S: [ah wow’s it going?
4 P: really good [I’m excited we’ve almost finished our:=
5 S: [((laughs))
6 P: presentation=
7 S: =ah: right right right
8 P: yeah:
9 S: um: (0.8) I just want to talk about the presentation?
10 P: mmhm
11 ((continues))
12 S: so am how about (1.0) yeah ah: giving questionnaire sheet (.) to these six guys and (0.5)
13 P: ah:=
14 S: =ah analyze again?
15 (0.5)
16 P: "wow" oh gosh a:m (0.7) I wish I could say yes
17 (0.6)
18 S: "yeah" 
19 P: [right now like I: feel I’ve got so many am things to do between now and then
20 S: now
21 (0.5)
22 P: as well as my part time job
23 (.)
24 S: ah yeah I [know
25 P: [a:m
26 S: mm yeah >but< I need your help because I don't know how
to use software for analyzing and
27 P: "oh" right okay
28 (0.5)
29 (continues))
30

In the greeting part, the test taker opens the conversation in lines 1 and 3 (hi Sharon and ah wow’s it going?) and the interlocutor responds by expressing the relief as instructed in the interlocutor’s task prompt (APPENDIX D). The test taker’s response in line 7 (right right right) to the interlocutor’s part is simple although it
may have been possible that the test taker has recognized the interlocutor’s implied unwillingness, as evident in the subsequent turns. The same tendency is seen after the test taker proposed re-reorganizing the presentation (lines 12-13), “yeah” (line 19) as a response to the interlocutor’s turn, I wish I could say yes (line 17). Also, when the interlocutor implies her inconvenience, right now like I: feel I’ve got so many am things to do between now and then (lines 20-21) followed by as well as my part time job (line 24), the test taker responds with, ah yeah I know (line 26), while in the same situation, the test taker at score 4 (Excerpts 23-25 above) explicitly expresses the sympathy, as in “yeah: sorry about the last minute (. ) yeah notice (lines 49-50 in Excerpt 24). These two utterances (the one for score 4 level in Excerpts 23-25 reviewed earlier and the other for score 3 in the current excerpts) are different in quantity. To evaluate them more qualitatively, ah yeah I know (the current test taker) is fairly general and can be used for multiple situations to show the speaker’s understanding, whereas sorry about the last minute (. ) yeah notice was explicitly tailored for the interlocutor’s utterance in the previous turn. Overall, simple responses are also repeatedly in the closing part of this interaction (Excerpt 27 below).

**Excerpt 27 (Task ID: D4, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

<table>
<thead>
<tr>
<th>Line</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 S:</td>
<td>okay (. ) so please um (2.3) let me know how to (0.3)</td>
</tr>
<tr>
<td>33</td>
<td>use our software an:</td>
</tr>
<tr>
<td>34</td>
<td>(0.3)</td>
</tr>
<tr>
<td>35 P:</td>
<td>yeah yeah I can [definitely help with that part [of it</td>
</tr>
<tr>
<td>36 S:</td>
<td>[Yeah yeah yeah [thank</td>
</tr>
<tr>
<td>37</td>
<td>you thank you</td>
</tr>
<tr>
<td>38 P:</td>
<td>yeah I’m really (. ) familiar with it [so:</td>
</tr>
<tr>
<td>39 S:</td>
<td>[“mhm”</td>
</tr>
<tr>
<td>40 P:</td>
<td>I’m quite happy to do that shouldn’t take so much</td>
</tr>
<tr>
<td>41</td>
<td>time (0.3) [“I think*</td>
</tr>
<tr>
<td>42 S:</td>
<td>[thank you (well) then (. ) I just try (. )</td>
</tr>
<tr>
<td>43</td>
<td>yeah=</td>
</tr>
</tbody>
</table>
44 P: all right (0.3) thank you so much
45 S: thank you
46 P: yeah I know it’s really important to have the gender balance
48 S: mmhm
49 P: as well [so
50 S: [um um right
51 P: am: all right (1.0) let’s do it ((laughs))
52 S: [((laughs)) yeah
53 P: okay am:
54 S: that’s right yeah
55 P: all right will talk to you soon [(  )
56 S: [mmhm thank you thank
57 P: [you
58 P: [okay thanks

After the test taker has reached an agreement with the interlocutor, yeah yeah I can definitely help (line 35), the interlocutor produces motivational utterances in respect of the collaborative project, yeah I’m really (. ) familiar with it so: (line 38), yeah I know it’s really important to have the gender balance (lines 46-47) and all right (1.0) let’s do it ((laughs)) (line 51). The current test taker in this excerpt responds simply by repeating the same and simple words all the way through the speech, yeah yeah yeah (line 36), thank you thank you (lines 36-37), and um um right (line 50). His concise and general responses in the interaction do not suggest that he does not follow the discourse context but are very different from the equivalent part for the score 4 in Excerpts 25, which were concrete and explicitly tailored for the discourse context. In Excerpts 25 (for score 4), the test taker interacted with the interlocutor in the same manner by mutually motivating each other in respect of the collaborative project, as reflected in yeah it’ll be interesting I think yeah (lines 83-84) and yeah I’ll get back when I get the data (lines 88-89). However, such strong evidence of engagement in interaction was not seen in performances equivalent to score 3, which is why score 3 is distinct from score 4.
Score 2 Level of Engagement in Interaction

Performances for score 2 were characterized by performances containing generally small responses and also by the fact that the shared understanding was undermined by absence of a response and/or turns that do not show the speaker’s understanding of the previous turn(s). Performance for score 2 were undermined by features such as responses to the interlocutor being generally small and by being at least once absent and/or by showing insufficient understanding of the ongoing context as reflected in their utterance. The descriptor for score 2 states as followings:

The test taker’s response is generally small. Response to the interlocutor (e.g., clarification backchannel, acknowledgement tokens, questions) is absent at least once or the test taker produces a turn that does not show understanding of a previous turn at least once.

Excerpt 28 below demonstrates turns showing the test taker’s insufficient understanding of the ongoing context. In this situation, a student (test taker) is addressing a receptionist at an administration office (interlocutor) seeking advice or support for a problem that the student has about changing his class. Excerpt 28 presents the post-opening part of the conversation, where the test taker has already explained to the interlocutor (the administrator) about the student’s problem. The interlocutor, as instructed in the task prompt, responds by providing the test taker with the advice (lines 1-17).

Excerpt 28 (Task ID: D2, Administrator topic, Dialogue)
S=Student (Test taker), A=Administrator (Interlocutor)

1  A: look there’s not- I can’t actually do anything myself I
know today is the cut off they have been a:m >look<
they have been in the past last year something similar
happened and my supervisor a:m signed form on it there
was a conditional (0.4) that you could (0.8) you could
a:m you could make the change and >we have to get<
email confirmation from the other lecture later. so
it’s a- I’m not sure the details my supervisor my
manager arranged it a:m and there’s one of it’s kind of
rare they do that (1.0) a:m he’s ah: he’d have to
tsigh it off again if he he decides it’s (0.7) you can
go ahead with that a:m is he’s actually in the meeting
now til until six so he’s not here right now so a:m
(0.7) ah:: just you’ve left a little bit late we are
running only like fifteen minutes we are closing at
five thirty so a:m (1.7) that’s all I can all I can
see at the moment sorry=
S: it’s (2.3) it’s okay it’s okay am: so am maybe I will
ah send ah: email again or
A: to to the lecturer or...
(0.8)
S: u:m (1.0) actually I want to find ah: directly (wait)
to meet the teacher or:
A: to meet the teacher?= S: =yeah to meet the [teacher:="
A: [yeah yeah
talk talk to the situation face to face face to face
ah: maybe so quickly
A: yeah you do have fifteen minute to try to arrange if
there’s five thirty though mostly five fifteen now
mostly the staff left campus >but you could< try to
arrange that if you could see them if they’re on campus
now >that’ll make it< easier
(0.5)
S: okay ok[ay]
(continues)

A: okay a::m (.) well look I tell you what I’ll I’ll a:m
(1.0) if ah: >if you wanna< come back like in thirty
min I’ll hang around for another fifteen minutes so
(0.6) it’s five fifteen now come back at quarter to six
likely they might come out around that time I’ll just
wait around I was gonna leaving at five thirty >but
I’ll Wait< around to quarter to six you are okay with
it? (0.6) in that in that half hour you can try (0.3)
to find a lecturer if you can (0.6) get their phone
number
S: ah: find the lecturer: yeah yeah maybe maybe it’s good
ah: but but do you think the lecturer have the all the
teachers’ (telephone) number?
(1.2)
A: the lecturer would have the:

After listening to the interlocutor’s advice, the test taker responds with a line,
it’s (2.3) it’s okay it’s okay am: so am maybe I will ah send ah:
email again or (lines 18-19) without further negotiating with or questioning the interlocutor to elicit further advice that could lead the test taker to the best possible option. This utterance indicates that the student’s action would have led to failure in the real situation. The task prompt states that the registration for class change needs to be completed by five thirty. This means that the student needs to obtain the teacher’s signature on the application form by that time, which would be impossible under the scenario’s circumstances. It is not clear whether the test taker’s utterance, it’s (2.3) it’s okay it’s okay (line 18) represents the test taker’s understanding of the interlocutor’s previous turn. The meaning of the test taker’s utterance, it’s okay, is unclear as to what is okay and the other part, so am maybe I will ah send ah: email again or, could be provided without understanding the intention of the interlocutor’s previous turn.

Regardless of whether the test taker understands the interlocutor’s turn, the test taker’s unclear utterance (lines 18-19) causes confusion for the interlocutor, as reflected in the interlocutor’s utterance, to to the lecturer or..? (line 20). In the test taker’s subsequent line, u:m (1.0) actually I want to find ah: directly (wait) to meet the teacher or: (lines 22-23), the test taker abruptly cancels what he has just said (what he was originally going to do) without responding to the interlocutor’s utterance, to to the lecturer or..? (line 20) and to his previous turn (the cancelled action). These parts do not necessarily show that the test taker understands the ongoing situation. At least they show that the shared understanding with the interlocutor is not achieved as reflected in the second confusion of the interlocutor, to meet the teacher? (line 24). The test taker’s next line, yeah to meet the [teacher:= =talk talk to the situation face to face face to face ah: maybe so quickly (lines 25 and 27-28) then shows that the test taker seems to
understand the interlocutor’s question, to meet the teacher? (line 24), but this proposed action of the test taker would not be realized in the real situation as implied by the interlocutor’s comment in the subsequent lines (lines 29-33).

The test taker’s utterances with his unclear intentions eventually elicits advice and support from the interlocutor in lines 37-46 (okay a::m...their phone number). However, the test taker’s next line, ah: find the lecturer: yeah yeah maybe maybe it’s good ah: but but do you think the lecturer have the all the teachers’ (telephone) number? (lines 47-49) indicates the test taker’s insufficient understanding of the interlocutor’s previous turn(s). Although the test taker implies his understanding, yeah yeah maybe maybe it’s good, the remaining part (but but do you think the lecturer have the all the teachers’ (telephone) number?) clearly demonstrates that the test taker does not understand the interlocutor’s previous turn. In fact, the test taker’s response (lines 47-49) does not make sense to the interlocutor, who resorts to confirming with the test taker, The lecturer would have the: (line 51). Even though the test taker’s language outputs may make the test taker appear to be engaged in the interaction with the interlocutor, the shared understanding is undermined.

**Score 1 Level of Engagement in Interaction**

In the case of performances given score 1, engagement in interaction was undermined as the negative features characterizing score 2 were more noticeable and/or a noticeable support from interlocutor was reflected in the performance. The descriptor for score 1 states as follows:

The negative features described characterizing score 2 are more noticeable
and/or,

the support from the interlocutor is noticeable (the test taker’s engagement in interaction is not active, which occasionally prompts the interlocutor to speak more or to initiate what the speaker is expected to say and/or to help the test taker to construct the conversation).

Excerpt 29 below illustrates a conversation which shows the support from the interlocutor afforded to the test taker. This excerpt is a part of Excerpt 22 reviewed for Language Use for Mitigation. The test taker in these excerpts showed a very passive engagement in interaction.

**Excerpt 29 (Task ID: D3, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

26 P:    yeah
27 S:    I [I’m inter-
28 P:    [tell me bit more=
29 S:    =I’m I’m interested in sports oh sports (1.3) and
30 some (you know the) (0.6) cooking (0.6) yeah I think
31 P:    um: [okay
32 S:    [u:m I think people can (1.0) improve their health
33 (0.5) ff- ah through (. ) doing sports (or) having a
34    good habi- ah good diet

47 S:    so will (you) join me?
48 P:    am:: (. ) okay that sounds pretty good (0.5) a:m (1.8)
49    yeah I think you ha- you share similar interests to me=
50 S:    =u:m

61 P:    [yeah
62    (1.7)
63 P:    um: yeah (. ) so am:: so long as you share the same
64    interest I think (0.5) we could work well together
65 S:    okay
66 P:    yeah:
67 S:    good
68 P:    ((laughs)) (1.4) great all right (1.0) am:: (0.7) so I
69    guess we can exchange details (0.6) la[ter
70 S:    [yeah yeah okay=
71 P:    =yeah okay thank you:
After the test taker has briefly explained his situation to the classmate (the interlocutor), he leaves his work to the interlocutor, who is forced to do what the test taker is expected to do. For example, as a response to the interlocutor’s utterance in lines 26 and 28 (yeah and tell me bit more), the test taker initiates to explain his interests (lines 29-30). In the later line, the interlocutor also utters, share the same interest (lines 63-64). These parts are what the test taker himself is supposed to deliver and to have stressed instead of making the interlocutor do so.

In addition to the effect on the interlocutor (Roever, 2011), the test taker’s response to the interlocutor is very simple. In particular, the test taker’s responses in line 65 (okay), in line 67 (good), in line 70 (yeah yeah okay) are evidently different from the pragmatically highly competent test taker in Excerpt 2 (thanks yeah it was good running into you ((laughs))). The test taker in the current excerpt does not even express his thanks to the interlocutor.

Effect on interlocutor (Roever, 2011) was also evident in performances of the two pilot test takers whose English proficiency levels were far below university-entry level. They showed even greater effect on interlocutor as exemplified in Excerpt 30 below. In this performance, the engagement in interaction could not be sustained without support from the interlocutor who performed work which was supposed to be performed by the test taker (e.g., explaining the situation) during the interaction. This task situation is the same as for Excerpts 1 and 4, where a student is addressing Professor Smith at the beginning of a semester for a class change. The following excerpt provides the middle part of the conversation.

**Excerpt 30 (Task D1, Professor topic, Dialogue)**

S=Student (Test taker), P=Professor (Interlocutor)

1  P: okay
The test taker’s utterance, a::nd (0.4) um:: (2.2) yeah (2.0) a::nd (3.5) m::: (line 2) shows that the test taker (as a student) has difficulty in continuing the negotiation with the interlocutor (as a professor) and forces the interlocutor to break with protocol (Al-Gahtani & Roever, 2012) and initiate what the test taker is supposed to explain to the interlocutor in lines 3, 5, and 7. The test taker’s subsequent utterance, okay (0.3) >so is that< finished? (line 8) shows that the test taker does not follow what is happening. Since what the interlocutor informed the test taker about (lines 3, 5 and 7) was all stated in the test taker’s task prompt, it is possible that the test taker initially did not understand the task situation adequately. The test taker’s inadequate understanding of the interlocutor’s utterance is also implied in the subsequent part of the test taker’s utterance, so (0.4) >so< from tomorrow I can (0.8) (line 11). Given the small amount of output in each of the two lines (lines 8 and 11) that the test taker produced, it may be equally possible that the test taker failed to produce outputs even if she grasped the meaning of the interlocutor’s turns. Regardless of the reasons, the test taker’s performance has required the second scaffolding from the interlocutor as in you can’t yeah after tomorrow you can’t change classes (line 12) and so you have to get the (0.5) the registration form singed
(lines 14-15), both of which are what the test taker is supposed to explain to the professor (the interlocutor). Line 17 indicates the closing of the conversation, which is closed without the test taker’s expected thanks to the interlocutor.

4.3.2.6 Turn Organization

Youn (2013) reported features to construct turn organization as relatively less demanding than language use to handle pragmatic expressions and linguistic resources. As predicted by Youn (2013), the test takers in the current study were capable enough to take turns generally naturally in one or more turns in a conversation. What was differentiating the test takers was whether pragmatically competent test takers were able to take turn without awkward delay and pauses and/or awkwardly overlapping an interlocutor’s utterance throughout the conversation from the opening to the closing. The most competent test takers in the current study received, kept and released a turn naturally, which allowed the interlocutor to judge when to initiate a turn. This criterion is one of the other assessment criteria used for dialogue tasks only.

Score 4 Level of Turn Organization

Turn taking behaviour of pragmatically highly competent test takers caused little confusion to the interlocutor. The descriptors for score 4 states as follows:

Smooth natural turn-taking throughout the conversation as:
- adjacency pairs are completed (e.g., question & answer, request & thanks) without awkward pauses and,
- the speaker (test taker) releases, receives, and keeps conversation flowing in a way that will not confuse the interlocutor (as to whether the interlocutor should
keep listening or take a turn to facilitate the speaker’s speech).

Even if turns overlap, the speakers are not confused.

Excerpt 31 below is a performance deserving score 4. This excerpt is a part of Excerpt 23 evaluated for Engagement in Interaction. The turns are taken without any noticeable pauses and transition of conversation flow (Ten Have, 2007) between the two speakers is smooth even though some parts of utterances are partially overlapped. Even if an overlap happened, two speakers mutually understand which of them is to take over the conversation flow and to continue the speech without being confused. Pragmatically competent test takers interacted with the interlocutor in ways that the interlocutor was able to recognize when the test taker’s current turn was to finish and when to take over the conversation flow. They also kept themselves ready to release the turn to the interlocutor whenever the interlocutor attempted to initiate a turn before they complete their turn.

**Excerpt 31 (Task ID: D4, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

1  S: hi Sharon((laughs))
2  P: [hi {(first name}) how are you?  
3  S: good how are you?  
4  .)
5  P: great  
6  .)
7  S: great?=  
8  P: =we’ve=  
9  S: =that’s good to he[ar {{laughs}})  
10 P: [finished data analysis so: (0.4) I’m really (0.7) I’m so happy {{(laughs)}}  
11 S: [yeah am: I actually have  
12 (0.5) I don’t know if it’s good news or not (.) but=  
13 P: =["oh"
15 S: =[I know we did all data analysis and (0.3) kind of started with our conclusion  
16 17 P: um:  
18 S: but I just found out that we could get six more:  
19 actually the male students in survey (0.7) a:[nd
20 P: [oh:]!
21 S: I know it’s gonna be additional data but I see that we
22 only have (.) two male and ah=
23 P: =yeah: that’s true:
24 S: there is a >bit of< bias in that (..) that was unrelated
25 to our “conclusions”

No awkward pauses nor the speakers’ confusion of which speaker should take over the turn are seen. In the introductory part, the speakers exchanges the greeting in lines 1-9 with short sentences without being overlapped, showing that both speakers allows each other to complete their utterance in the given turn. In lines 8 and 10-11, the interlocutor (as the classmate) starts to express her relief, we’ve finished data analysis so: (0.4) I’m really (0.7) I’m so happy ((laughs)). The test taker (as the student) listens to and receives the interlocutor’s expressed feeling until the interlocutor completes the interlocutor’s turn. After listening to and receiving the interlocutor’s feeling, the test taker initiates a line to project the upcoming proposal and request, Yeah am: I actually have (0.5) I don’t know if it’s good news or not (lines 12-13). A little after that, the interlocutor expresses her surprise, "Oh" (line 14) in a very low voice, which may be recognized by the test taker. There is a slight pause between the test taker’s utterance (line 13) and the surprise, "Oh" (line 14), of the interlocutor, which in turn implies the test taker is aware that the interlocutor may take over the conversation flow and is ready to release the turn when it is cut in by the interlocutor. In fact, in lines 15-25, no obvious overlap was seen. The test taker’s turn, but I know we did all data analysis and (0.3) kind of started with our conclusion (line 13 and 15-16) is followed by the interlocutor’s reaction, Um: (line 17). Then the test taker resumes her utterance, but I just found out that we could get six more: actually the male students in survey (0.7) a:nd (lines 18-19). The explicit completion of the interlocutor’s part, Um: (line 17) may indicate that test taker consciously or unconsciously leaves some space where the
interlocutor could have taken the conversation flow. The test taker resumes her utterance (line 18) eventually because the interlocutor simply responds, Um: (line 17) and does not initiate her turn.

The test taker’s awareness of maintaining or releasing her turn is further confirmed in lines 18-20. The test taker puts a more explicit but not awkwardly prolonged nor confusing pause at the last of line 19, in survey (0.7) a:nd, when the interlocutor could have initiated his or her turn. The interlocutor expresses a surprise, Oh:↑ (line 20), which is not overlapped by the test taker’s resumed utterance, I know it’s gonna be additional data but I see that we only have (. ) two male and ah (lines 21-22). This could confirm the test taker’s readiness for being cut in her own turn by the interlocutor and for receiving or initiating a turn. The last five lines (lines 21-25) provide more explicit evidence of the test taker’s readiness for turn transition. The test taker’s utterance, I know it’s gonna be additional data but I see that we only have (. ) two male and ah (lines 21-22) was resumed in line 24, there is a >bit of< bias in that (. ) that was unrelated to our °conclusions°, with the interlocutor’s turn, Yeah: that’s true: (line 23) inserts between them. This interlocutor’s turn has no gap with the test taker’s proceeding turn in lines 21-22, which clearly shows that the test taker has stopped speaking (at the end of line 22) and released her turn to the interlocutor in order to switch to becoming the listener. The test taker’s utterance is resumed in line 24 without any awkward pause between this utterance and the interlocutor’s previous line, but this inserted line of the interlocutor could represent the test taker’s awareness of the point when a turn transition happens and of when to initiate her own turn, which leaves space open for the interlocutor to initiate or take over the turn.
Score 3 Level of Turn Organization

Performance deserving score 3 generally showed smooth turn taking.

Performances of test takers at this level were similar to those at score 4 in most parts of a conversation. The descriptors for score 3 states as follows:

- Generally smooth turn-taking although the performance may contain parts where
  - adjacency pairs are not completed (e.g., question & answer, request & thanks)
  - and/or,
  - the test taker’s turn is a bit delayed.

Test takers’ ways of turn organization for score 3 was not as consistently positive as those at score 4, which showed strong evidence of the positive features discussed above. Excerpts 32 and 33 below demonstrate turn organization patterns which were very rarely seen in performances at score 4.

Excerpt 32 below shows an example of possible turn delay. The tasks situation is the same as Excerpt 28, where a student (test taker) is addressing the administrator (interlocutor) to seek assistance for the student’s class change. The excerpt provides the middle of the conversation.

Excerpt 32 (Task ID: D2, Administrator topic, Dialogue)  
S=Student (Test taker), A=Administrator (Interlocutor)

1 A: so you have any other ideas of (. ) what we might be able to do for it?
2 1.0
3 S: maybe I can phone her again or you can ask me for her (2.0)
4 1.0
5 A: a:m (0.6) you could try phoning her again but we’ve we are really running it’s five thir:ty now isn’t it?
6 0.3 5 five fifteen okay=
7 S: =[ah↑
8 10 A: =[probably they’re gonna be gone it’s after hours
9 already

11
In line 1, the interlocutor asks a question, so you have any other ideas of (. ) what we might be able to do for it?, which shows that the interlocutor explicitly has moved the conversation flow to the test taker. Then a pause is laid before the test taker initiates the utterance, maybe I can phone her again or you can ask me for her (line 2). In addition, when the interlocutor is offering help in lines 14-18, (ah: look it’s the possibility you can wait for (0.5) ah: I can you know hang I could hang around here if the meeting is finishes early (0.6) a:m yeah the possibility >we could< we could sort that out if you want to wait here), another pause is laid before the test taker initiates the next turn in line 21 (so (0.4) maybe I just (0.3) can wait here). The test taker’s turn (line 21) is taken with delay which consequently generates the interlocutor’s part at the end of line 20 (a:m). This part, (line 20, a:m) may show the interlocutor’s expectation for the transition of the conversation flow to the test taker because the interlocutor as the administrator has told the test taker (the student) all of what could be done for the test taker (the student). The test taker herself utters so (0.4) (line 21) instead of smoothly initiating the main part (maybe I just (0.3) can wait here). In the context equivalent to this, test takers whose turn organization performance deserved score 4 were very likely to express thanks to the interlocutor without laying a pause. It was not clear whether or not this pause was an indicator of the test taker’s insufficient understanding of the discourse context. However, this turn
organization was distinguishable from that seen in the highest level.

Except 33 below also provides an example of performance for score 3. In test takers’ performances deserving score 3, the boundary of which speaker should hold the conversation flow was not always clear, which made their performance no more than “generally smooth.” The task situation that this excerpt describes is the same for Excerpts 1 and 4, where a student is addressing Professor Smith about the class chance. The excerpt shows the middle of the conversation.

**Excerpt 33 (Task ID: D1, Professor topic, Dialogue)**

S=Student (Test taker), P=Professor (Interlocutor)

1 S: I’m taking your elective a (.) class?
2 P: right [yes
3 S: [ye: (.) so do you have time now?
4 P: ah: yes [I’ve got a couple of minutes
5 S: [ye ah okay
6 thank you because (1.0) actually I have problem?
7 P: ah really=
8 S: =ye:s I’m taking your class on (.). Wednesday
9 P: ah “yes that’s right”
10 S: [ye: (0.6) but (1.0) from week four I have to take compulsory units
11 (0.5)
12 P: oh I [see
13 S: [So I have to change- (0.6) your class so I need your signature
14 P: oh: I [see
15 S: [so ye:=
16 P: so you’ve got a timetable class happening
17 S: ye: timetable:
18 P: I [see
19 S: [I’m so panicking ye:((laughs))
20 P: oh: these [things happen
21 S: [ye: so- yeah so I have to- (0.5) ye- (0.6)
22 P: 23 change subject
24 S: right yeah
25 S: and then- (0.5) I have to change (.). by five thirty
26 tomorrow so I’m very rushing
27 P: ah ye[s yes the registration
28 S: [Ye: I’m so lucky to find you here
29 S: ye:s that’s good
30 P: ye:s that’s good

In this excerpt, overall conversation overlap is seen in multiple places. The test taker makes short and noticeable pauses. These pauses are identified in line 6 (because
(1.0) actually I have problem?), in line 14 (so I have to change- (0.6) your class) and in lines 23-24 (ye: so- yeah so I have to- (0.5) ye- (0.6) change subject), which are both in the middle of the sentences or a clause. The first one, because (1.0) actually I have problem?, may reflect the test taker’s attempt to conduct a repair.

This pattern rarely occurred in performances at score 4. In the performance in Excerpt 31 earlier (showing the performance for score 4), the test taker laid the short pauses at the places which he should clearly recognize as the end of the sentences or the clause. As discussed earlier, these places of pause appeared to help the interlocutor to understand whether the test taker should keep speaking or whether the interlocutor should take a turn or insert something. In contrast to this, the test taker in the current excerpt (showing an example of score 3), abruptly stops the speech (as in I have to- (0.5)) or puts the pauses in the middle of the sentences or the clauses in multiple places for unclear reasons. It is possible that these pauses have made it difficult for the interlocutor to judge what the test taker is going to do next. In addition, in line 28, the interlocutor’s utterance, ah yes yes the registration, is largely overlapped by I’m so lucky to find you here. Also the test taker does not give a comment back to the interlocutor as the test taker’s part, ye: I’m so lucky to find you here (line 29), overlapped with the interlocutor’s utterance (line 29). Performances for score 4 also contained test takers’ parts overlapping the interlocutor’s parts but not as noticeable as the test takers in the current excerpt (score 3). The test taker started speaking (line 29, ye: I’m so lucky to find you here) almost at the same time as the interlocutor’s part (line 28, ah ye[s yes the registration) and both utterances in parallel, where the test taker did not respond to (and possibly did not pay attention to) the interlocutor’s part.
The features contained in performances at score 3 were not as negative as those at score 2 (discussed next) and in large parts of a performance, turn taking was smooth. However, performances deserving score 3 did not show strong evidence to deserve score 4, which was characterized by completing adjacency pairs without awkward pauses, and releasing, receiving or keeping conversation flowing in a way that would not confuse the interlocutor.

*Score 2 Level of Turn Organization*

As discussed earlier, pragmatically competent test takers were more likely to release and receive conversation flow in a way that the interlocutor understands who should speak next and thus engage in the conversation smoothly. However, this was not the case for performances deserving score 2. The descriptor for score 2 states as follows:

The test taker puts awkward pause(s) between turns or confuses the interlocutor (as to whether the interlocutor should keep listening or take a turn to facilitate the speaker’s speech).

Excerpt 34 below illustrates the interactional behaviour leaving little space that does not allow the interlocutor to take a turn and which confuses the interlocutor of when to take his turn. The task situation is the same as the situation for Excerpts 28 and 32, where a student (test taker) is addressing an administrator (interlocutor) to seek assistance.
Excerpt 34 (Task ID: D2, Administrator topic, Dialogue)

S=Student (Test taker), A=Administrator (Interlocutor)

1  S:  hi hello I’m ((first name)) I’m a student of the uni
2      (0.4) of the university [and ah yeah and in the you=
3      [I see
4  S:  =know I I’m here to reason I’m here is because I need
5      ah professor’s signature who is not in the campus right
6      now=
7  A:  =[ah
8  S:  =[and (0.5) and to be more specific [you] know=
9  A:  [um]
10 S:  ='cause u:m I I want to exchange my (.). my class u:m
11  currentl from cla- electi- elective class a to (.). the b (and)
12  fortunately I already get one one professor
13  signature but the other (.). who is (.). off campus
14  currently I need his (1.3) signature
15 A:  right [( ]
16 S:  [yeah and today is the last day for my
17  apply[mation ((transcribed as pronounced))
18 A:  [yes that’s ri[ght
19 S:  [so I I I’m here I ( ) for (.)
20  ask some help if I can do if there another way to make
21  this happen and:

The test taker initiates the conversation with a greeting and by introducing
herself to the interlocutor, hi hello I’m ((first name)) I’m a student of
the uni (0.4) of the university (lines 1-2). The interlocutor inserts, I see
(line 3), while the test taker holds the conversation flow to deliver the explanation. The
interlocutor’s inserts an utterance, I see (line 3) is fully overlapped by the continued
utterance of the test taker, and Ah yeah and in the you know I I I’m here to
reason I’m here is because … (lines 2 and 4-6). This part indicates that the
interlocutor is entirely prevented from producing an utterance even a simple greeting
such as “Hello” and “What can I do for you?” This test taker does not let the
conversation flow move to the interlocutor as the interlocutor barely utters, ah (line 7)
and right ( ) (line 15), words which are fully or mostly overlapped by the test
taker’s utterance, indicating that the current test taker continues speaking. In the end, the
interlocutor is prevented from taking turns or from producing an utterance. Fluency is
lost and little sound variation is seen, which also makes it difficult for the interlocutor to
judge where to take a turn.

The test taker showed the same tendency of holding the conversation flow regardless of the task situation. Excerpt 35 below describes a situation where a student (test taker) is addressing the classmate working on the same project as for Excerpt 31 representing score 4.

Excerpt 35 (Task ID: D4, Peer student topic, Dialogue)
S=Student (Test taker), P=Peer student (Interlocutor)

1 S: hi Sharon [ye-
2 P: [hi [({first name})] how are you:? 
3 S: [yeah yeah 
4 fine you know we’ve been wor work about the our 
5 assessment for the: (0.5) ah: about the (.1) the the 
6 social issues (0.6) recently=
7 P: =[yeah 
8 S: =[and and I know we we do the questionary there’s a 
9 group=
10 P: =[umm 
11 S: =[we invite people more more girls and (0.4) u:m you-
12 I I’ve been worry about the (0.6) this because you know 
13 I think the gender should should (0.6) should should 
14 equal (0.6) both (0.4) women and men so I found more

After the greeting exchanged between the test taker and the interlocutor in lines 1-2, the interlocutor initiates another greeting phrase, how are you:? (line 2). The test taker responds to this, Yeah yeah fine you know...(lines 4-6). She answers the interlocutor’s question, but immediately continues explaining the encountered issue to the interlocutor, you know we’ve...(line 4) without asking the interlocutor back or allowing the interlocutor to speak. This was markedly different from the performance of the test taker in Excerpt 31(representing score 4) that asked the interlocutor back, Good how are you? (line 3 in Excerpt 31 above) and subsequently allowed the interlocutor to respond, great (line 5 in Excerpt 31) and to express a relief about the progress of their collaborative research, we’ve finished data analysis so: (0.4) I’m really (0.7) I’m so happy ((laughs)) (lines 8 and 10-11 in Excerpt 31).
By contrast, the current test taker in Excerpt 35 (score 2) does not release the conversation flow. The interlocutor inserts utterances twice, yeah (line 7) and umm (line 10), however they were fully overlapped by the test taker’s utterances (lines 8 and 11), which implies that the test taker does not intend to have the conversation flow move to the interlocutor. This relatively long utterance composed by lines 4-6, 8-9, and 11-14 combined with the ways of delivering the speech makes it difficult for the interlocutor to judge where the end of the sentence is, where a pause can be laid and thus when to cut in or take a turn in the test taker’s utterance.

Although the reason was different from the effect on interlocutor (Roever, 2011) discussed in Engagement in Interaction, the interactional behaviour in turn taking discussed in this section was also found to be affecting or changing the interlocutor’s ways of interaction. As discussed in the previous section (for Engagement in Interaction), due to the limited ability, the pragmatically least competent test taker forced the interlocutor to provide the considerable scaffolding or to do the work that the test taker was supposed to do. The turn organization revealed in Excerpts 34 and 35 above made it difficult for the interlocutor to interact and thus prevented the interlocutor from doing what the interlocutor was expected to do.

Score 1 Level of Turn Organization

As for the other criteria, score 1 was given because negative features were more noticeable than those in performances for score 2. The descriptor for score 1 states as follows:

- The negative features described characterizing 2 are more noticeable throughout the conversation and/or,
-some turns are noticeably delayed or awkwardly initiated without response to the interlocutor’s comment.

Excerpt 36 below describes a conversation with the classmate working on the same project as for Excerpt 31 and 35 in this section. In this excerpt, the test taker awkwardly attempted a turn without responding to comments, in addition to confusing the interlocutor of when to take (as for score 2).

**Excerpt 36 (Task D4, Peer student topic, Dialogue)**

S=Student (Test taker), P=Peer student (Interlocutor)

```
1 S: hello Sharon
2 P: hi ((first name)) how are you?
3 S: ah fine thanks how are you? (0.3)
4 5 P: _really good [we almost finished our assignment=
6 S: [ha(.)oka:y
7 P: =so[I’m excited-
8 S: =thank you for your help thank you join thank
9 you to be partner to (0.6) finish this presentation
10 but=
11 P: =[um
12 S: =[ah- actually I found (.) some problem about the
13 questionnaire (0.6) we are only two: male interviews so
14 so I need (0.5) maybe I need to interview more male
15 (0.7)
16 P: [all right
17 S: [more male to ah collect more data about this:(1.7) ah
18 so ah you know I: don’t ah: (1.2) good at analyze all
19 the information so(1.2) do- would you mind that us-
20 (0.5) that we: (;) separates work I interview the
21 person and you: (0.7) ah [analyze the data?
22 P: [”oh”
23 oh really?
```

After the greetings which show smooth and not overlapped turns between the speakers, the test taker responds, *fine thanks how are you?* (line 3) to the interlocutor’s previous utterance, *how are you?* (line 2). The next utterances (line 6 and 8) display the test taker’s abrupt initiation of her turns in the middle of the interlocutor’s utterance without responding to it. In line 6, the test taker inserts an
utterance, *oka:y*, which overlaps the most part of the interlocutor’s speech, we almost finished our assignment (line 5). The inserted, *oka:y* (line 6), might possibly indicate a response to the interlocutor’s utterance, *really good* (line 5). However, it equally makes sense to suggest that the test taker does not pay attention to the interlocutor’s part and that the inserted, *oka:y* (line 6), is a projection of the test taker’s upcoming speech. In fact, the meaning of *oka:y* (line 6) is ambiguous and not tailored as a response to the interlocutor’s part. The test taker’s ignorance of the interlocutor’s turn is further confirmed by the test taker’s following part, thank you for your help thank you join...but (lines 8-10) which explicitly overlaps the interlocutor’s continuing speech to express a feeling of excitement, *I’m excited* (line 7). Unlike the pragmatically competent test taker in Excerpt 31, the current test taker does not wait until the interlocutor completes or produced even a response to the interlocutor. Rather, the test taker awkwardly initiates her turn in the middle of the interlocutor’s speech without the interlocutor’s expressed feeling. In the remaining lines in this excerpt, the interlocutor is not given a conversation flow until line 23 (*oh really?*) and produces little utterances *um* (line 11), *all right* (line 16) and °oh° (line 22), with two utterances, *um* (line 11) and *all right* (line 16), fully overlapped by the test taker’s continuing speech.

The turn organization revealed in the test takers’ discourse data was also confirmed in the two interlocutors’ voluntary comments about interactional behaviours of particular test takers. According to the comments, these test takers were likely to (a) stick to written notes taken on the task prompt, (b) have scripts rehearsed in their memory before initiating the tasks and have difficulty when interaction is started, (c) make little eye contact even at the very beginning of the task, and (d) make themselves look afraid that the conversation flow moves to the interlocutor. One of the interlocutors
commented that therefore (d) the interactional behaviour of the test takers eventually made it difficult for the interlocutor to interact and to cut into the conversation. These comments were not elicited in a structured interview with the interlocutors as their comments were not originally expected. These comments thus may not be adequate enough to be an independent source of data. However, they served to confirm the findings from the discourse data.

4.3.3 Do the Rating Criteria Provide a Stable Assessment of L2 Pragmatics Defined in the Study?

A multi-faceted Rasch analysis based on all of the collected data was conducted to provide a summary of four facets (test taker, rater, rating criterion, and task) specified in the analysis. Figure 4.1 below represents the FACETS ruler map, which locates four facets in a single frame of reference. This map also allows for comparisons to be made between and within the facets and facilitates interpretation of the results.
Figure 4.1. FACETS ruler map.
S.1= Social Actions to Achieve the Communicative Goal; S.2=Facility with the Language; S.3=Language Use to Deliver the Intended Meanings; S.4=Language Use for Mitigation; S.5=Engagement in Interaction; S.6=Turn Organization.
The first column indicates the logit which locates the test takers, the raters, the tasks and the rating criteria on the same scale expressed in logit, which is used to indicate the ability of the test takers, the degree of severity the raters in rating test takers’ performances, the difficulty of the tasks and the difficulty of the rating criteria. The second column identifies the test takers, each of whom is indicated as an asterisk. The higher the test taker’s ability, the higher in the column he or she is located on the logit scale. The third column shows the raters, indicated by a number (1, 2, and 3), and these are located according to their level of severity in the rating. The raters located higher in this column were estimated to be more severe in rating the test takers’ performances. Since the rater facet was specified as the non-centered facet, its logit mean score was not set as zero. In the fourth column, the test tasks (the role plays) are plotted according to their estimated difficulty. More difficult test tasks appear higher while the less difficult ones appear lower. The fifth column displays the rating criteria. The more difficult criteria (thus more challenging for the test takers) of them are plotted higher than those less difficult (thus less challenging) in the column.

The sixth to eleventh columns show how the assessment band levels were used for each criterion, from left to right, Social Actions to Achieve the Communicative Goal (S.1), Facility with the Language (S.2), Language Use to Deliver the Intended Meanings (S.3), Language Use for Mitigation (S.4), Engagement in Interaction (S.5), and Turn Organization (S.6). The horizontal line in each column indicates a 50% of likelihood that either of two level bands is awarded.

Under the assessment that the current study designed and administered, the total of 67 test takers were widely separated as visually presented, which was also indicated

---

3 FACETS advises users to define one non-centered facet. Since the two groups of test takers on task set 1 and those on set 2 were group-anchored (as described in section 3.6.3 in chapter 3), the test taker facet cannot be non-centered. Thus, the rater facet was selected as the non-centered facet for the analysis (M. Linacre, personal communication, September 10, 2015).
by the separation index (10.36) with its reliability (0.99) as in Table 4.5 below, which summarizes the extent to which the elements in each facet were separated. The separation index refers to how many statistically distinct levels test takers are separated into by the data. The separation index and the reliability index indicate that the test takers were fairly reliably separated into over 10 statistically distinct levels of ability. This widely separated test takers in oral pragmatic performance was support by the analysis of the test takers’ discourse, which also differentiated the test takers. In addition, the range of logit values for the test takers was between -3.63 (the lowest) and 5.44 (the highest), with a total spread of 9.07 logits, exhibiting a wide range of test takers’ ability levels.

The logit difference between the most severe rater and the least severe rater was 0.76 and those of the tasks (the logit distance between the most and the least difficult tasks) and the rating criteria (the logit distance between the most and the least challenging rating criteria) were 0.91 and 1.30 respectively. Table 4.5 below provides overall descriptive statistics of each facet of the Rasch analysis.

<table>
<thead>
<tr>
<th>Facet</th>
<th>Number of elements</th>
<th>Logit</th>
<th>Separation (Reliability)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test taker</td>
<td>67</td>
<td>0.00 2.13 5.44</td>
<td>-3.63 9.07 10.36 (0.99)</td>
</tr>
<tr>
<td>Rater</td>
<td>3</td>
<td>-0.61 0.32 -0.30</td>
<td>-1.06 0.76 7.86 (0.98)</td>
</tr>
<tr>
<td>Task</td>
<td>12</td>
<td>0.00 0.30 0.52</td>
<td>-0.39 0.91 4.94 (0.92)</td>
</tr>
<tr>
<td>Criterion</td>
<td>6</td>
<td>0.00 0.48 0.55</td>
<td>-0.75 1.30 7.75 (0.98)</td>
</tr>
</tbody>
</table>

Notes.
- “Reliability” refers to the extent to which the elements in each facet were reliably separated.
- “Difference” refers to the difference between Maximum and Minimum logit values.

Stable fit statistics indicates internal consistency of the characteristics in performance assessment (McNamara, 1996). The FACETS program provides infit and
outfit mean square statistics of each facet. Infit values were information weighted thus considered as a chief concern for the study whereas outfit statistics is more sensitive to outliers. The current study refers to infit values as an indicator of the variability of the test takers’ behaviour on the tasks as “Aberrant infit statistics usually cause more concern than do large outfit statistics” (Bond & Fox, 2015, p. 67). Table 4.6 below presents the summary of infit values and Table 4.7 reports the details of each rating criterion.

Table 4.6
**FACETS Descriptive Statistics: Internal Consistency**

<table>
<thead>
<tr>
<th>Facet</th>
<th>Number of elements</th>
<th>Mean score</th>
<th>S.D.</th>
<th>Acceptable infit range</th>
<th>Number of elements outside of the acceptable range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test taker</td>
<td>67</td>
<td>1.00</td>
<td>0.20</td>
<td>1.40 to 0.60</td>
<td>3 (all underfitting)</td>
</tr>
<tr>
<td>Rater</td>
<td>3</td>
<td>1.00</td>
<td>0.14</td>
<td>1.28 to 0.72</td>
<td>0</td>
</tr>
<tr>
<td>Task</td>
<td>12</td>
<td>0.99</td>
<td>0.09</td>
<td>1.17 to 0.81</td>
<td>1 (underfitting)</td>
</tr>
<tr>
<td>Criterion</td>
<td>6</td>
<td>1.00</td>
<td>0.11</td>
<td>1.22 to 0.78</td>
<td>0</td>
</tr>
</tbody>
</table>

*Note. The acceptable ranges were computed according to McNamara (1996).*

Table 4.7
**Criteria Measurement Report (Arranged by Criterion Difficulty)**

<table>
<thead>
<tr>
<th>Rating Criterion</th>
<th>Difficulty logit</th>
<th>Model S.E.</th>
<th>Infit mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Use for Mitigation</td>
<td>0.55</td>
<td>0.05</td>
<td>0.97</td>
</tr>
<tr>
<td>Facility with the Language</td>
<td>0.34</td>
<td>0.05</td>
<td>0.86</td>
</tr>
<tr>
<td>Language use to Deliver the Intended Meanings</td>
<td>0.32</td>
<td>0.05</td>
<td>0.94</td>
</tr>
<tr>
<td>Social Actions to Achieve the Communicative Goal</td>
<td>0.08</td>
<td>0.05</td>
<td>1.16</td>
</tr>
<tr>
<td>Engagement in Interaction</td>
<td>-0.55</td>
<td>0.08</td>
<td>0.94</td>
</tr>
<tr>
<td>Turn Organization</td>
<td>-0.75</td>
<td>0.08</td>
<td>1.13</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.00</td>
<td>0.06</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>0.48</td>
<td>0.01</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Values larger or smaller than mean (set at 1.00) indicate more or less variability than was predicted by the model that FACETS computed. A generally acceptable range of infit value is 0.5 to 1.5 (Lunz, Wright & Linacre, 1990). A more strictly defined
acceptable range was proposed by McNamara (1996), which provided the formula: 
Mean score ±2 standard deviation (SD). As seen in Table 4.6, The FACETS provided 
the mean scores of 1.00 and SD of 0.20 for the test taker facet. Applying the formula, an 
acceptable range was between 1.40 and 0.60, which was given by the formula, 1.00 
(Mean) ±0.20 (S.D.) times two. Three test takers’ infit values computed by the Rasch 
analysis were above 1.40 and no test takers’ infit values were below 0.60, indicating 
that these three test takers displayed more variability or unpredictability than was 
expected by the model that FACETS computed. Two test takers’ infit values (1.41 and 
1.46) were slightly over the upper-control limit of 1.40, which was within the range of 
0.5 to 1.5 (Lunz, Wright & Linacre, 1990). One other test taker with infit value of 1.71 
was the only one far beyond the upper-limit control, indicating the test taker’s 
unpredictable performances. The rest of the 64 test takers showed stable fit values, 
f falling within the acceptable range stated above. Regarding the task facet, one task (M5) 
marked its infit value of 1.19, which was slightly over the acceptable range. Overall, 
most of the elements in each of the facets, test taker, rater, task, and criterion involved in 
the current assessment, maintained a certain degree of internal consistency, which 
indicates that those elements in each facet behaved reasonably as predicted by the 
model that the Rasch analysis estimated.

To examine the rating criteria in closer detail, Table 4.7 reports the varied 
difficulties (0.55, the highest for Language Use for Mitigation and -0.75 for Turn 
Organization) as well as the infit values of each criterion. Generally, the estimated 
relative difficulty of each criterion was similar to that of Youn (2013), who reported that 
more interactionally-orientated criteria were less demanding for her test takers while the 
criteria related to language use were more demanding.

As reported in Youn (2015) in her validity argument for the Evaluation
Inference, the current study confirmed that the estimated infit values widely and stably differentiated the pragmatic competence of the test takers. Also, the stable infit values and the varied difficulty of each rating criterion “indicate its unique contribution to tap into the distinct aspects of pragmatic competence” (Youn, 2015, p. 218) defined in the current study. In the following sections, I will discuss the quality of the rating provided by the Rasch analysis.

4.3.4 To What Extent Did the Raters Reliably Assess Test Taker Performance?

Under the fully-crossed rating design, all of the three raters gave scores to all audio-recorded samples collected from 67 test takers. This means that each of the three raters gave the total of 2010 judgments respectively (for dialogue, 3 tasks x 6 criteria x 67 test takers; for monologue, 3 tasks x 4 criteria x 67 test takers).

Inter-rater reliability was first estimated using Person Product Moment correlation for each of twelve tasks. The correlation coefficients used as inter-rater reliabilities for the tasks ranged from 0.63 to 0.91 depending on the pair of raters and on the task itself. For this reliability estimation, a test score for each test taker on each task was defined as the total score of the scores of all criteria given by a rater.

The raters’ rating performances were further evaluated by the Rasch-estimated indices indicating inter-rater reliability (Linacre, 2013). As reported in Table 4.5 earlier, the separation index and the severity range (in logit) of the three raters were substantially smaller than those for the test takers. The raters’ infit values were within the acceptable range. Table 4.8 below presents the logit values and the infit values of each rater estimated by the single Rasch analysis based on all the combined data. The logit values of the three raters indicate relative severity of rating among the raters. The higher the value is, the more severe the rater is.
### Table 4.8
*Rater Measurement Report (Arranged by Rater Severity)*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Severity logit</th>
<th>Model S.E.</th>
<th>Infit mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.30</td>
<td>0.04</td>
<td>0.96</td>
</tr>
<tr>
<td>1</td>
<td>-0.48</td>
<td>0.04</td>
<td>0.85</td>
</tr>
<tr>
<td>3</td>
<td>-1.06</td>
<td>0.04</td>
<td>1.18</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.61</td>
<td>0.04</td>
<td>1.00</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.32</td>
<td>0.00</td>
<td>0.14</td>
</tr>
</tbody>
</table>

*Notes.* Separation: 7.86; Reliability (not inter-rater reliability): 0.98; Exact agreement: 52.6%

Rater 2 was found to be the most severe with a logit value of -0.30 and Rater 3 was the most lenient with the logit value of -1.06. The severity difference between Rater 1 and Rater 2 is smaller than that between rater 1 and rater 3. The severity difference among the raters was also confirmed by the separation index (7.86) and the separation reliability (0.98) in Table 4.5 which indicates that the raters’ judgments were reliably spread. This degree of rater difference in severity for assessing oral pragmatic performances and interaction is not surprising (Youn, 2013) so long as human raters are involved to assess speaking performances (McNamara, 1996).

The severity difference among the raters was smaller than the difference reported in Youn (2015). The infit values of all three raters were within the acceptable range (McNamara, 1996) of estimated infit values. Also, simple comparisons of the current results with those reported in previous studies were, of course, not possible because the rating design (e.g., the number of raters, fully-crossed rating design or not, the number of band levels) was not identical to the methods used in the previous studies. However, it would be reasonable to conclude that the obtained logit values and the infit values in the current research were within the expectations of the literature.

In addition, as reported in Table 4.5 in the previous section, the logit difference between the most severe and the most lenient raters was much smaller than the logit
difference between the most and least competent test takers. The logit difference indicating severity in rating between the most and the least severe raters was 0.76, which was about one-twelfth as small as the logit difference (9.07) indicating test taker ability between the most and the least competent test takers.

Therefore, impact of rater’s severity difference is small as suggested by Myford and Wolfe (2000) as follows:

The finding that the range of TSE examinee proficiency measures is about five times as wide as the rage of TSE rater severity is an important one, because it suggests that the impact of individual differences in rater severity on examinee scores is likely to be relatively small. (p. 11)

The percentage of the exact agreement was 52.6 %, indicating that in respect of the raters’ judgments, all of the three raters gave the same score for about half the cases. However, the exact agreement percentage needs to be interpreted with caution (Knoch, 2009). An extreme case would be, for example, that all raters gave score 1 (the lowest score among the four band levels) to all 67 test takers. In this case, all raters completely agreed on the score, but it would be highly questionable whether the awarded score reflected the reality. As discussed in Figure 4.2 (the probability curve) in the next section, this hypothetical case was not realized in the current study, where the raters operated the assessment band levels appropriately. Future studies on assessment of the similar construct of oral pragmatics will be necessary to compare these results with and to evaluate the current findings. However, the obtained percentage could affirm a certain degree of the shared understanding of the raters, which the researcher as the rating rubrics developer confirmed in the rater training session.
4.3.5 How Appropriately Were the Rating Rubrics Operated by Raters?

The FACETS program provides the probability curves that visually present how the rating criteria functioned. Figure 4.2 below provides the probability curves of *Language Use to Deliver the Intended Meanings* as an example. The probability curve was provided for each rating criterion in the same Rasch analysis as conducted in sections 4.3.3 and 4.3.4.

*Figure 4.2. Probability curve: Language Use to Deliver the Intended Meanings.*

Each of the rating criteria had four level bands into which the test taker could be classified. Each curve was labeled according to the band levels used for this study (1 to 4). The horizontal line indicates test taker ability increasing from left to right with the minimum value of -6.0 and the maximum value of 6.0. The vertical line denotes the probability (from 0 to 1) of a score being given to a test taker at a particular ability level on the horizontal line. For instance, score 4 was more likely to be awarded to the more competent test takers and the probability of receiving score 4 increases according to the test takers’ estimated ability. In contrast, the probability of receiving score 1 was the highest with test takers at the lowest ability (closest to -6.0).
The primary concern, when looking at probability curves, is whether each score level (1 to 4) shows distinct peaks (displayed vertically) and whether curves equally spaced (Knoch, 2009). Although band level 3 or 2 were slightly higher or lower depending on the criterion, very few problems were identified in the probability curves for the other rating criteria as well as for Language Use to Deliver the Intended Meanings, in which each band level showed different peaks, and they were generally evenly spaced. This implies that the rating band levels were operated by the raters appropriately to separate the test takers according to their competence. The similar tendency was confirmed for those relating to the other rating criteria, suggesting that all raters reasonably complied with the rating rubrics as expected by the current study.

4.4 Generalization

The Generalization inference rests on the understanding of whether similar test results of the test are produced by the assessment under similar conditions when the test is conducted for a similar test taker population. It is also concerned with whether paralleled task sets have comparability and whether the items are representative of possible items from the Universe of Generalization. In other words, the Generalization inference relies on an assumption that test results obtained under individual tests are generalizable beyond a particular test occasion but are representative of equivalent test contexts. This section thus addresses the following question:

To what extent does the assessment yield test results consistent across assessment contexts: task administration conditions, raters, and task sets?

The investigation into this inference was guided by the specific backings and the
sub-questions listed in Table 4.9 below.

Table 4.9
Backings and Sub-questions to Address the Generalization Inference

<table>
<thead>
<tr>
<th>Backings</th>
<th>Source of supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ Sub-questions</td>
<td>Test reliability (indicated by Cronbach Alpha)</td>
</tr>
<tr>
<td>1. What are the reliability estimates for the test items (test tasks)?</td>
<td>Comparability evaluated by the results of separate Rasch analyses using the data on task set 1 and task set 2 respectively</td>
</tr>
<tr>
<td>2. How and to what extent do the two task sets function similarly?</td>
<td>The practice of the test administration in this study</td>
</tr>
<tr>
<td>3. Test administration conditions were consistent for all test takers.</td>
<td></td>
</tr>
</tbody>
</table>

As Kane (2013) suggested, the Generalization inference and the Evaluation (Scoring) inference have shared a certain aspect in terms of relevant backings that these inferences rest upon. Evidence for the Generalization inference can be sought through “observations that can vary in a number of ways, involving, for example, samples of tasks, testing contexts, occasions in which the test is administered, and possibly raters who score the responses” (Kane, 2013, p. 26). In this section, I will discuss the reliability of the tasks, the comparability of the two task sets, as well as the issues of test contexts (test administration condition and sampling). The raters’ rating performances examined for the purposes of addressing the Evaluation inference in the previous section can also support the Generalization inference which involves an issue of variation in raters and thus in raters’ performances.

4.4.1 What Are the Reliability Estimates for the Test Items?

As an overview of how the designed test functioned, test reliability was examined for each task in each task set first. For this reliability estimation, each test taker’s total score for one task was calculated by adding all the scores on the rating
criteria assigned from three raters. Table 4.10 reports the results of the reliability estimates. Each task was labeled with an ID (e.g., D1, M2). “D” and “M” refer to dialogue and monologue respectively. The numbers (e.g., “1” as in “D1”) show the number of the order of the tasks within the six tasks assigned to the test takers in each task set.

Table 4.10  
*Test Item Reliability*

<table>
<thead>
<tr>
<th>Addressee</th>
<th>Task</th>
<th>Mean (S.D.)</th>
<th>Task</th>
<th>Mean (S.D.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Professor</td>
<td>Dialogue (D1)</td>
<td>46.03 (13.37)</td>
<td>Monologue (M1)</td>
<td>30.29 (8.83)</td>
</tr>
<tr>
<td>2 Administrator</td>
<td>Monologue (M2)</td>
<td>31.70 (8.14)</td>
<td>Dialogue (D2)</td>
<td>45.44 (13.35)</td>
</tr>
<tr>
<td>3 Peer student (unfamiliar)</td>
<td>Dialogue (D3)</td>
<td>48.24 (12.00)</td>
<td>Monologue (M3)</td>
<td>30.41 (8.52)</td>
</tr>
<tr>
<td>4 Peer student (close)</td>
<td>Monologue (M4)</td>
<td>32.21 (7.47)</td>
<td>Dialogue (D4)</td>
<td>48.50 (11.71)</td>
</tr>
<tr>
<td>5 Administrator</td>
<td>Dialogue (D5)</td>
<td>49.33 (11.31)</td>
<td>Monologue (M5)</td>
<td>29.70 (8.37)</td>
</tr>
<tr>
<td>6 Professor</td>
<td>Monologue (M6)</td>
<td>32.06 (7.71)</td>
<td>Dialogue (D6)</td>
<td>50.32 (9.99)</td>
</tr>
</tbody>
</table>

Cronbach’s Alpha 0.967 0.967

It was confirmed that the estimated alpha was high (0.967) and each test task contributed to the reliabilities in each task set (set 1 and set 2). This was also reflected in the inter-item correlations ranging from 0.850 to 0.929 for set 1, and from 0.766 to 0.926 for set 2. Each test taker’s total score on each test task were defined as being the same for estimation of the test reliability.

The obtained indices discussed in the study can indicate that the test items measured the same construct and went in the same direction. For the current study, it was impossible to give a number of speaking tasks to the test takers, given the time-consuming nature of performance-based assessment (Kane et al., 1999). Although a
range of possible tasks simulating university settings were brainstormed and reviewed in the phases of task development to avoid under-representing the Universe of Generalization, the number of tasks assigned to each test taker was limited to six. During the process of the task development, the degree of imposition, one of the three contextual variables (Brown & Levinson, 1987) was designed to be middle or higher as described in Chapter 3, because lower imposition may not be able to elicit test takers’ oral discourse as extended as the current study intended. Therefore, in the methods practiced in the current study, it would be hardly imaginable that a particular task among a limited number of tasks functioned differently in measuring the test takers’ oral pragmatic abilities. Therefore, it can be concluded that the high reliability and the inter-item correlations were attained as the task design intended.

4.4.2 How and to What Extent Do the Two Task Sets Function Similarly?

The current study used two parallel task sets. The similarity of the developed two test sets were examined by two separate Rasch analyses, which provided another backing which allowed for a stronger inference of the Generalization. The constructs embedded in the two sets were designed to be the same, and either set 1 or set 2 (consisting of six tasks respectively) was assigned to each test taker. The total of 67 test takers in the main study were divided into two groups assigned with either set 1 or set 2 respectively, under an assumption that the average ability of the two groups was the same. Table 4.11 below shows the assignment of the test takers to each task set.
Table 4.11
Assignment of the Test Takers

<table>
<thead>
<tr>
<th>Test taker</th>
<th>Number of test takers on Task set 1</th>
<th>Number of test takers on Task set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced university students (N=9)</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Other university students (N=35)</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Pre-entry students (N=23)</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td><strong>Set total</strong></td>
<td><strong>33</strong></td>
<td><strong>34</strong></td>
</tr>
</tbody>
</table>

The test takers for each of the three groups was divided into halves respectively, thus the number of the test takers assigned to task set 1 and task set 2 was made to be the same. For the current research sub-question, the two separate Rasch analyses were performed with one analysis using the data of set 1 and the other using the data of set 2 respectively. Table 4.12 below summarizes the test taker measurements of the separate Rasch analyses.

Table 4.12
Test Taker Measurements of the Two Separate Rasch Analyses

<table>
<thead>
<tr>
<th></th>
<th>Task set 1</th>
<th>Task set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separation index (Reliability)</td>
<td>10.38 (0.99)</td>
<td>10.34 (0.99)</td>
</tr>
<tr>
<td>Logit mean (S.D.)</td>
<td>0.00 (2.14)</td>
<td>0.00 (2.12)</td>
</tr>
<tr>
<td>Fair average mean (S.D.)</td>
<td>2.71 (0.62)</td>
<td>2.65 (0.63)</td>
</tr>
<tr>
<td>Observed average mean (S.D.)</td>
<td>2.66 (0.62)</td>
<td>2.61 (0.63)</td>
</tr>
<tr>
<td>Number of test takers overfitting</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(below the lower limit)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of test takers underfitting</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>(above the upper limit)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The separation indices were 10.38 and 10.34 for the two sets respectively with reliability indices of 0.99 for both cases. This indicates that both task sets were able to separate the assigned test takers similarly widely and also reliably. The other values such as S.D. for the logit values, the mean scores of the fair average of each test taker with their S.D. as well as the mean scores of the observed average for each test taker with S.D. were all similar between set 1 and set 2. With respect to infit value of each
test taker, one test taker on set 1 and two test takers on set 2 were above the upper limit of an acceptable infit range calculated following McNamara (1996). This indicates that these three test takers behaved somewhat unpredictably, but most of the test takers in both groups were within the acceptable infit ranges. This, in turn, implies that both of the two sets similarly minimized unpredictable behaviour by the test takers or by the raters as a whole for each group.

In addition to the degree of the test taker separation, how the test takers are separated in each task set group was investigated by the test takers’ estimated competence by the separate Rasch analyses. Figure 4.3 and Figure 4.4 visually present how the test takers were separated in the case of task set 1 group and that of task set 2 group respectively. As shown in Table 4.11 above, the test takers were divided into the three groups (Advanced university students, Other university students and Pre-entry students) in almost equal numbers in each task set group.

Figure 4.3. Test taker distribution (task set 1).
Both cases revealed a similar tendency in which the Advanced university students exceeded the other university students and the Pre-entry students. The estimated competency range of the Other university students and the Pre-entry students were, to a certain extent, overlapped and the latter was found to be the least competent as a group. The similarity between the two task sets was further supported by the raters’ performances as summarized in Table 4.13 below.

Figure 4.4. Test taker distribution (task set 2).
Table 4.13

*Rater Measurements of the Separate Rasch Analyses (Arranged by Rater Severity)*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Task set 1</th>
<th></th>
<th>Task set 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Severity logit</td>
<td>Infit mean square</td>
<td>Rater</td>
<td>Severity logit</td>
</tr>
<tr>
<td>Rater 2</td>
<td>-0.38</td>
<td>0.97</td>
<td>Rater 2</td>
<td>-0.23</td>
</tr>
<tr>
<td>Rater 1</td>
<td>-0.56</td>
<td>0.80</td>
<td>Rater 1</td>
<td>-0.41</td>
</tr>
<tr>
<td>Rater 3</td>
<td>-1.12</td>
<td>1.21</td>
<td>Rater 3</td>
<td>-1.00</td>
</tr>
<tr>
<td>Mean</td>
<td>-0.68</td>
<td>0.99</td>
<td>Mean</td>
<td>-0.55</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.31</td>
<td>0.17</td>
<td>S.D.</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Note.
Reliability: 0.97; Separation: 5.36 for Task set 1
Reliability: 0.97; Separation: 5.67 for Task set 2
Rater facet was set as non-centered, thus the logit mean of rater facet (severity) was not set as zero^4^.

The raters’ performances in both cases were in general consistent with the findings estimated based on the combined data of the two sets (as discussed for the Evaluation inference above). Regardless of the cases, Rater 2 was found to be the most severe and Rater 3 was the most lenient in the rating with similar severity logit of one rater on task set 1 to the counterpart on set 2. The infit values of the all raters in both cases were within the acceptable ranges, indicating that all of the three raters rated with internal consistency on both task sets. It could therefore be assumed that the difference in the task set did not substantially change the raters’ performance in consistency and severity.

The difficulty and the infit values for the tasks in the two sets were also examined. Table 4.14 below summarizes the results.

---

^4^ For these separate Rasch analyses, it was possible to set test taker facet as non-centered instead of rater facet because the test takers on task set 1 and those on set 2 were not group-anchored here. However, rater facet was selected as non-centered facet, in order to be consistent with the Rasch analysis using all the combined data reviewed in the previous sections.
Table 4.14
*Task Measurements of the Separate Rasch Analyses (Arranged by Estimated Difficulty for Each Set)*

<table>
<thead>
<tr>
<th>Task item</th>
<th>Difficulty Logit</th>
<th>Infit mean square</th>
<th>Task item</th>
<th>Difficulty Logit</th>
<th>Infit mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 Professor</td>
<td>0.50</td>
<td>1.11</td>
<td>D2 Administrator</td>
<td>0.43</td>
<td>1.06</td>
</tr>
<tr>
<td>D3 Peer student</td>
<td>0.09</td>
<td>1.02</td>
<td>M5 Administrator</td>
<td>0.19</td>
<td>1.20</td>
</tr>
<tr>
<td>M2 Administrator</td>
<td>-0.08</td>
<td>0.94</td>
<td>M1 Professor</td>
<td>0.03</td>
<td>0.89</td>
</tr>
<tr>
<td>D5 Administrator</td>
<td>-0.11</td>
<td>0.90</td>
<td>M3 Peer student</td>
<td>-0.01</td>
<td>1.01</td>
</tr>
<tr>
<td>M6 Professor</td>
<td>-0.18</td>
<td>0.94</td>
<td>D4 Peer student</td>
<td>-0.14</td>
<td>0.92</td>
</tr>
<tr>
<td>M4 Peer student</td>
<td>-0.22</td>
<td>1.02</td>
<td>D6 Professor</td>
<td>-0.49</td>
<td>0.91</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>0.00</td>
<td>0.99</td>
<td><strong>Mean</strong></td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>S.D.</strong></td>
<td>0.24</td>
<td>0.99</td>
<td><strong>S.D.</strong></td>
<td>0.28</td>
<td>0.11</td>
</tr>
</tbody>
</table>

*Note.*
Reliability: 0.88; Separation: 2.75 for Task set 1
Reliability: 0.91; Separation: 3.28 for Task set 2

The tasks in task set 2 were somewhat more widely and reliably separated in the estimated difficulty than those in task set 1 according to the separation indices (2.75 for task set 1 and 3.28 for task set 2 with the reliability indices 0.88 for task set 1 and 0.91 for task set 2). The logit difference between the most and the least difficult task items was less than 1 logit for both cases of task set 1 and task set 2. These figures indicate that the tasks, regardless of the sets, were separated within a much smaller range than compared to the separation of the test takers.

As Iwashita, McNamara and Elder (2001) argued, variables affecting task difficulty are complex, which could make it difficult to predict and adjust the difficulty of the tasks as intended. A difference in difficulty between the two task sets may exist. However, it should be safe to conclude that each task set, taken as a whole task set of six task items, had a similar degree of demand on the test takers, given the fairly smaller range of the separation for the tasks.

With regard to the infit values, task set 2 showed more varied values among its task items, according to each of the six infit values (0.43 to -0.49) in task set 2 and its
S.D. (0.28). The infit values of all task items in both task sets were found to be acceptable according to McNamara’s (1996). Combined with the obtained results of the test takers and the raters, it can be indicated that both task sets functioned similarly in eliciting pragmatic performance from the test takers and discriminating them according to the abilities underlying the test construct of pragmatics.

To summarize, overall, the two task sets seemed to be functioning similarly, as marked differences between the two sets were not indicated in the results. In the current study, the test takers did one or the other of the two task sets. In the process of the assignment of the task sets to each test taker, the average abilities of the test takers on both task sets and the constructs embedded in both task sets were operationalized so as to be the same. The similarity of the two task sets can provide a backing for a part of Generalization inference that similar assessment outcomes can be obtained by test takers on equivalent versions of the test.

4.4.3 Consistency of the Test Administration for the Test Takers

Consistency of the task administration conditions across all test takers was important to the current study as it was logistically not possible to administer the oral performance-based tasks to all the test takers in the same room simultaneously. The task administration required a room suitable for audio-recording and a trained interlocutor as a task administrator.

As well, the data collection also entailed the post-task meetings (for the questionnaire) with the individual test takers by the researcher. Therefore, no single test taker completed the test tasks and the post-task questionnaire at the same time with another test taker. However, the trained the interlocutors (as the test administrators) and the created similar test conditions adequately served to ensure the consistency of the test
conditions across all test takers.

4.5 Explanation

The Explanation inference allows us to estimate the extent to which the test results are reflective of the target construct of pragmatics, which is, in the current study concerned with interacting appropriately in speaking contexts in an academic environment. In other words, this phase of validation examines the extent to which the test takers’ task performances are accounted for by the test construct of L2 pragmatics, as reflected in the assessment instruments. The Explanation inference, as reviewed in Chapter 2, can be distinguished from the Extrapolation (Chapelle, 2008; Roever et al., 2014; Xi, 2008) or treated as integrated into the Extrapolation inference (Kane, 2006, Knoch & Elder, 2013; Youn, 2013). These inferences can be identified by research methodology as proposed in the literature. More importantly, they were distinguished by the extent to which the obtained evidence offers an account for the construct underlying the test (Explanation) or for linking the test performances to the real life situations (Extrapolation).

The previous stages of validation (Domain Description, Evaluation, and Generalization) examined the path of the observed test performance converted into the test score or “observed score” (concerned with the Evaluation inference) and whether the observed score can be generalized in and made representative of the targeted language activity settings, where students handle pragmatic demands. The Explanation and the Extrapolation inferences allow us to estimate the extent to which the test scores indicate their pragmatic abilities in the target domain. The current study differentiated the Explanation inference from the Extrapolation inference as the former primarily relates to whether, how and to what extent the test results are reflective of the test
construct of pragmatics, whereas the latter essentially relates to how and to what extent
the results are indicative of the reality. The current section will discuss the evidence
contributing to the backings for the Explanation inference, followed by section 4.6
addressing the Extrapolation inference. Thus, the following question is addressed in this
section:

To what extent can test takers’ test results be attributed to the construct of
pragmatic abilities utilized for language activities at university?

This question was addressed by investigating to what extent the test takers’ task
performances relate to English proficiency and experience in respect of the target
settings. Three types of supporting evidence were integrated in an effort to address this
question. These are outlined in Table 4.15 below.

Table 4.15
Backings and Sub-questions to Address the Explanation Inference

<table>
<thead>
<tr>
<th>Backing</th>
<th>Source of Supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ Sub-question</td>
<td></td>
</tr>
<tr>
<td>Do the test takers’ task performances relate to English proficiency and experience to the target settings?</td>
<td>A group comparison of the test takers’ pragmatic abilities estimated in the Rasch analysis</td>
</tr>
<tr>
<td></td>
<td>A correlation between the test takers’ proficiency levels and their estimated pragmatic abilities</td>
</tr>
<tr>
<td></td>
<td>A group comparison in the test takers’ task preparation time</td>
</tr>
</tbody>
</table>

The first evidence was provided from a comparison of the test takers groups that
were expected to differ as to their pragmatic abilities. It was examined whether the test
takers’ estimated pragmatic abilities were separated by the groups as hypothesized in the
current study. Subsequently, another piece of evidence was sought from a correlational analysis. This piece of evidence was then used further so as to investigate the relationship between English proficiency and the Rasch-estimated pragmatic abilities that Youn (2013) relied upon to address the Extrapolation inference (treated as the Explanation inference in the current study). As supplementary evidence, how the test takers used the task preparation time was investigated in the discussion.

4.5.1 A Group Comparison in the Test Takers’ Pragmatic Abilities Estimated in the Rasch Analysis

The Explanation inference requires evidence in how and to what extent the test results can be explained as reported in the literature. Although research designs (e.g., the definitions, quantification methods, classification methods of proficiency, exposure and pragmatic performance as well as definition of pragmatics) were different for each study, it appears that the literature on pragmatics generally has reached the consensus that proficiency and exposure to the target settings were two major factors accounting for pragmatic performances (Roever et al., 2014).

The total of 67 test takers in the main study were divided into three groups as in Table 3.7 (see section 3.4.1 in Chapter 3). For statistical analyses, their proficiency levels and the experience of language activities at English-medium university were quantified as in Table 4.16 below. The boundary of the target experience of language activities at English-medium university was drawn between the Pre-entry students (as prospective students) and the two groups of university students with the latter being actually in the target domain. In view of English proficiency, the groups of Advanced university students, Other university students and Pre-entry students were quantified as 3, 2 and 1 according to the proficiency levels.
According to the assumption that the literature has provided (Roever, et al., 2014), Advanced university students, being more proficient with experience at the target domain, were likely to outperform the other groups. It was also hypothesized that the group of Other university students as a whole was pragmatically more competent than the group of Pre-entry students who were separated by both variables from the Other university students. It was difficult to predict precisely to what extent a particular test taker group was more strongly or weakly differentiated from another and from all the other groups.

However, it was deemed unlikely that the Advanced university students as a group would underperform the group of Pre-entry students since the former was in the target language activity domain and substantially distant from the latter concerning English proficiency. As described in section 3.4.1, the group of Pre-entry students lacked the experience of language activities at English-medium university, which the current study identified as one of the factors accounting for the pragmatic ability. In addition, the lengths of stay in English speaking regions they reported was limited (to a few months), suggesting that they had little exposure to English-speaking environments.

To investigate how and to what extent the test takers’ pragmatic competence is related to the rank of assumed pragmatic abilities (in Table 4.16 above), the test takers’
pragmatic abilities estimated by the Rasch analysis (based on all the data) were examined. A group comparison was subsequently conducted by examining the ranges of the pragmatic competence (expressed in logit) for each group. Figure 4.5 below visually represents the ranges of the estimated abilities of the three test taker groups.

![Boxplot of Pragmatic Competence for Three Groups](image)

*Figure 4.5. Test taker distribution (all test takers).*

The boxplots in Figure 4.5 above revealed that the two groups of university students were more pragmatically competent. The median values (indicated as a line in each boxes), and the 50 percent of the cases (indicated as the boxes themselves) were higher for the two groups of university students than those of the group of Pre-entry students. The group of Advanced university students outperformed the Other university students and the Pre-entry students although the lower limit of the Advanced university student group was close to the upper limit of the Other university student group. The Other university student group and the Pre-entry student group overlapped although the
average of the former was to some extent higher than that of the latter. Some test takers in the group of Other university students were found to be lower in their pragmatic abilities than the abilities of some Pre-entry students. The separation of the three groups was generally similar both to the separation of the test takers on task set 1 (Figure in 4.3) and to that on task set 2 (Figure 4.4) in section 4.4.2.

The ranges of the estimated ability as seen in Figure 4.5 are reported in Table 4.17 below. The competence range of the whole group of test takers was divided into four ability areas as seen in the table, which shows how many test takers in the group of Other university students and the group of Pre-entry students were overlapped.

Table 4.17
Test Taker Distribution According to their Estimated Pragmatic Abilities

<table>
<thead>
<tr>
<th>Ability area range</th>
<th>Pragmatic competence (logit value)</th>
<th>Advanced university student (N=9)</th>
<th>Other university student (N=35)</th>
<th>Pre-entry student (N=23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>5.44 to 2.76</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2nd</td>
<td>2.45 to -0.67</td>
<td>0</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>3rd</td>
<td>-0.72 to -1.88</td>
<td>0</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>4th</td>
<td>-1.92 to -3.63</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

All of the Advanced university students were not overlapped by the other groups and fell within the first and the highest ability range (5.44 to 2.76 in logit), as visually presented in Figure 4.5 above. The second highest area (2.45 to -0.67) indicates the ability range of the Other university students who were not overlapped by the other groups. The area overlapped by Other university students and Pre-entry students was the area indicating the third highest ability range with the upper limit of the logit value of -0.72 (the highest among the Pre-entry students’) and its lower limit of the logit value of -1.88 (the lowest among the Other university students’). The fourth area as the lowest ability comprised exclusively the remaining Pre-entry students.
The only part which overlapped with different groups was the third area which included 10 Other university students and 13 Pre-entry students. More than two-thirds of the Other university students in the second area exceeded all of the Pre-entry students, although more than half of the Pre-entry students fell within the third area with about one-third or less test takers of the Other university student group. The test taker found pragmatically the most competent in the group of Other university students almost reached the lower limit of the group of Advanced university students.

The results show that all nine test takers in the group of Advanced university students differed from the other test takers, in particular from the group of Pre-entry students, whose general English proficiency was clearly and markedly lower than that of the Advanced university students. The nine test takers in the group of Advanced university students were revealed as both proficient in English and pragmatically competent. This was not always the case for the test takers in the other two groups which overlapped to a certain extent. Regarding the groups as a whole, the Other university students outperformed the Pre-entry students as predicted by the assumed rank of the three groups in pragmatic abilities (in Table 4.16 above).

4.5.2 A Correlation between the Test Takers’ Proficiency Levels and their Estimated Pragmatic Abilities

The test taker groups in Table 4.16 above indicated the rank of pragmatic competence (3, 2 and 1) assumed to differ according to the discussion in the literature. Simultaneously, the group category can also be used as an indicator of their English proficiency levels. Proficiency has been recognized in the literature as a major factor that accounts for L2 speakers’ pragmatic competence, although measurements and definitions of the two variables (proficiency and pragmatic ability) were not consistent
in different empirical studies. Although the reported strengths of the association between proficiency and pragmatic ability were also different in these studies, the literature has generally agreed that pragmatic performances are accounted for and are relate to their proficiency levels in degrees between “to a certain extent” and “to a large extent.”

In order to confirm the strengths of the two variables, a correlational analysis was performed using the test takers’ estimated pragmatic abilities and their proficiency levels (3, 2 and 1) as indicated in Table 4.16. The proficiency levels defined in the table could be treated as ordinal data because the distance between two adjacent levels was not always assumed to be equal, thus the data were not always assumed to be interval data. Therefore, the Spearman’s rho correlation analysis was performed. The results of the analysis showed a strong and statistically significant correlation between the test takers’ ability and their proficiency level (rho=0.801, p < 0.01), although the correlational strength was weaker than that in Youn (2013), who reported a considerably high correlation (r=0.90) between proficiency and pragmatic performance.

Generally, as predicted by the literature, the test results were accounted for by proficiency and experience of language activities in the target domain. To summarize the findings, the pragmatic performance differentiated the groups of university students and the group of Pre-entry students, as the former groups were more proficient in English and were in the target domain. The group of Advanced university students with higher proficiency outperformed the Other university students who were less proficient. The results also showed a substantial distance in the pragmatic competence between the group of Advanced university students (with the highest proficiency and the experience of language activities in the target domain) and that of the Pre-entry students who were presumably the least proficient in English as a whole group. Ten out of the 44 university
students overlapped with thirteen out of the 23 Pre-entry students in the shared area, where one Pre-entry student or more found to be more pragmatically competent than these ten university students. This may be due to the fact that unlike the general ESL learners in the pilot study (as described in section 3.3.4.5 in Chapter 3), the Pre-entry students were preparing for university entry, thus expected to (a) be highly aware of language use for university activities and to (b) have already reached adjacent levels of English proficiency for university study. Therefore, it should not be impossible for Pre-entry students to perform better than some university students, although the estimated pragmatic abilities of the majority of university students exceeded the Pre-entry students’ abilities. If the group of Pre-entry students as a whole group collectively had exceeded Advanced university students or the majority of the Other university students in the estimated pragmatic abilities, which was against the expectations of the literature, further investigation would be necessary to provide an account for such outcomes. However, it was confirmed that the reverse outcomes were not the case for the obtained results in the current study.

4.5.3 A Group Comparison in the Test Takers’ Task Preparation Time

The time that each test taker spent to prepare for the assigned test tasks was also examined. Their task preparation time was expected to play a role as supplementary evidence to support an assumption that the task preparation time reflects their readiness for the simulated university settings and thus the three test taker groups are differentiated according to the time required to prepare for the tasks. More specifically, it was hypothesized that Advanced university students required less time and Pre-entry students required more time to make themselves ready for the simulated university settings, although individual differences in how to use the time were assumed to be
equally possible.

As described in section 3.5.1 in Chapter 3, each test taker performed six role plays. Each test taker was given five minutes (300 seconds) plus an extra two minutes (120 seconds) if necessary, to prepare for each task. Therefore, they were allowed to use 30 minutes (1800 seconds) plus extra twelve minutes (720 seconds) in total to prepare for the assigned six tasks. They did not need to spend all of the given preparation time since they were instructed to start each task whenever they were ready (see Appendix A for the general instructions to the test takers). Thus, it depended on the test takers how they use the given time. The test takers were allowed to ask questions to the role play conductor (playing the role as the interlocutor at the same time) to clarify the described task situations if necessary. The recorded preparation time included the time when some test takers were asking questions and receiving responses from the interlocutor.

Shorter preparation time may indicate the test taker’s high ability or it may possibly lead to the test taker’s insufficient understanding of the task situation, or it may generate a performance not adequately addressing the task goal. Although these possible varied outcomes and somewhat unpredictable personal behaviour of test takers in the task preparation time were acknowledged, it was difficult to expect that the Pre-entry students as a whole needed less preparation time than the Advanced university students.

Table 4.18 shows the time (in seconds) that test takers spent for preparing for the tasks. The time that the test takers in each group spent is also visually represented in Figure 4.6.
Table 4.18
Total Amount of Time Spent for Task Preparation

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean (S.D.)</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced university student (N=9)</td>
<td>874 (238)</td>
<td>1300</td>
<td>647</td>
</tr>
<tr>
<td>Other university student (N=35)</td>
<td>1368 (425)</td>
<td>2125</td>
<td>356</td>
</tr>
<tr>
<td>Pre-entry student (N=23)</td>
<td>1542 (333)</td>
<td>2085</td>
<td>1047</td>
</tr>
<tr>
<td>All (N=67)</td>
<td>1362 (425)</td>
<td>2125</td>
<td>356</td>
</tr>
</tbody>
</table>

Figure 4.6. Test taker distribution by task preparation time.

Overall, the Advanced university students were more likely to start the tasks more quickly than the test takers in the other two groups, simply by referring to the mean time (in seconds) of each group. The Pre-entry students as a whole group required
substantially more time than the group of Advanced university students as a whole. The largest variation in use of the preparation time was seen in the group of Other university students with the minimum of 356 and the maximum of 2125. This may not be surprising as they comprised the largest number of test takers (N=35).

In order to initiate each task, the test takers were required to input the information in the task prompt and to prepare for the output, and, if necessary, for the structuring and the rehearsing of the speech. The results in Table 4.18 could suggest that, generally speaking, the two groups of university students, and in particular the group of Advanced university students, initiated the tasks more quickly, which may indicate the Advanced university students’ greater readiness for the simulated language activities at university.

The evidence from the test takers’ task preparation time itself may only allow for a limited conclusion about the test takers’ pragmatic abilities. The group boundaries in use of task preparation time were, to some extent, blurred. However, the obtained results in the current section did not contradict the results of the group comparison in the previous section. At the very least, the results in this section suggest that the group of Pre-entry students, compared to that of Advanced university students, perceived greater challenges when handling pragmatic demands in the simulated university situations, which was also reflected in their self-assessments (to be discussed in the next section) and in their pragmatic abilities as estimated by the Rasch analysis. These pieces of evidence and the results of the Rasch analysis, to a certain extent, mutually supported the findings of each analysis, which eventually could support an assumption that the test takers were differentiated according to their pragmatic ability. Thus, the evidence from the test takers’ task preparation time could support the findings of the previous section to allow for a stronger inference of the Explanation.
4.6 Extrapolation

The Extrapolation inference links the test results to a broader domain in reality and is supported by understanding to what extent the test results are informative of the test takers’ pragmatic abilities to utilize in the authentic situations at university. Thus, the current section addresses the following research question:

To what extent are the performances observed in the tasks indicative of the test takers’ real pragmatic performances in the target domain?

It was virtually impossible for the current study to observe the test takers’ authentic performances in the target domain. The relevant evidence was sought from the perceptions of the test takers as summarized in Table 4.19 below.

Table 4.19
Backings and Sub-questions to Address the Extrapolation Inference

<table>
<thead>
<tr>
<th>Backing</th>
<th>Source of supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>↑ Sub-questions</td>
<td>Test takers’ perceptions of similarities between their task and authentic performances summarized for descriptive statistics</td>
</tr>
<tr>
<td>1. To what extent do the test takers perceive their task performance to be similar to their real performances in the university domain?</td>
<td>Test takers’ perceptions of similarities between their task and authentic performances summarized for descriptive statistics</td>
</tr>
<tr>
<td>2. To what extent does the test administration procedure affect the test takers’ performances?</td>
<td>Test takers’ perceptions elicited through the questionnaire and summarized for descriptive statistics</td>
</tr>
<tr>
<td>3. To what extent do the performances on the measure relate to the test takers’ own assessment of their performances?</td>
<td>Correlation between their pragmatic abilities estimated by Rasch analysis results and their self-assessment results</td>
</tr>
</tbody>
</table>

All three backings relied on the test takers’ perceptions. As described in Chapter 3, the test takers’ perceptions were summarized statistically. The first backing was sought from a comparison made by the test takers’ perceptions about their task and the authentic performances. The second backing was assured by investigating the extent to
which construct-irrelevance in the test takers’ performance was minimized. The takers’ self-assessed task performances were integrated as the third baking. The test takers were in reality the language users in the domain of higher education or those who aimed to enter the domain. Hence, their own perceptions and self-assessment were, in effect, an invaluable source of evidence to address the Extrapolation inference.

4.6.1 To What Extent Do the Test Takers Perceive their Task Performance to be Similar to their Real Performances in the University Domain?

The test takers provided their perceived similarity between their language use in the task performances and that in reality. The data were statistically summarized for descriptive statistics by the groups. For this sub-research question, no presumption was made about what backgrounds of test takers were more likely to perceive the similarity between the tasks and the reality because there were very few theoretical assumptions in the related literature to rely upon. The test takers’ perceived similarity between their language use in the task performance and their real performance was measured by a four-point Likert scale. They expressed their perceived task similarity by four scales, “4: Very similar,” “3: Similar,” “2: Not so similar,” “1: Not similar at all.” Table 4.20 below presents the mean scores and standard deviation of each group. The responses were collected from all of 67 test takers.
## Table 4.20

### Perceived Similarity between Task and Actual Performances

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean (S.D.)</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced university student</td>
<td>3.67 (0.43)</td>
<td>4.00</td>
<td>3.00</td>
</tr>
<tr>
<td>(N=9)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other university Student</td>
<td>3.21 (0.63)</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>(N=35)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-entry student</td>
<td>2.97 (0.64)</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>(N=23)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>3.19 (0.64)</td>
<td>4.00</td>
<td>2.00</td>
</tr>
<tr>
<td>(N=67)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Generally, the test takers recognized their language use in task performance as similar to their possible real tasks as confirmed by the mean scores for each of the three test taker groups (3.67, 3.21, and 2.97) and for the whole group (3.19), with the mean score of the Pre-entry students being slightly below 3.00. Some test takers reported different perceived similarities in different task(s) or in different modalities. For instance, they regarded their dialogue performances as “4: Very similar” to their real performance while they judged their monologue performances as “3: Similar.” In this case, the test taker’s perceived similarity for the overall performance was computed as 3.5.

The mean scores of two groups of university students (3.67 and 3.21 respectively) were higher than the group of Pre-entry students (2.97), showing that university students with higher proficiency, combined with experience of the target settings, perceived a stronger similarity between task performance and actual performance. It was also confirmed that the mean scores of the Advanced university students (3.67) were markedly higher than those of the Other university students (3.21) and of the Pre-entry students (2.97). The group of Other university students came closer to the group of Pre-entry students than to the group of Advanced students.
Of the whole group of test takers, 49 out of the 67 test takers perceived their language use in their task performance to be “3: Similar” or “4: Very similar” to their actual performances. No test takers chose “1. Not similar at all.” As confirmed in the mean score in Table 4.20 above, Advanced university students (N=9) were likely to perceive markedly stronger similarity between their task and real performances as all of them chose “4: Very similar” or “3: Similar” The responses of the groups of Other university students and Pre-entry students were spread more widely than those of Advanced university students. As to the groups of Other university students and Pre-entry students, the former perceived a stronger similarity than the latter as the mean scores of the test takers’ responses of these two groups showed.

It would therefore be reasonable to conclude that, overall, the test takers’ task performances reflect likely real world performance to an understandable extent. This finding was also affirmed by the results for the test takers’ perceived construct-irrelevant factors discussed in the next section, which implied that the test takers’ task performances were not unnecessarily affected by the task conditions. Methodologically, the designed role play tasks allowed the test takers to develop their performances from the opening to closing in their own ways, and also did not constrain their performances with time limitations and any pre-planned scenarios. This task design might have minimized the test takers’ perceived discrepancy between task and real performances.

In addition, although not conclusive due to the limited numbers of the test takers, Advanced university students (N=9), with experience of activities in the target settings combined with the high proficiency level and in fact estimated as the pragmatically most competent speakers in particular, were likely to be able to link their task performances with the reality more strongly. For the Advanced university students, experience of the language activities in the university domain may have allowed them to
imagine their performances in the real domain. Furthermore, their high proficiency of English, not surprisingly, would have enabled them to utilize their linguistic resources as they intended.

What may have possibly threatened the validity of the Extrapolation inference in this data would be the results where university students who were experiencing language activities in the real domain on a daily basis perceived a weaker similarity between the task performance and the reality. However, this case has been rejected for the current research, which found that the test takers, in particular university students, perceived a stronger similarity between their task and the authentic performances. The stronger task similarity recognized by the university students in the actual target domain would indicate the test takers’ task performances under the measurement of the current study can be extrapolated to the authentic domain.

The Extrapolation inference is also indirectly enhanced by the test takers’ awareness of pragmatic demands in university settings, which was integrated as the evidence for the Domain Description inference. The test takers’ awareness of the pragmatic features reflects the extent to which they consider the pragmatic features to be important and that they were aware of these features in the real settings in the university domain, to which the test takers task performances were expected to extrapolate, rather than in universal situations.

To summarize, the test takers’ test performances observed through the current assessment instruments, can to some extent predict test takers’ real world performance as the evidence obtained through the test takers’ perspectives was evaluated. Given the fact that the test takers performed on artificially created assessment tasks (Messick, 1994) which would inevitably constrain extrapolation of their performances, it is not unnatural that test takers report substantial discrepancy between task and real
performances. In the literature of pragmatics and pragmatic assessment, there have been few available criteria for the current study to allow it to draw any conclusions as to whether or not perspectives expressed by the test takers are satisfactory. To evaluate the obtained results by the researcher’s initial own expectation, the test takers’ perceived similarity between task and real performances were stronger than had been expected, given that the test takers’ performances were elicited under a test condition. The obtained test results, therefore can account for a substantial portion of their real performance in the domain of higher education, where they were aware of pragmatic demands in language activities.

4.6.2 To What Extent does the Test Administration Procedure Affect Test Takers’ Performances?

Construct-irrelevance (Messick, 1989) refers to factors which are irrelevant to the test constructs and which unnecessarily change test takers’ task performances. In general, possible construct-irrelevant factors underlying language assessment could be different depending on the types of assessment, design of the assessment or how the assessment is administered. The current study identified two possible construct-irrelevant factors rising from the task administration design. The study attempted to investigate their existence in the test takers’ task performance through the test takers’ own perceptions drawn in the post task interviews.

The two questions given to the test takers were concerned with designing the task administration procedure in the current study. Unlike the language use situation in reality, the test takers’ speaking performances in this study were observed in a testing situation where (a) the test takers dealt with six situations at the same time and (b) they performed the multiple role play tasks with the same role play conductor as an
interlocutor playing three different roles: that of a professor, of an administrator, and of a peer student.

Question (a) was associated with tiredness rising while performing the multiple tasks. As the current research employed the tasks eliciting oral productive performance, it was deemed necessary to investigate the extent to which engaging in six oral role play tasks reduced the quality of the test takers’ performance. With regard to question (b), unlike to reality the test takers in the role play situations encountered six problem-solving situations simulating university activities at the same time. Question (b) also rose from the logistical constrains applied to the current study. The dialogue role plays simulated situations where a student interacts with three different kinds of people (professor, administrator, and peer student) at university, but the roles of these different people were all played by a single interlocutor as a role play conductor for each test taker.

Therefore, it was deemed possible that these two factors unnecessarily affected the test taker’s performances. In reality, of course, it may not always happen that students would encounter six situations where they would need to handle pragmatic demands at the same time. In addition, the intended imaginary interlocutors (e.g., professor, peer student) should be different persons.

Table 4.21 and Table 4.22 below show the summaries of the test takers’ responses to the two questions about whether (a) tiredness from doing six tasks at the same time and (b) performing role plays with the same interlocutor changed their performance. The test takers were asked to express their perceptions dichotomously (Yes or No), but they were also allowed to comment in other ways if they could not explicitly decide on a yes or a no. The majority of them provided their views by answering “Yes” or “No.” Unclear answers and any answers that implied even a slight
possibility of presence of the construct-irrelevant factors were classified into “Unclear answer or possibly to some extent.” Table 4.21 below first presents the distribution of the test takers’ responses about question (a) regarding their perceived tiredness affecting their performance.

Table 4.21
Test Takers’ Perceived Tiredness Affecting their Performances

<table>
<thead>
<tr>
<th>Response category</th>
<th>Advanced university student (N=9)</th>
<th>Other university student (N=35)</th>
<th>Pre-entry student (N=23)</th>
<th>All (N=67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>7</td>
<td>30</td>
<td>19</td>
<td>56 (83.6%)</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Unclear answer or possibly to some extent</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>11 (16.4%)</td>
</tr>
<tr>
<td>Response total</td>
<td>9</td>
<td>35</td>
<td>23</td>
<td>67 (100%)</td>
</tr>
</tbody>
</table>

The distribution was summarized by the response categories and by the groups comprising Advanced university students, Other university students and Pre-entry students. The primary investigation was how the test takers as a whole responded to the question. Although the test takers’ responses are presented according to the groups, a group comparison was not the central question to this backing as there is little argument, theoretical assumptions and empirical evidence in the literature of what backgrounds of test takers perceived tiredness under the current test conditions.

The majority of the test takers as a whole (83.6%) and across the groups (seven out of nine for Advanced university students, 30 out of 35 for Other university students and nineteen out of 23 for Pre-entry students) did not recognize doing six role play tasks at the same time as unnecessarily changing their performance. A small number of test
takers in each group (two out of nine for Advanced university students, five out of 35 for Other university students and four out of 23 for Pre-entry students) provided unclear answers that may imply a potential irrelevant factor on their performances to a small or some extent. However, no test takers reported explicit presence of a construct-irrelevant factor generated from the tiredness as indicated by the category, “Yes.” The small numbers of test takers classified into “Unclear answer or possibly to some extent expressed” provided varied views and varied degrees of tiredness, which were difficult to generalize. Under any testing conditions, it would not be surprising that tiredness unnecessarily affects the quality of the test takers’ performance regardless of whether or not they perceived it as construct-irrelevance. The current test condition required the test takers to produce a substantial amount of spoken output. However, the substantial construct-irrelevance was not confirmed in the test takers’ perceptions in this regard.

The other possible construct-irrelevant factor was performing six role plays with the same interlocutor unlike reality, which may be awkward for the test taker and may consequently change their task performance unnecessarily. Table 4.22 below summarizes the test takers’ answers as the whole and by the groups.

Table 4.22
Test Takers’ perceptions of the Extent to Which Doing Six Role Plays with the Same Interlocutor Affected their Performances

<table>
<thead>
<tr>
<th>Response category</th>
<th>Advanced university student (N=9)</th>
<th>Other university student (N=35)</th>
<th>Pre-entry student (N=23)</th>
<th>All (N=67)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>9</td>
<td>30</td>
<td>20</td>
<td>59 (88.1%)</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1 (1.5%)</td>
</tr>
<tr>
<td>Unclear answer or possibly to some extent</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>7 (10.4%)</td>
</tr>
<tr>
<td>Response total</td>
<td>9</td>
<td>35</td>
<td>23</td>
<td>67 (100%)</td>
</tr>
</tbody>
</table>
Except for one test taker in the group of Other university students who explicitly recognized this construct-irrelevance and answered “Yes,” most test takers (88.1%) did not perceive their performances as being affected by the administration manner. A small number of test takers in the two groups of Other university students (four out of 35) and Pre-entry students (three out of 23) provided answers that were unclear or that implied potential effect on their performances to a small or some extent. All of the Advanced university students (N=9) explicitly rejected this construct-irrelevance expressed as “No.”

There are few criteria and no theoretical assumptions in the literature to conclude whether or not the test takers’ pragmatic performances were substantially changed by the possible construct-irrelevant factors as discussed above. However, it would be reasonable to conclude that the construct-irrelevance rising from the test administration concerned with the two possible factors was adequately minimized, given that very few test takers explicitly expressed the perceived construct-irrelevance for both factors. Perceived construct-irrelevance underlying some test takers’ unclear answers might be negligible or substantial enough to be classified into “Yes.” Regardless of how the unclear answers were interpreted, the numbers of test takers as a whole and across the groups rejecting these two kinds of possible construct-irrelevance were smaller than the researcher’s own expectation.

Substantial construct-irrelevance reflected in the test takers’ perception would be a threat to the inference of extrapolating the test takers’ task performances to the reality. However, under any testing conditions, it would be virtually impossible and unrealistic to completely eliminate every construct-irrelevant factor. What test developers can do is to make an attempt to identify possible factors irrelevant to the test construct and to investigate the extent to which they unnecessarily affect test takers’ test performances,
which may lead to a revision of the test in the future.

4.6.3 To What Extent Do the Performances on the Measure Relate to the Test Takers’ Own Assessment of their Performances?

The third backing for the Extrapolation inference was provided by the association between the test takers’ self-assessed performance and their estimated ability (rater-assessed performances and computed by Rasch analysis). Self-assessment is a non-test criterion and has been recognized as an alternative measurement for language testing research with its strengths of relatively lower cost for administration and of enhancing learners’ awareness and motivation. The utilization of self-assessment as a method to measure L2 speakers’ abilities requires careful interpretation of the results as the literature using self-assessments has provided varied results (Ross, 1998). Also, the outcomes of the test takers’ subjective judgements were unpredictable. Thus, the results may possibly undermine the validity inference. However, the test takers’ own perspectives about their performances were considered to be an important source of evidence because they were the real or prospective language users in the target domain. They self-assessed their performances on the following scale: “4: I could achieve this as well as I wanted to,” “3: Sometimes I couldn’t do this,” “2: I could not do this well even though I wanted to,” “1: I did not do this,” and, “0: Others.”

The test takers judged their task performances according to the pragmatic features as summarized in Table 4.23 below (the same features to examine the test takers’ awareness of pragmatic demands in section 4.2.2). Firstly, the results were summarized for descriptive statistics. For each self-assessment item, a few test takers chose responses “1: I did not do this,” or “0: Others,” which were excluded from computing the mean scores in Table 4.23 below and from the correlational analysis (to
be discussed later), because these responses did not necessarily indicate the degree of self-assessed achievement and were thus difficult to interpret.

Table 4.23
Mean Scores of Test Takers’ Self-assessment Items

<table>
<thead>
<tr>
<th>Feature</th>
<th>Advanced university student</th>
<th>Other university student</th>
<th>Pre-entry student</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Language use depending on social status</td>
<td>3.71 (0.49)</td>
<td>3.28 (0.68)</td>
<td>3.23 (0.61)</td>
<td>3.31 (0.65)</td>
</tr>
<tr>
<td>(b) Language use depending on familiarity (social distance)</td>
<td>4.00 (0.00)</td>
<td>3.33 (0.66)</td>
<td>3.57 (0.59)</td>
<td>3.50 (0.62)</td>
</tr>
<tr>
<td>(c) Language use to deliver intended meaning</td>
<td>3.13 (0.35)</td>
<td>2.79 (0.64)</td>
<td>2.45 (0.51)</td>
<td>2.72 (0.60)</td>
</tr>
<tr>
<td>(d) Language use for mitigation</td>
<td>3.78 (0.44)</td>
<td>3.12 (0.74)</td>
<td>3.04 (0.64)</td>
<td>3.18 (0.71)</td>
</tr>
<tr>
<td>(e) Repair to mitigate and to stress the information</td>
<td>3.56 (0.73)</td>
<td>3.06 (0.67)</td>
<td>2.87 (0.76)</td>
<td>3.06 (0.73)</td>
</tr>
<tr>
<td>(f) Repair for language mistakes</td>
<td>3.22 (0.67)</td>
<td>2.71 (0.69)</td>
<td>2.74 (0.75)</td>
<td>2.79 (0.72)</td>
</tr>
</tbody>
</table>

Table 4.23 reports the average scores of each group on each self-assessed item. Overall, Advanced students self-assessed their task performances much more positively than Other university students and Pre-entry students did across all items. The difference between the groups of Other university students and Pre-entry students was not noticeable nor did either one of the groups self-assessed more positively than the other.

Subsequently, a correlational analysis, which was the main analysis to address the sub-research question in this section, was performed to investigate relationship between the test takers’ abilities indicated by test takers’ logit values and their self-assessment. For this analysis, each test taker’s pragmatic ability in logit was estimated based on data excluding performance on the two criteria used for dialogue only.
(Engagement in Interaction and Turn Organization) as the self-assessment items did not explicitly cover the features described in the two rating criteria.

For each test taker’s self-assessed performance was computed by averaging his or her responses of 4, 3 or 2 on each self-assessment item in Table 4.23 above. Fifty-two out of 67 test takers chose either response of 4, 3 or 2 on all self-assessment items. The remaining test takers chose responses 1 (I didn’t do this) and 0 (Others) to one question (feature) or more, and these test takers’ data were excluded for the correlational analysis. Therefore, this correlational analysis was performed by using 52 test takers’ data. As the self-assessment data were not always assumed to be interval data, the Spearman correlation analysis was conducted. The results show the test takers’ perspectives about their pragmatic performances on the tasks and their abilities estimated by the Rasch analysis, were moderately correlated (rho=0.475, p< 0.01).

A possible threat to the correlational findings above would be that the construct of pragmatics covered in the self-assessment and that of the rating criteria were not completely identical. However, the construct coverage underlying the self-assessment was not deviated from the construct represented in the rating criteria as sub-features underlying the construct of pragmatics are inter-related (Roever et al., 2014). Another concern would lie in the nature of self-assessment, which relies on test takers’ subjective judgments, which could possibly underestimate or overestimate the test takers’ own abilities and which are not necessarily congruent with the raters’ views. Also, the raters judged the test takers’ performances according to the features described in the rating rubrics whereas the test takers judged their perceived degree of achievement without any descriptors although meaning of each question (each pragmatic feature) was explained to the test takers by the researcher himself. It would be the most ideal exercise to be able to elicit the judgments of the test takers and those of the raters exactly in the
same way, but it would be unrealistic to expect that the test takers were able to assess their own task performances themselves, retrospectively and analytically, according to the descriptors, as did the raters.

Although the current self-assessment involved the above-mentioned limitations, it was at least found to be clear that these two types of evidence (the test takers’ abilities estimated by the measurement and by their self-assessed performances) did not contradict each other. Thus, the obtained results of the correlation between self-assessed performance and rater-assessed performance would not undermine the Extrapolation inference. The correlational analysis showed that the test takers’ self-assessment and their abilities estimated based on the data reflecting the raters’ judgments, were moderately correlated as expected for non-test criteria (Weigle, 2010). If the correlation had been very weak or negative, the disagreement or discrepancy between the test takers’ perspectives about their pragmatic performances and the raters’ judgment according to the rating criteria would be a serious threat to the validity of the Extrapolation inference. The moderate correlation between these two variables was predicted in the literature, and at least the correlational strength was higher than the researchers’ own expectation, given the subjective nature of self-assessment. Little guidance has been provided in the literature on pragmatic assessment about how speakers engage in self-assessment and how they self-assessed their performance. However, the correlational results imply that the test takers’ perceptions adequately reflect what the designed test assesses and thus can be integrated to enhance the Extrapolation inference.

4.7 Utilization

The previous sections have demonstrated the defensibility of the test results as
representing L2 students’ oral pragmatic abilities for language activities at university through the inferences of Domain Description to Extrapolation. In the frameworks of argument-based approach to validation (Chapelle et al., 2008; Kane, 2006; Knoch & Elder, 2013), the chain of the inferences of an interpretive argument has been extended to the Utilization inference which covers the issues of decision-making based on the test results (Kane, 2013) and the consequences that the test and the test results lead to.

The current study, being in line with the proposed frameworks, examined the Utilization inference as to how the test results and the test are utilized for decision-making for educational purposes and how the decisions based on the test results and the test are beneficial for test users. To adequately address the Utilization inference, sufficient empirical evidence needs to be documented from a certain period during the actual operation of the assessment instruments. As the test results of the current study have not actually been utilized in reality, nor have the test instruments developed in the current study, evaluating evidence for the test use is not possible for the current study.

In this section, I will re-evaluate the obtained results as discussed in the previous sections (for the test score interpretation), through the lens of the Utilization inference. In spite of the dearth of evidence to directly address the Utilization inference at this stage, I, as a researcher and as the test instrument developer, have decided nevertheless, to remain responsible for providing future test users with some guidance towards their educational decisions. Thus, this section aims to address a portion of the Utilization inference by addressing the following research question:

What are possible cut-off score points for educational purposes?

I will also discuss what consequences could be expected from the decisions taken. It is
hoped that future studies extended from the current study will empirically refine the tentative conclusion drawn in this section about this research question.

4.7.1 A Possible Cut-off Point to Indicate Developed Pragmatic Ability for University Activities

Test scores of official English proficiency tests have been utilized as a proficiency indicator for administrators and pertinent staff at universities to assist them in judging whether or not students’ English levels are satisfactory for university study in English-medium contexts. Users of the test scores would also include English learners, English teachers, and coordinators of English learning programs who are informed by the test scores of the test takers’ English proficiency levels and who will, based on that information, then decide what actions to take. While test scores of those proficiency tests have played the crucial role of being an English proficiency indicator for university studies, it is debatable whether and to what extent those test users would be adequately informed from the official test scores and their rating criteria about L2 students’ pragmatic abilities utilized for university activities (Roever, 2011). The lack of empirical information available for test users about L2 students’ pragmatic abilities provided the current study with a strong rationale for developing the assessment instrument of L2 pragmatics, the test scores of which are useful for (a) inferences as to their pragmatic abilities for university activities and for (b) educational decisions based on the test scores.

The previous sections have evaluated the evidence required when discussing how useful the test scores were for inferences as to the pragmatic abilities of L2 students. This section, by reviewing the results gained in the previous sections, will discuss a possible cut-off point to differentiate the test takers for educational purposes.
Educational purposes would include decisions about providing educational treatments for students at university and pre-entry EAP courses to nurture their pragmatic abilities for university activities.

In order to seek possible cut-off points between the test takers, the two indices of the test takers’ pragmatic abilities computed by the Rasch analysis (logit and observed average) were summarized in Table 4.24 below. The categorization of the rank is more detailed than that of Table 4.17 (reviewed above for the Explanation inference). The test takers’ abilities were estimated by the Rasch analysis based on all of the collected data.

Table 4.24
_Distribution of Pragmatic Ability and Possible Cut-off Points_

<table>
<thead>
<tr>
<th>Rank</th>
<th>Logit</th>
<th>Observed average score</th>
<th>Advanced university student (N)</th>
<th>Other university student (N)</th>
<th>Pre-entry student (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 8</td>
<td>5.44 to 3.47</td>
<td>3.93 to 3.67</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2.76</td>
<td>3.49</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 19</td>
<td>2.45 to 1.27</td>
<td>3.40 to 3.04</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 34</td>
<td>1.09 to -0.67</td>
<td>2.99 to 2.49</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35 to 57</td>
<td>-0.72 to -1.88</td>
<td>2.41 to 2.02</td>
<td>10</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>58 to 67</td>
<td>-1.92 to -3.63</td>
<td>2.07 to 1.51</td>
<td></td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

_Notes._
- Rank refers to the test takers’ ability, based on the logit value estimated by the single Rasch analysis (using all the data).
- For rank 35th to 57th, the upper limit (35th) is the Pre-entry students estimated to be the most competent among the Pre-entry students, and the lower limit (57th) is the Other university student estimated to be the pragmatically least competent among the university students.

The test takers were ranked from 1 to 67 according to their estimated ability (in logit) by the Rasch analysis. Logit value is the index estimated by adjusting the impact of severity or difficulties of the other facets (rater, task, and rating criterion). To address the inference in the previous sections, the current study used the logit values as an indicator of test taker ability. The _observed average score_ is another index that The FACETS program provides and it refers to “the average rating received by the element”
In the case of the current study, each test taker received 90 raw scores as they did three dialogue tasks rated by six criteria, and three monologue tasks rated by four criteria. They were given the scores from the three individual raters. As described in section 3.5.3 in Chapter 3, there were four band levels for the scoring. Therefore, each test taker’s observed average was calculated by adding the observed 90 raw scores and dividing the result by 90, thus the range of the observed average score was 1 (minimum) to 4 (maximum).

The 67 test takers’ logit values and observed average scores were highly correlated with the strengths of 0.996 (p < 0.01) for Spearman\(^5\). Thus, both indices could be used as useful indicators of the test takers’ ability. The rationale for providing the observed average scores in Table 4.2 above was, firstly, so that the current study can interpret the scores according to the meaning of the four band levels defined in the rating rubrics and, secondly, because of the possibility that not all the real world users of the current assessment instruments (to be discussed later) may have access to the Rasch analysis and to the proper software to perform it. However, calculation of the observed average should be manually possible by the users themselves. Thus, I will make a suggestion regarding a cut-off score, considering the observed average scores, and will provide a justification for that suggestion.

The university students ranked 34th or higher could be evaluated as a possible indicator because none of their abilities were estimated to be below any of the Pre-entry students. However, the university students ranked from 20th to 34th marked their observed average scores below 3.00, ranging from 2.99 to 2.49. The rating was designed to evaluate test takers’ pragmatic performances in simulated university situations and classify the performances into four band levels according to the criteria. Generally,

\(^5\) Spearman correlational analysis was performed because the observed average score was not always assumed to be interval data.
across the six assessment criteria, the two lower scores (score 2 and score 1) were applied to test takers’ performances where negative features were more noticeable and outweighed positive features. The observed average scores in this range would therefore imply, according to the intention of the rating criteria, that their performances had a tendency to be more noticeable in the negative features, decreasing the confidence when deciding upon their level as the justifiable cut-off point.

If the borderline had been drawn at the observed average of 3.00, eight of the Advanced university students, those ranked from 1st to 8th with the observed average scores of 3.93 to 3.67, one Advanced university student ranked 9th with the observed average of 3.49, as well as ten Other university students ranked from 10th to 19th with the observed average scores of 3.40 to 3.06 would be considered to be ideal indicators of university students’ developed pragmatic abilities. The observed average scores of the eight (out of nine) Advanced university students were higher than 3.50, showing stronger evidence of the positive features. The values of the remaining one Advanced university student and ten Other university students were below 3.50 decreasing toward 3.00, implying that their positive features were more or less inconsistent but should be satisfactory.

Decision-making, in some cases, may be subject to the decision-makers’ circumstances. For example, the number of test takers who were judged as passing a test may depend on the decision maker’s capacity of accommodating the number of students, the decision maker’ personal belief or some other in-house or official criteria. In this sub-section, I have attempted to draw a cut-off line which would indicate the satisfactory performance, simply using the test takers’ test results and backgrounds as well as the descriptive information of the rating rubrics developed based on the empirical data in this study. I have regarded the nineteen test takers, those ranked from
1st to the 19th as having met the satisfactory level based on the test results and the rubric developer’s intention, I being the rubric developer. This tentatively drawn borderline was also in agreement with my personal view as a student in the target domain. The suggestion made in this section has motivated future studies to empirically address a more justifiable cut-off line for pedagogical purposes.

4.7.2 Expected Consequences

Following the discussion above, the pragmatic performances of the nineteen test takers consisting of all of the nine Advanced university students and the ten Other university students were considered to be satisfactory in the higher education domain. They were thus found to be both proficient and pragmatically competent. By contrast, the pragmatic abilities of the remaining 48 test takers, including 25 Other university students and all of the 23 Pre-entry students, remained questionable.

What was worthy of note was the fact that a certain number of Pre-entry students outperformed some university students. Based on the ranking of the test takers alone, this result may be positive and encouraging for those Pre-entry students. However, their observed average scores in Table 4.24 in the previous section above should not make the remaining 48 test takers of both the university students and the Pre-entry students optimistic about their pragmatic performance.

As the lowest ranks in Table 4.24 were dominated by Pre-entry students, and as it was evident in their observed average scores, all being below 2.50, their abilities could be differentiated from the majority of university students’ and could be regarded as still being in the process of developing towards reaching the university levels. To my best understanding, teaching or diagnosing pragmatics and self-guided learning for pragmatic ability have been very rarely practiced at pre-entry level EAP courses or
equivalents. The obtained results, including the test takers’ perceived demands of pragmatics in language activities, their self-assessed performances, and the test results led me to recognize most vividly, the significance of offering to those pre-entry students who are heading towards full engagement with their future university studies, opportunities of nurturing their pragmatic abilities. Thus, it is expected that test scores generated from the current study and the performance-based instrument provide positive washback for students’ learning and motivation for improving their pragmatic abilities.

The opportunities for learning pragmatics ought to be available for the 25 post-entry university (Other university) students whose performances tended to be or were more negative than positive. In particular, the observed average score of the university student who was ranked at 57th (the least pragmatically-competent university student) was quantitatively far too distant from those of the Advanced university students’ even though they were all engaged in language activities at university in real circumstances. It is thus hoped that in particular, those whose observed scores were close to 2.00 are provided with opportunities (e.g., at a post-entry EAP program), so that they can be aware of their current levels of pragmatics. The limitation in their pragmatic ability was evident in the illustrated discourse performances when compared with higher levels of performances in section 4.3.2. These empirically highlight our motivation for providing learning opportunities for those students regardless of post-entry or pre-entry levels, as reflected in recent practices of workshops on teaching L2 pragmatics (e.g., Roever, 2012).

4.7.3 Towards the Use of the Test Instruments Developed in the Study

The Utilization inference is concerned with how useful the test scores are for making educational decisions such as selecting a particular number of test takers,
classifying test takers into groups, and determining the appropriate amount of educational treatment for the test takers. The current study assumed that the test scores generated in the measurement designed in this study were useful for language educators and learners to take educational actions. In reality, however, to implement a performance-based assessment of pragmatics for educational treatments to nurture L2 students’ pragmatic ability, it would require weighty consideration about human as well as financial resources. In the subsequent section, I will discuss the possibility of utilizing monologue assessment as an alternative to dialogue assessment and for English teaching practitioners to administer the assessment activities more practically, while maintaining a certain degree of construct coverage. This issue is, as argued in Chapter 2, particularly important because it is possible that impractical instruments cannot be utilized before the Utilization inference is considered.

4.8 Practicality

The aim of this section is to address the question of how the designed assessment can be implemented practically, while avoiding construct under-representation of pragmatics (McNamara & Roever, 2006). A strength of oral performance-based assessment on dialogue modality such as the instruments in this study, is to cover a broader construct of pragmatics for assessment. Its shortcoming is to increase the quantity of necessary resources, which decreases instrument practicality. If both dialogue and monologue assessment of pragmatics can yield similar test results, then monologue instruments could serve as an alternative to dialogue instruments.

Monologue instruments were assumed to be more practical, but it remains uncertain whether monologue instruments can elicit and assess the construct of pragmatics in the same way as dialogue instruments can do. This section addresses the
following research question:

How and to what extent can monologue tasks serve as an alternative to dialogue tasks?

Specifically, this research question was addressed by investigating how and to what extent the test takers’ pragmatic ability under the dialogue condition and their ability under the monologue condition are correlated.

Correlational analyses were conducted using the test takers’ pragmatic abilities estimated by separate Rasch analyses using the dialogue data and using the monologue data respectively. As described in section 3.5.3 in Chapter 3, both the dialogue and the monologue performances were assessed by the four shared criteria (*Social Actions to Achieve the Communicative Goal, Facility with the Language, Language Use to Deliver the Intended Meanings, Language Use for Mitigation*). In addition, two more interactionally-orientated criteria (*Engagement in Interaction and Turn Organization*) were used as assessment criteria exclusively for the dialogue performances.

Proceeding to the correlational analyses, the test takers’ pragmatic abilities (expressed in logit) were estimated by three separate Rasch analyses, based on: (a) all dialogue data including the data on two criteria (*Engagement in Interaction and Turn Organization*), (b) the dialogue data excluding the data on these two criteria, and (c) all monologue data.

The results showed that irrespective of including or excluding the data on the two interactional criteria (*Engagement in Interaction and Turn Organization*) in the dialogue data, the test takers’ estimated abilities under both the dialogue and the monologue conditions were highly correlated in addition to being statistically
significant. The correlational strength between the test takers’ abilities under the
dialogue condition (estimated including the data on the two interactional criteria) and
those under the monologue condition was high ($r=0.937$, $p<0.01$; $r^2=0.878$). The similar
result was obtained for the case between takers abilities under the dialogue condition
(estimated excluding the data on the two interactional criteria) and those under the
monologue condition ($0.939$, $p < 0.01$; $r^2=0.872$). The results show about 88% of the
shared variance between the two modalities of performances. It was therefore suggest
that the pragmatic ability observed under the monologue condition were constructed
largely by the variance, which also accounted for the pragmatic ability observed under
the dialogue condition.

Subsequently, the strength of the correlation between the Rasch-estimated
abilities (in logit) under the two modality conditions on each rating criterion were
investigated. Table 4.25 below summarizes the results.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Pearson’ $r$</th>
<th>Pearson ($r^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Actions to Achieve the Communicative Goal</td>
<td>0.853, $p &lt; 0.01$</td>
<td>0.728</td>
</tr>
<tr>
<td>Facility with the Language</td>
<td>0.983, $p &lt; 0.01$</td>
<td>0.966</td>
</tr>
<tr>
<td>Language Use to Deliver the Intended Meanings</td>
<td>0.931, $p &lt; 0.01$</td>
<td>0.867</td>
</tr>
<tr>
<td>Language Use for Mitigation</td>
<td>0.889, $p &lt; 0.01$</td>
<td>0.790</td>
</tr>
</tbody>
</table>

The highest correlation coefficient was identified for *Facility with the Language*
($r=0.983$), followed by the coefficient for *Language Use to Deliver the Intended
Meanings* ($r=0.931$). The other two criteria, *Social Actions to Achieve the
Communicative Goal* and *Language Use for Mitigation*, showed slightly lower
correlation coefficients ($r= 0.853$ and $r=0.889$, respectively), which retained a strong
association between the abilities seen under the dialogue and the monologue conditions.
All correlation coefficients were statistically significant (p < 0.01). The results indicate that abilities exercised under the dialogue and monologue conditions were explained by approximately 72.8% (the lowest \( r^2 \) for *Social Actions to Achieve the Communicative Goal*) to 96.6% (the highest \( r^2 \) for *Facility with the Language*) of the shared variance. The test takers’ abilities assessed by the two rating criteria, *Engagement in Interaction* and *Turn Organization*, were not targeted since they were only used for dialogue performances.

Overall, it was implied that the test takers’ abilities on each of the assessment criteria under one task modality condition can, to a large degree, be predicted by those under the other. No assessment criteria generated disproportionately weaker correlations between the two modalities than the other criteria. The quality of the test takers’ performances on two different modalities would not be exactly the same. However, pragmatic features underlying the four assessment criteria shared for both modalities were identified in the test takers’ performances on both modalities, and these four criteria functioned adequately enough to discriminate the test takers. Regardless of whether performing under the dialogue condition or the monologue condition, pragmatically competent test takers were able to determine the amount of output and social actions that they found appropriate in the given context, and were able to control their language use adequately enough to serve their intended communicative goal.

As confirmed by the high correlation between test takers’ abilities estimated from the dialogue data and those from the monologue data, both types of assessment instruments functioned generally in a similar manner, in terms of ranking L2 speakers according to the pragmatic abilities. It would then suggest a possibility that monologue tasks, which are less resource intensive, can serve as an alternative to dialogue instruments in assessment of L2 pragmatics which the current study defined. The
correlational analysis indicated a strong association between performances on the two modalities with a large shared variance accounting for these types of performances. Therefore, with regard to test takers’ overall pragmatic competence defined in the criteria for this study, it is very likely that L2 speakers who perform positively under the monologue condition will perform under the dialogue condition just as positively.

Balancing instrument practicality and construct coverage of pragmatics involves a trade-off. As reviewed in Chapter 2, multiple-choice tasks and DCTs, for example, can offer practical advantages in their test implementation and in gathering a large amount of data, but they elicit a limited aspect of pragmatics isolated from interaction. Prioritizing instrument practicality may result in oversimplifying the construct of pragmatics, namely, construct under-representation. This may ultimately underestimate test takers’ abilities and limit a conclusion about their pragmatic abilities as utilized in reality. In contrast, performance-based tests can elicit their online performances reflecting broader aspects of their abilities. However, the implementation essentially involves higher resource intensity. Therefore, the trade-off makes us aware of what will be gained and what will be lost in operationalizing assessment instruments.

Monologue assessment of pragmatics has an advantage in reducing resource intensity because they could elicit test takers’ extended discourse data without using a conversation partner (an interlocutor). A downside is that diagnostic information of test takers’ abilities specified in the two interactionally-orientated criteria (Engagement in Interaction and Turn Organization) would be lost because the test takers would not have performed interactively with an interlocutor. Needless to say, the dialogue assessment of pragmatics would allow for providing diagnostic information more comprehensively, including that on these interactional criteria, at the expense of increased resource intensiveness in its test implementation.
As Bachman & Palmer (1996) suggests, practicality rests on available resources (e.g., human resources) relative to the required resources. Because the availability of resources and what to prioritize differ depending on individual assessment contexts, it would not be possible to provide a universally applicable solution for the trade-off between instrument practicality and construct coverage of pragmatics. To summarize, the results in this section could suggest that, except in cases where diagnostically informing test takers’ abilities on *Engagement in Interaction* and *Turn Organization* is crucial for the assessment, the monologue assessment could serve as an alternative to the dialogue assessment. In a case of simply separating and ranking test takers according to their oral pragmatic abilities, the monologue assessment can be an alternative to the dialogue assessment, and the advantages of using the monologue modality would outweigh its shortcomings.

### 4.9 Chapter Summary

This chapter has discussed how the evidence and the backings obtained from the evidence allowed for the validity inferences of Domain Description, Evaluation, Generalization, Explanation, Extrapolation, and Utilization. Through the extensive validation, the study suggests that test scores produced under the current measurements are useful for inferences as to L2 students’ pragmatic abilities in the target domain and are useful in making educational decisions. The practicality issue was also integrated into the discussion so as to explore the practical implementation of the assessment while avoiding the construct under-representation. The findings allowed the study to draw the conclusions, to provide the implications, to identify the limitations and to make the suggestions for the future research, which I will discuss in the subsequent chapter.
Chapter 5 Conclusion

In this chapter, I will summarize the key findings in the study and discuss implications, after which I will discuss limitations and provide suggestions for future studies.

5.1 Summary of the Key Findings

This study set out to develop an oral pragmatic assessment instrument for L2 students’ language activities at English-medium university. It evaluated the multiple sources of evidence as well as the different phases of research activities to structure an argument for the extensive validation and for addressing the practicality issue.

For the Domain Description inference, which examined the extent to which the assessment contents reflect L2 students’ oral pragmatic abilities for use in university settings as the target domain, the practice of the task development and the test takers’ perceptions were evaluated. The test task specifications were made and refined through the multiple phases of piloting, and the domain experts’ perspectives were integrated into the test tasks. The study also confirmed the test takers’ substantial awareness of pragmatic demands in language activities at university.

In the phase of validation for the Evaluation inference, which examined the extent to which the procedure for assigning test scores is appropriate, was addressed by the practice of the rating and of the design of rating criteria with the descriptors developed to differentiate the test takers and the quality of the rating. The raters showed similarities in rating as indicated in the inter-rater reliabilities (correlation), although the strength of correlation varied depending on the pair of raters. The Rasch analysis showed that the raters’ internal consistency in rating and the severity difference in rating
were within the expectation of the literature. It was also confirmed that the raters operated the assessment band levels appropriately.

The Generalization inference examined to what extent the assessment yields test results consistent across assessment contexts: task administration conditions, raters, and task sets. High test reliability indicated by Cronbach alpha and the high comparability of the two task sets were confirmed to support the Generalization inference. Additionally, the study reviewed the test administration conditions consistent for all test takers.

After examining the appropriateness of converting the test takers’ performances into test scores (the Evaluation inference) and the extent to which the test results can be generalized (the Generalization inference), the study investigated the extent to which the test results are informative of the test takers’ pragmatic abilities in a university domain (supported by the Explanation and the Extrapolation inferences). The phase to address the Explanation inference involved the group comparison and the correlational analysis. The three groups of the test takers differed generally as hypothesized according to their pragmatic abilities although the two less competent groups overlapped to some extent in their pragmatic abilities. The test takers’ proficiency level and their estimated pragmatic ability were strongly associated. The test takers’ utilization of the task preparation time supported the group difference according to the pragmatic abilities. The results were used as the backings for the assumption that test takers’ test results can be attributed to the construct of pragmatic abilities utilized for language activities at university.

In the phase of addressing the Extrapolation inference, the study examined the extent to which the performances observed in the tasks are indicative of their real pragmatic performances in the target domain. The study investigated and evaluated the evidence obtained from the test takers’ perceived similarity between their task and
authentic performances, their perceived construct-irrelevance in their task performances as well as their self-assessed performances. The test takers expressed similarities between their task and authentic performances. Advanced university students especially perceived a stronger similarity between the test and the target language use domain. The two possible construct-irrelevant factors that the current study recognized were not substantial as suggested in the test takers’ perceptions. The test takers’ self-assessed performances were moderately correlated with their pragmatic abilities estimated by the Rasch analysis. Considering the subjective nature of self-assessment, the strength of the correlation was higher than the researcher’s expectation.

Overall, it was suggested that the test instrument evaluated in this study measures L2 students’ oral pragmatic abilities integrating discursive abilities. Test results that the current assessment instrument generates are seen as useful for inferences as to L2 students’ pragmatic abilities. As the test instrument has not been operated by users in real circumstances, the validity inference could not be directly extended to the Utilization inference, which is concerned with what decision can be made based on the test results and what consequences the decision and the test itself yield. Addressing the consequential aspect of validity requires a certain period of test operation and empirical evidence gained from test users. Thus, as an effort that could be made at this stage, the current study has proposed a possible cut-off point for educational decisions and discussed the expected consequences.

For the practicality issue, the possible utilization of monologue instrument for pragmatic assessment was discussed. It was revealed that the test takers’ pragmatic ability under the monologue condition were highly correlated with their pragmatic ability under the dialogue condition across the rating criteria shared for both modalities. This suggested a possible utility of discursively-orientated assessment of pragmatics.
under a monologue condition. Although information about two internationally-orientated criteria (Engagement in Interaction and Turn Organization) are not available from monologue assessment, monologue assessment can be an alternative to dialogue assessment in ranking L2 students according to their oral pragmatic abilities defined in the current study.

5.2 Implications

The current study has several implications: educational implications, theoretical implications, methodological implications, and implications for research on testing of L2 pragmatics, which will be discussed in the following sections.

5.2.1 Educational Implications: Significance of Practice of L2 Pragmatic Assessments for University Activities

Pragmatics is a vital part of communication as it is recognized as an integral part of communicative competence in applied linguistics. Pragmatic abilities are what we use to achieve our communicative goals in certain contexts. This is true in the higher education domain as shown in the test takers’ perceptions about their awareness of pragmatic demands in university activities. Currently, pragmatics is not a central consideration in the English proficiency tests (Roever, 2011; Roever et al., 2014) and very few practices of assessment of L2 pragmatics have been documented in EAP classrooms (Youn, 2013). This means that L2 students already enrolled in and those aiming to enter English-medium education are insufficiently informed of their readiness for communicative settings where they handle pragmatic demands at university. The data analyzed and discussed in the current study suggest (a) that those students were, in fact, clearly aware of the pragmatic demands in university activities, (b) that a
substantial number of university students struggle with handling it (as reflected in their self-assessment), and (c) that their pragmatic abilities still reserved much room for improvement (as reflected in their task performances and in their abilities estimated by the Rasch analysis). For this reason, the current study has re-confirmed meaningfulness of L2 pragmatic assessment for those in English-medium education.

The study estimated the test takers’ pragmatic abilities as defined in the current rating criteria. The study has revealed that the Advanced university students and some Other university students being advantaged by both higher proficiency level and experience at the target domain were found to be pragmatically highly competent. However, the readiness of the majority of Other university students and all of the Pre-entry students for handling pragmatic demands in language activities at university is questionable. The current study highlights the significance of assessing L2 students’ pragmatic abilities as well as that of providing educational treatments for those with insufficient abilities at both pre-entry and post-entry stages.

All of the test takers in the main study, as argued above, should be adequately proficient in English as they were admitted to pursue English-medium education or near university-entry level. However, the difference in pragmatic performance between the highest and the lowest levels among those test takers and the separation of the test takers according to the pragmatic abilities were noticeable. A substantial number of university students and all of the Pre-entry students produced pragmatic performances where negative features outweighed positive features. The insufficiently controlled pragmatic performances of those test takers were revealed quantitatively and qualitatively. The identified difference in pragmatic performances among the test takers stresses the significance of offering educational treatments to L2 students and pragmatics assessment, in order to give awareness raising to students.
As disused in Chapter 4, some particular test takers’ interactional behaviours when handling pragmatic demands may reflect their lesser degree of confidence and readiness for university settings. Those Pre-entry students’ interactional behaviours included sticking to the task prompts, a tendency to deliver what they might have memorized, turn-taking, and insufficient social actions, which affected the interlocutors’ interactional patterns (Crystal, 1997; Roever, 2011). Based on the pragmatic performances discussed both qualitatively and quantitatively and upon the interlocutors’ voluntary comments, the readiness of a substantial number of Other university students and all of the Pre-entry students for university situations where they handle pragmatic demands is questionable. Put differently, the pragmatic abilities of these students still reserved substantial room for improvement before entering university.

The significance of pragmatic assessment is also highlighted for post-entry English education programs by the pragmatic abilities of the Other university students. Hypothetically, Other university students exceeded Pre-entry students in the assessment as discussed in section 3.4.1 (Chapter 3). The overlap in the pragmatic abilities between the groups, Advanced university students, Other university students, and Pre-entry students, was deemed possible, but the revealed width of the overlapped ability range between Other university students and Pre-entry students was outside of the researcher’s expectation. The Pre-entry students might have unexpectedly performed better than they were usually able to. Whatever the reasons, it has become clear that being a university student in an English-medium context does not necessarily mean their pragmatic abilities are sufficient for language activities in that domain.

The empirical evidence would contribute to the literature pertaining to L2 pragmatics in view of how L2 students at English-medium university are differentiated in L2 pragmatics as defined in the study. The findings discussed in the study are also
directed to English language education for university study. The test takers’ perceptions elicited in the post task interviews suggests that handling pragmatic demands in university activities matters to L2 students. In addition, it challenges them, as reflected in the test takers’ self-assessed performances. Therefore, the test takers’ awareness of pragmatic demands, the challenge that they perceived in handing pragmatic demands and their pragmatic abilities discussed qualitatively and quantitatively, adequately support the meaningfulness of pragmatic assessment and educational treatment practiced for both pre-entry and post entry students.

Although the evidence obtained and the discussions developed in the current study could not be extended to how pragmatics should be taught to students, it is suggested that students are able to be aware of pragmatics for their communication. Therefore, at least it would be possible in an EAP course to provide a session where students self-assess their own performance for their awareness raising or a session where students are shown comparative discourse performances to discuss features that they note and why they think the features are positive or negative.

5.2.2 Theoretical Implications

As reviewed in Chapter 2, pragmatics is a multi-faceted construct. The literature has not reached a consensus of definition of pragmatics and respective studies have their own focus to investigate. It goes without saying that a single assessment instrument is not able to elicit and assess every single aspect of pragmatics discussed in the literature. In the current study, the target pragmatic features to assess were limited to the assessment criteria and sub-features and the target L2 speakers were limited to university students and pre-entry students in English-medium education (and therefore relatively high in proficiency compared to general ESL and EFL learners. The findings
of the current study, however, confirmed what has been reported in the literature and would contribute to refining a picture of L2 speakers’ pragmatic performance at different levels. Firstly, as hypothesized in the literature, the current study has shown that the test takers’ proficiency and the experience of language activities at the target domain (English-medium university) aided their pragmatic performances. Secondly, as reported in Youn (2013), the criterion measurement produced by the Rasch analysis in the current study suggested, overall, that managing interaction (as measured in *Engagement in Interaction* and *Turn Organization*) is relatively less demanding for L2 students at the English-medium university level or the pre-entry level, whereas the criteria related to language use (e.g., *Language Use for Mitigation* and *Language Use to Deliver the Intended Meanings*) are more challenging for them when handling pragmatic demands.

At more micro levels, the test takers’ several features such as pre-expansions before the main topic and the main action and use of pragmatic expressions (e.g., modal verbs “could” “would,” and expressions using bi-clausal structures) were also consistent with the literature which recognizes these as an indicator of advanced features. As expected in the literature, pragmatically competent test takers in the current study were able to use these types of expressions (though not limited to these). They were also able to manage the opening as argued in the previous studies such as Al-Gahtani and Roever (2012) and Youn (2013), but furthermore, the current study has identified the quality difference in speakers’ pre-actions, in addition to whether or not the pre-actions are taken.

English language abilities of the test takers (including the Pre-entry students) in the main study were not at an elementary level of EEL learners. Therefore, it is not surprising that they were able to produce a certain amount of English-language output
and to engage in interaction to some- to a large extent. However, the current study has captured their language use and interactional behaviours, which indeed differentiated the test takers. The features and sub-features discussed in the current study would thus serve to further refine the picture of L2 students’ pragmatic performances that the accumulated insight of previous studies showed. Also, the findings of the test takers’ performances can provide guidance of how and why observed performances of students at English-medium higher education levels can be judged as more positive or negative when their performances are examined for an assessment context.

5.2.3 Methodological Implications

Methodological implications can be provided from utilization of the self-assessment in the current study. The self-assessment of test takers’ own pragmatic features in their oral extended discourse was one of the few attempts, although a self-assessment for speech acts was conducted in earlier studies such as Hudson et al. (1995). The test takers’ self-assessed performance in the current study was moderately correlated with their ability estimated based on the scoring of three trained raters. A moderate correlation between self-assessment as a non-test criteria and test score may not be surprising (Weigle, 2010). However, it also highlights a consideration for utilizing self-assessment for both research purposes or for the purpose of seeking evidence for test validation. The self-assessment as a device itself requires researchers’ careful interpretation and handling it with caution (Ross, 1998). Subjective judgement of research participants may be beyond the control of the researcher, but at least the meaning of pragmatic features to self-assess must be clearly delivered to the participants to elicit the target perspectives from them. Self-assessment should have the potential of providing a valuable source of evidence for the research as well as for motivating and
awareness-raising for L2 learners in nurturing pragmatic abilities for language activities. The obtained results of the self-assessment in this study would not allow the study to draw a strong conclusion more than “an encouraging conclusion.” However, at least it has become clear that the test takers’ perceptions did not contradict the test results that the current study produced. This would suggest the potential of utilizing self-assessment for pragmatic research and for educational purposes, although simultaneously further elaboration and careful development of a self-assessment instrument are necessary in future studies.

5.2.4 Implications for Research on Testing of L2 Pragmatics

Testing of L2 pragmatics is a relatively new enterprise (Roever, 2013), compared to the other components of language competence and is still under-explored. First and foremost, the current study has highlighted the significance of testing of L2 pragmatics and its responsibility in contributing to better practices of pragmatic assessment in educational contexts. Because students and English educators, both of whom are involved in university activities, have rarely been informed of students’ pragmatic abilities for language activities at university, the research on testing of L2 pragmatics has played and will continue to play a vital role in realizing the assessment practices. The current study re-confirmed the rationale for practicing such assessments at both pre-entry and post entry levels. To realize this, the accumulated efforts in the area of testing pragmatics from the earliest studies to the most recent studies should be fully utilized. At the same time, the study suggests that the pragmatic features including the discursive features focused on in the current study are measurable. This in turn suggests that testing of pragmatics would be able to better inform teachers and L2 students of L2 students’ pragmatic abilities beyond offline knowledge of speech acts,
which was predominantly targeted in the past research.

In response to the insights which the literature in L2 pragmatic assessment provided, the current study attempted to explore test construct of L2 pragmatics expanded from traditionally forced speech acts to more interactional features including conversational management. As Youn (2015) has asserted, more research on testing the expanded construct of pragmatics will be needed. Reflecting a broader construct in an instrument is indeed a strength for an assessment because it serves to avoid construct under-representation. This strength of a broader construct coverage designed in the assessment can simultaneously threaten the instrument practicality. An assessment instrument designed to elicit and assess a broader construct inevitably involves greater resource (e.g., cost, human resource) in order to implement the assessment. Therefore, equally importantly, consideration for practicality should be taken into account at the same time in respect of available resources. The findings of the current study have addressed the challenges that the social dimension of language testing faces (McNamara & Roever, 2006). The demonstration of the current study is expected to encourage further studies to explore an appropriate balance between practicality and the construct coverage of the assessment and to realize the goal of “keeping it social but practical” (McNamara & Roever, 2006, p. 63) for test users and test practitioners of oral pragmatic assessment.

In addition to the implications above, this study has also exemplified the validation in the argument-based framework, which allows us to inform the literature and test users of what the test scores mean and how useful they are and why the proposed score interpretation and test use can be justified. The argument-based approach to validation has been demonstrated in a small number of studies in testing of L2 pragmatics (Roever, et al., 2014). As reviewed in Chapter 2, two previous studies,
Roever, et al. (2014) and Youn (2013) adapted this approach comprehensively. The current study is in line with these previous studies in that the argument-based framework is applicable to and can make significant contribution to testing L2 pragmatics.

5.3 Limitations and Suggestions for Future Research

Whilst this study provided valuable implications, limitations should be acknowledged. Firstly, the test construct of pragmatics focused in the current study did not cover every single feature discussed in the related literature, although the attempt was made to expand the test construct. Essentially, it should not be possible for a single type of task to elicit every aspect of pragmatics. The target assessment features and the criteria were selectively defined for a purpose of assessment as they were identified as discriminating the test takers. In addition, for the dialogue tasks, their performances were elicited in conversations with the native speakers (interlocutors). This reflects a part of reality but not all of the reality of university domain, where L2 students address other L2 students in their activities. In a scenario where an imaginary interlocutor’s English competence was lower than the test taker’s, interactive features such as providing support for the interlocutor (Galaczi, 2013) may have been able to be elicited and integrated into the criteria if the features are discriminating the test takers.

With regard to the procedure of administering the test, the current study did not set any time limitation for the test takers to perform the tasks. The target test construct integrated their ability to manage the extended discourse from the opening to the closing. Therefore, terminating performances within a time limit was deemed unnatural and, most of all, distorting the essence of pragmatic performance. However, time limitation is a practical issue in the administration of a test. Therefore, it would be
interesting in a future study to investigate how time limitation affects the quantity and possibly the quality of the test takers’ language production, compared to those tested with no time limitations.

Another limitation also relates to the test takers’ performances which may be affected by the test methodology. The current study did not inform the test takers of what the test attempted to measure and what the specific rating criteria were. With regard to this, it has been little reported in the previous studies as to whether or not they provided assessment criteria for the research participants beforehand. As practiced by Youn (2013), where the data-driven rating criteria were developed after her test takers’ task samples were collected, the current study did inform the test takers about the criteria. In contexts of high-stakes official tests of English, test takers are, in general, supposed to prepare for the tests by referring to the assessment criteria publically available for them. Therefore, it can be suggested that a future study should investigate how the test takers with explicit knowledge of the assessment criteria could perform differently or similarly, when compared with those without the knowledge. This issue would enhance the Utilization inference which is concerned with the utilization of the test instrument.

Further research may also investigate the extent to which the difference of interlocutors affects test takers’ performances (Brown, 2005, Filipi, 2014) in the case of oral pragmatic assessment. The current study employed multiple interlocutors. The assignment of the trained interlocutors to each test taker was determined according to the availability of the test takers and the interlocutors as well as of rooms suitable for the administration of the test. Therefore, the number of test taker participants assigned to each interlocutor was not equal.

It would be challenging to isolate and extract pure interlocutor effect in the
performance-based assessment that the current study conducted, given that several elements (task, interlocutor, and rater) were involved in the process of eliciting test takers’ performances and converting them into test scores (McNamara, 1996). As Brown (2005) demonstrated, an attempt to identify interlocutor effect in the test scores would require research carefully designed for its research question, which the current study could not realize. However, the present study can draw a tentative conclusion on a ground that the interlocutor effect was minimized as reflected in the correlational analyses results showing that the test takers’ performances under the dialogue condition with the presence of an interlocutor and those under the monologue condition without interlocutor were highly correlated. Unlike a generic task requiring test takers to express an opinion about a given topic and then discuss it with an interlocutor, which allows the test takers to develop their speaking performances in numerous ways, performances on a contextualized role play for the current study are assumed to be more constrained and thus predictable in terms of what actions test takers are to take in the given communicative goal. For the current study, even though what actions to take and when to take them may vary depending on the test taker (as the task design allowed the test takers to), the range of possible social actions was found predictable in the series of the pre-pilot and the pilot studies. In addition, the researcher wrote the instructions to the interlocutors in what to say and how to say it. Furthermore, for the interlocutors, the number of dialogue tasks to conduct was limited, which should be far less demanding for the interlocutors than preparing for test takers’ performances showing numerous ways on a number of topics.

Finally, it should be acknowledged that the backings to structure and to support the validity argument were limited to those discussed in Chapter 4 and there would be several other backings that enhance the validity for the assessment. For instance, an
investigation of raters’ cognitive process can be a study to ascertain the extent to which they are focusing on the assessment features as the test developer(s) intended. Rating serves to enhance or undermine the overall validity beyond being concerned with the Evaluation (Scoring) inference as suggested by Knoch and Chapelle (in press). A possible method would be to provide sessions with a rater where audio-recorded samples are played and the rater is allowed to stop and resume the audio as he or she wishes in order to verbalize his or her thoughts leading to the rating judgments. In the current study, the raters left notes in their rating work sheet (Excel files) on a volunteer basis giving reasons for their assigning scores, although they were not asked to do so. So far as the researcher could ascertain from these notes and comments, he concluded that the raters, even though their assigned scores were not always the same, adequately understood the intentions of the rating rubrics. The method suggested above would enable us to investigate rating processes more thoroughly and comprehensively. It would also help to identify the extent to which as well as the reasons why the designed rating is demanding.

The Extrapolation phase relied on the test takers’ perceptions and it was concluded that overall, their task performances allow us to predict their authentic performance in the target domain. A possible study to strengthen the Extrapolation inference would be one investigating and analyzing test takers’ authentic performances in naturally occurring settings, if this investigation was ethically and logistically allowed. Alternatively, it would be possible to review the recorded or transcribed role play performances with the test takers and to ask them to verbalize specifically in what way and why their task performances are similar to their real performances.

The Utilization inference could not be directly addressed, although a suggestion was made from the rating rubrics developers’ perspectives for a possible cut-off score
point to judge whether or not L2 students’ pragmatic abilities are satisfactory. This tentative conclusion was reflected in the perceptions of Other university students and Pre-entry students about their own task performances. These two groups self-assessed their task performances substantially negatively than the group of Advanced university students. Thus, the justification for the current conclusion of the cut-off score point has been made based on the data that the current study collected. The overall picture of L2 students’ pragmatic ability would be further refined by the data of a broader population of students at university. In addition, a more justifiable conclusion could therefore be drawn from perspectives of those involved in authentic university settings (e.g., English teachers, academic staff, administration staff, and students). Their perspectives could be elicited by some possible methods, including their intuitive judgments of the role-play samples collected in the current study. Another possible method would be that where they act as a role play interlocutor in role plays in a new data collection as well as a rater, scoring the newly corrected samples based on their intuitive judgments. Whatever the methods, additional data of the pragmatic performances and the perspectives of those in authentic settings in the academic domain (at both post and pre-entry levels) should be crucial for refining the cut-off score point.

The Utilization inference is concerned with issues broader than setting a cut-score discussed in this study. They include issues regarding how useful the test instruments are and how beneficial the decision made based on the test scores are for stake-holders of the test. The consequential aspect of validity will be addressed in future studies for which, again, would require the perspectives of those in the authentic settings at university.
5.4 Concluding Remarks

Addressing the challenge that testing of L2 pragmatics faces (Kasper & Ross, 2013; Roever, 2011; Roever & McNamara, 2006; Youn, 2015) involved multiple phases of research activities, a number of research participants contributing to these activities in each phase, and adaptation of multiple types of analytic approaches. It is hoped that this study will contribute to the enhancement of L2 students’ pragmatic abilities for university study and to the continued discussion in the literature concerned with L2 pragmatics. For L2 students, the findings of this study are expected to provide indicators for them to move forward from their current levels of pragmatic abilities, with an assistance of assessment practices in educational contexts as proposed in the study. For the literature pertaining to testing of pragmatics, which has reached “the third generation” (Roever, et al., 2014, p. 51) encompassing discursive features, the current study will serve to move the discussion forward to the next generation, where assessment of the broader construct of pragmatics can be realistically feasible.

This study was inspired by the previous and the recent empirical studies and by the arguments reviewed in the literature. Although the focus of pragmatic features, target learners, and target medium of communication are different among the existing studies, they have a common role of addressing a crucial component of human communication, which should ultimately contribute to enriching L2 learners’ language activities. The current study has addressed the same goal as the literature, from the perspective of language assessment, by targeting L2 students’ pragmatic abilities for language activities at English-medium university. Although the scope of this study is limited, considering a wider range of topics in the literature of testing of pragmatics, the findings in this study would provide a step for future studies on the testing of pragmatics, which is an under-explored area of second language assessment but one
which is constantly advancing, going forward.
References


University Press.


277
Frankfurt am Main: Peter Lang.


Fernandez Guerra (Eds.), *Pragmatic competence in foreign language teaching* (pp. 161-184). Gastello: Servei de Publicacions de la Universitat Jaume I.


DOI: 10.1002/9781405198431.wbeal0049


Searle, J. (1975). Indirect speech acts. In P. Cole & J. L. Morgan (Eds.), *Syntax and


Xi, X. (2015, September). *Revisiting constructs- Recent theoretical advances and applications to practical test design, development and validation*. Workshop conducted at University of Melbourne, Melbourne, Australia.
APPENDIX A: Task General Instruction to Test Takers

Instruction to the test taker

■ You have

3 dialogue role-plays with a conversation partner.
and
3 monologue role-plays (leaving a voice mail message).

All of the role-plays simulate situations that you would encounter at university from the beginning to the end of the semester.

■ Some of the role-play situations are connected, but each role-play is independent. Therefore, your performance in the previous tasks will not affect what you are going to do in the next tasks.

■ In all of the role-plays:

- Your role is a student. Please use your own name for the role-play.
- You will address a Professor, Administrator, and Peer student.
- Please do the role-plays as naturally as possible. Feel free to use your own words.
- You are given a situation and context in each role-play.
  - Some situations may not have happened or may not happen to you. But imagine yourself in exactly the same given situations and how you would address each situation.

  - You can create more specific information (e.g., reasons for why you are asking the professor) within the given context.

- In the context, what to say and how to say it are up to you.
- You can make notes before the role-play starts.
- You can speak while looking at your notes and the task instruction.
- You can conclude the conversation or speech if you think it’s finished.
APPENDIX B: Role Play Task Set 1

Task ID: D1 Professor topic

- Get a signature to change classes (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
**Task ID: D1**

**Situation (at the beginning of the semester):**

You are currently taking a compulsory class and an elective class A. But you found out that you have a timetable clash with these two classes from week 4 (as in the class schedule circled in red below). Therefore, you need to change from Elective A to Elective B.

**Class schedule**

<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Week 1</strong></td>
<td>Compulsory</td>
<td>Elective A</td>
<td>Elective B</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Week 2</strong></td>
<td>Elective A</td>
<td>Elective B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Week 3</strong></td>
<td>Compulsory</td>
<td>Elective A</td>
<td>Elective B</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Week 4</strong></td>
<td>Compulsory</td>
<td>-Elective A 2pm</td>
<td>-Compulsory 2pm</td>
<td>Elective B</td>
<td></td>
</tr>
<tr>
<td><strong>Week 5</strong></td>
<td>Elective A 2pm</td>
<td>-Elective A 2pm</td>
<td>-Compulsory 2pm</td>
<td>Elective B</td>
<td></td>
</tr>
</tbody>
</table>

*Vocabulary Note:
Compulsory = a class that you must take
Electives = classes that you can choose

For this change, you need to submit an application form signed by the Elective A teacher and the Elective B teacher to the administration office by 5:30 pm tomorrow.

The other day, you emailed the teacher of Elective A (Prof. Smith) to explain your situation, but you have not got a reply yet.

Now, you are walking on campus, and you encounter and talk to Prof. Smith.

You do not have the application form with you now.

**Task:**

- Explain your situation to the professor, and negotiate in order to get her signature.
Task ID: M2, Administrator topic

- Change the registration of the class (Monologue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M2, Administrator topic

\textbf{Situation (at the beginning of the semester):}

You are taking Elective A, but you need to change it to Elective B for the reason(s) in task 1.

Today is the last day that you are allowed to change classes.

For this change, you need to get signatures of both teachers on the application form. You have already got the signature from the Elective A teacher, but the Elective B teacher has been off-campus and you have not met her yet.

You found the phone number of the General Enquiries of the administration office and call them because the office is far from where you are now. You are not sure if you can directly reach a staff member who can help you with this matter.

The phone line is busy and an automated message is asking you to leave a message in order for a staff member in charge to understand your situation and call you back.

It is 5:15 pm now and both the administration counter and the registration close at 5:30 pm.

Your personal information is below:

\begin{center}
\begin{tabular}{|l|}
\hline
\textbf{Student ID:} & 5028FF  \\
\textbf{Mobile:} & 0423-121-7735  \\
\hline
\end{tabular}
\end{center}

\textbf{Task:}

- Explain your situation to the administration office by leaving a message to get help for this problem.

There is no time limit for this voice message.
Task ID: D3, Peer student topic

- Make pairs for a presentation (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
Task ID: D3, Peer student topic

Situation (at the beginning of semester):

You have to make a presentation with another student about a topic in a class “Societal issues today”.

The presentation is very important for your grade in the class.

Most classmates have already made pairs, but you have not.

After the first session, outside of the classroom, you approach a classmate named Sharon to propose working together for the presentation. She is the only classmate that you have talked to in class.

Below is information about Sharon:

* You saw her in another class for the first time but only greeted each other.
* You know that she is interested in “People’s perceptions of health and what people do to stay healthy”
* You are not sure that she knows your interests.

You want to do a presentation with Sharon which includes questionnaire data of students at your university or people around you.

Doing the presentation with other classmates is very inconvenient for you.

Task:

- Think about what you are interested in (e.g., sports, cooking, science, education and so on) and how it relates to Sharon’s interests.

- Think about the benefit(s) of doing a questionnaire survey for the presentation.

- Convince Sharon to work together and do a questionnaire for the presentation by sharing what you thought about above.
Task ID: M4, Peer student topic

- Reorganize a presentation (Monologue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M4: Peer student topic

■ Situation (in the middle of semester):

You and your classmate (= Sharon, the same classmate in the last task) are doing a project for a class “Societal issues today”.

Your topic is “University students’ perceptions of health and what they consciously do to stay healthy”. You will present it in 4 days. This presentation is very important for your evaluation.

You and Sharon developed a questionnaire and have given it to 15 students, but only 2 of them were male. So you are worried that your data may not represent all students. However, you and Sharon have finished the data analysis and written some conclusions.

Today you found out that you could get 6 more male student participants to answer the questionnaire. You want to include these additional 6 participants for your research even if it takes time to collect the data, analyse it and modify your paper.

In order to do that, you need Sharon’s help to analyse the data because you don’t know how to use the analysis software well.

You called Sharon, but couldn’t reach her. You decide to leave a voice message because you are not sure when she will call you back and you may need to meet the 6 male participants soon.

■ Task:

- Think about why including 6 additional male student participants is good for your presentation.
- Leave a voice message to convince Sharon to reorganize the presentation by sharing what you thought about above.

There is no time limit for this voice message.
Task ID: D5, Administrator topic

- Solve a projector problem (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
Task ID: D5, Administrator topic

Situation (in the middle of semester):

You are going to make a presentation in the class “Societal issues today”. You have come to the classroom (Arts building Room 101) 20 minutes before the class to set up your laptop and connect it to the data projector with a cable to display your material on the screen. The projector is switched on but nothing is displayed on the screen and you do not know the reason for that.

Your presentation includes several visual materials so you need to be able to operate the device. There is no guarantee that your teacher and classmates can help you with this matter.

You found information about the Technical Support office on the projector, and go there, hoping that you can borrow some kind of device or that they can help you in some other way.

Task:

- Explain your situation to the staff member, and negotiate to get help for this problem.
Task ID: M6, Professor topic

- Avoid losing marks for an assignment (Monologue) -

You have 5 minutes to prepare for this task at maximum.
Please read the situation and the task carefully.
You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M6, Professor topic

■ Situation (At the end of semester):

You are working on the final assignment of “Societal issues today”, however you have been sick and could not finish it by the due date (yesterday).

The course syllabus says that:

- Marks will be lost at the rate of 15% per day for late submission.
- Hard-copy only for submission.

*Some special consideration may be applied.

Before the due date, you emailed the teacher (Prof. Brown) to explain your situation, but you have not got a reply yet.

You thought your email might have automatically gone into the Junk mailbox or might not have reached the professor. Therefore you decide to leave a voicemail message to explain the your situation again.

Your current circumstances are:
- You need one more day to complete the assignment. (You cannot finish it today).
- You are going to provide a medical certificate, although you have not got it yet.

■ Task:

- In order to avoid losing any marks, explain your situation to the professor by leaving a message.

There is no time limit for this voice message.
APPENDIX C: Role Play Task set 2

Task ID: M1, Professor topic

- Get a signature to change classes (Monologue) -

You have 5 minutes to prepare for this task at maximum.
Please read the situation and the task carefully.
You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M1, Professor topic

**Situation**  (at the beginning of the semester):

You are currently taking a compulsory class and an elective class A. But you have found out that you have a timetable clash with these two classes from week 4 (as in the class schedule circled in red below). Therefore, you need to change from Elective A to Elective B.

### Class schedule

<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>Compulsory</td>
<td>Elective A</td>
<td></td>
<td>Elective B</td>
<td></td>
</tr>
<tr>
<td>Week 2</td>
<td>Elective A</td>
<td>Elective A</td>
<td></td>
<td>Elective B</td>
<td></td>
</tr>
<tr>
<td>Week 3</td>
<td>Compulsory</td>
<td>Elective A</td>
<td></td>
<td>Elective B</td>
<td></td>
</tr>
<tr>
<td>Week 4</td>
<td>Compulsory</td>
<td>-Elective A</td>
<td>-Compulsory</td>
<td>-Elective B</td>
<td>2pm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2pm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Week 5</td>
<td>-Elective A</td>
<td></td>
<td></td>
<td>Elective B</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-Compulsory</td>
<td></td>
<td></td>
<td></td>
<td>2pm</td>
</tr>
</tbody>
</table>

*Vocabulary Note:
Compulsory = a class that you must take
Electives = classes that you can choose

For this change, you need to submit an application form signed by the Elective A teacher and the Elective B teacher to the administration office by 5:30 pm tomorrow.

The other day, you emailed the teacher of Elective A (Prof. Smith) to explain your situation, but you have not got a reply yet.

You thought that your email might have automatically gone into the Junk mailbox or might not have reached the professor. Therefore, you decide to leave a voice message to explain your situation again.

**Task:**

- Explain your situation and what you need to the professor by leaving a message.

There is no time limit for this voice message.
Task ID: D2, Administrator topic

- Change the registration of the class (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
Task ID: D2, Administrator topic

**Situation** (at the beginning of the semester):

You are taking Elective A, but you need to change it to Elective B for the reason(s) in task 1.

Today is the last day that you are allowed to change classes.

For this change, you need to get signatures of both teachers on the application form. You have already got the signature from the Elective A teacher, but the Elective B teacher has been off-campus and you have not met her yet.

You visit the General Enquiries counter of the administration office with the application form and talk to a staff member. You are not sure if the staff member can help you with this matter.

It is 5:15 pm now and both the administration counter and the registration close at 5:30 pm.

**Task:**

- Explain your situation to the staff member, and negotiate to get help for this problem.
Task ID: M3, Peer student topic

- Make pairs for a presentation (Monologue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M3, Peer student topic

■ Situation (at the beginning of semester):

You have to make a presentation with another student about a topic in a class “Societal issues today”.

The presentation is very important for your grade in the class.

Most classmates have already made pairs, but you have not.

The only classmate you know is Sharon, but she is not on the campus today.

Below is information about Sharon:

- You saw her in another class for the first time but only greeted each other and exchanged your mobile phone numbers.
- You know that she is interested in “People’s perceptions of health and what people do to stay healthy”
- You are not sure that she knows your interests.

You want to do a presentation with Sharon which includes questionnaire data of students at your university or people around you.

Doing the presentation with other classmates is very inconvenient for you.

You tried to call Sharon before she makes a pair with someone else, but couldn’t reach her. You decide to leave a voice message.

■ Task:

- Think about what you are interested in (e.g., sports, cooking, science, education and so on) and how it relates to Sharon’s interests.

- Think about the benefit(s) of doing a questionnaire survey for the presentation.

- Leave a voice message to convince Sharon to work together and do a questionnaire for the presentation by sharing what you thought about above.

There is no time limit for this voice message.
Task ID: D4, Peer student topic

- Reorganize a presentation (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
Task ID: D4, Peer student topic

**Situation (in the middle of semester):**

You and your classmate (= Sharon, the same classmate in the last task) are doing a project for a class “Societal issues today”.

Your topic is “University students’ perceptions of health and what they consciously do to stay healthy”. You will present it in 4 days. This presentation is very important for your evaluation.

You and Sharon developed a questionnaire and have given it to 15 students, but only 2 of them were male. So you are worried that your data may not represent all students. However, you and Sharon have finished the data analysis and written some conclusions.

Today you found out that you can get 6 more male student participants to answer the questionnaire. You want to include these additional 6 participants for your research even if it takes time to collect the data, analyse it and modify your paper.

In order to do that, you need Sharon’s help to analyse the data because you don’t know how to use the analysis software well.

You may need to meet the 6 male participants soon, and before that, you are going to meet Sharon now.

**Task:**

- Think about why including 6 additional male student participants is good for your presentation.
- Convince Sharon to reorganize the presentation by sharing what you thought about above.
Task ID: M5, Administrator topic

- Solve a projector problem (Monologue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.
Task ID: M5, Administrator topic

■ Situation (in the middle of semester):

You are going to make a presentation in the class “Societal issues today”. You have come to the classroom (Arts building Room 101) 20 minutes before the class to set up your laptop and connect it to the data projector with a cable to display your material on the screen. The projector is switched on but nothing is displayed on the screen and you do not know the reason for that.

Your presentation includes visual materials so you need to be able to operate the device. There is no guarantee that your teacher and classmates can help you with this matter.

You now call the staff on the Technical Support mobile number printed on the projector, hoping that you can borrow some kind of device or that they can help you in some other way, but you couldn’t reach the staff.

An automatic message is asking you to leave a message in order for the staff in charge to understand the situation and call you back.

Your personal information is below:

| Student ID: | 5028FF |
| Mobile:     | 0423-121-XXX |

■ Task:

- Explain your situation to the staff member by leaving a message to get help for this problem.

There is no time limit for this voice message.
Task ID: D6, Professor topic

- Avoid losing marks for an assignment (Dialogue) -

You have 5 minutes to prepare for this task at maximum.

Please read the situation and the task carefully.

You can start anytime you are ready. Tell the examiner when you are ready.

You speak first.
Task ID: D6, Professor topic

**Situation (At the end of semester):**

You are working on the final assignment of “Societal issues today”, however you have been sick and could not finish it by the due date (yesterday).

**The course syllabus says that:**

- Marks will be lost at the rate of 15% per day for late submission.
- Hard-copy only for submission.
*Some special consideration may be applied.

Before the due date, you emailed the teacher (Prof. Brown) to explain your situation, but you have not got a reply yet.

You have recovered now and want to visit the teacher’s office to avoid losing any marks.

**Your current circumstances are:**
- You need one more day to complete the assignment. (You cannot finish it today).
- You are going to provide a medical certificate, although you have not got it yet.

Now, you encounter Prof. Brown on your way to the office.

**Task:**

- In order to avoid losing any marks, explain your situation to the professor and negotiate.
APPENDIX D : Task Prompts for Interlocutor

Task 1: Professor topic
- Get a signature to change classes (Dialogue) -

You are Professor Smith of elective class A. It is the beginning of semester. You are walking on campus, and a student (= the test taker) comes to you to ask about changing classes from your class to another one. You have seen the student in class but do not remember his/her name.

In order for the student to change the class, s/he needs your signature on the application form for the class change. This application is handled by Administration (not you) but the student needs your signature first.

- Please lead the conversation naturally in such a way that a negotiation for mutual agreement occurs by providing what you can/cannot do for the student.
- Do not be too accommodating or direct towards the test taker.

In the conversation, you are required to:

- say to him/her you can sign the form if s/he has the form now.
- say you don’t remember receiving the email it or not if the student asks if you have read it
- respond to the student according to what s/he says. You are on your way to a meeting off-campus and do not have time to wait for him/her to see you again today.

Your schedule is below:

Now until tomorrow at around 5 pm:
Off campus for a meeting

After 5pm tomorrow: all-day meeting until 5pm. But, say that it’s possible you’ll be available in your office (though inconvenient for you).

The day after tomorrow: Available all day (tell the student this availability first)

- if the student wants to meet you tomorrow around 5pm onward, accept the request.
- some test takers might ask you to print out the form and sign it and put it in your mailbox. If they ask to do so, accept it, however the signed form won’t be ready until you go back to your office at around 5pm.
- if the student cannot continue the conversation, tell him/her to visit your office after 5pm

The test taker has 5 minutes to prepare for this task at maximum.
Please start the recording when s/he is ready.

The test taker speaks first.
You are a staff member at the General Enquiries counter of the administration office. You deal with matters relating to course subjects. A student comes to you to ask for a class change. The school rule states that students need to get signatures of both teachers from the new and current class on the class change application form. The student has not got a signature from the new teacher yet.

Today is the last day that students are allowed to change classes.

It is 5:15 pm now and the administration counter (and the registration) closes at 5:30.

- Please lead the conversation naturally in such a way that a negotiation for mutual agreement occurs by providing what you can/cannot do for the student.
- Do not be too accommodating towards the test taker
- You provide information for the student, but all of the information you provide is beyond your control and not guaranteed.

In the conversation, you are required to:

- imply that last year there was a case where an exception was made for a student submitting the application form at that time without the new teacher’s signature. *That student could then ask the new teacher to send an email to the Admin Office giving permission for the class change. However, that was only possible because the manager of the student administration office (your boss) granted the exception.

- imply to the student that this exception is not guaranteed.
- if asked, say that your boss is currently in a meeting which may finish before 6pm according to the schedule, but you not sure.
- ask the student what s/he wants you to do if s/he stops negotiating.
- if the student asks you to call the manager now or later, accept the request, and tell the student to come back in 30 minutes. Imply his/her that you will stay here just for his/her.
- if the student cannot continue the conversation, tell him/her to come back in 30 minutes.
You are taking a class “Societal issues today”. Your first name is Sharon (if you are a female) or Paul (if you are a male).

You have to make a presentation in pairs about a topic in the class. Most classmates have already made pairs. After the first session, one of them (= the test taker) approaches you to do the presentation together and propose his/her presentation plan. You know s/he is in another class but do not know them well and have so far only greeted each other.

Your interest is “people’s perceptions of health and what people do to stay healthy” and you want to do a presentation with a partner who will work on a topic related to yours.

Please lead the conversation naturally in such a way that a negotiation for mutual agreement occurs by sharing your ideas.

You are required to:

- if the classmate asks you if you have already made a pair with someone, imply that you were thinking of contacting Lucy (another classmate) or doing a presentation with someone who has a similar topic. Do not show your preference to his/her proposal quickly.
- ask how his/her interest is related to your interest, if not mentioned.
- ask the test taker why he/she hasn’t contacted others and if s/he is really ok to work with you
- accept the test taker’s request after sharing your interests with each other unless the test taker says something totally irrelevant, since your topic can be approached from various angles.
- if the test taker proposes doing some questionnaire work, show your unwillingness by implying that you are happy with the minimum grade (or effort) for the class.
- accept the test taker’s proposal if he/she provide some reasons for doing the questionnaire.
- if the test taker cannot continue the conversation, ask s/he wants to do and accept his/her request.

The test taker has 5 minutes to prepare for this task at maximum.
Please start the recording when s/he is ready.

The test taker speaks first.
You (‘Sharon’ or ‘Paul’) and your classmate (= test taker) are working on a project for a class “Societal issues today”. Your topic is “University students’ perceptions of health and what they consciously do to stay healthy” which you will present in 4 days.

You and your partner developed a questionnaire and have given it to 15 students but only 3 of them were male. You have both completed the data analysis and written some conclusions.

Today, your partner found out that they could get 6 more male student participants to answer the questionnaire.

You are not willing to include these additional 6 participants in your research because you will need to re-analyse it and modify the paper with the additional participants. You are busy with your part-time job and other assignments due next week. In order to modify the data, the test taker needs your help because he/she does not know how to operate the analysis software as well as you do.

You are required to:
- express your relief at having almost finished the analysis during the greeting exchange.
- show your reluctance to the classmate
- accept his/her request if a reason or a compromise of the idea is provided.

*Do not resist the test taker’s idea too much. You recognise that gender balance of the participants in research is important, even if you are busy with other things. Imply that you will analyse the data either by yourself or do it together with the test taker, if he/she provide a reason for doing so.

*if the test taker cannot continue negotiating with you, tell him/her that you will think about it and call him/her later.

The test taker has 5 minutes to prepare for this task at maximum. Please start the recording when s/he is ready.

The test taker speaks first.
Task 5: Administrator topic
- Ask for help for a projector problem (Dialogue) -

You are a staff member of the Technical Support office. A student comes to you asking for help.

- Please lead the conversation naturally in such a way that a negotiation for mutual agreement occurs by providing what you can/cannot do for the student.
- Do not be too accommodating towards the test taker

<Information you provide to the student>

- A technical support staff can be sent out to check the device and fix the problem such as checking the connection between the cable and the PC or fixing display problems.
- Lending devices (laptops, projectors, cables) to students is not a matter for your office. It may be a Facilities-related Office matter but you’re not sure, nor do you know where the Facilities Office is.
- The only technician available now is fixing a problem elsewhere but will come back to the office in 20 minutes (but you are not 100% sure)
- You cannot leave the office.

In the conversation, you are supposed to:

- explain what you can/cannot do, though not all of the information at the same time. Provide the pieces of information if asked or if the test taker seems to have difficulty working out what to do.
- accept the request if the student asks you to send the staff to his/her classroom when the staff returns
- if the student cannot decide what to do even with the information above, please tell them that you will contact the staff by mobile phone and ask him to go to the student’s class directly as soon as he finishes his current work. In that case, ask the test taker for their room number.
You are Prof. Brown teaching a class “Societal issues today.”
You are walking on campus, and a student comes to you to ask for Special Consideration for a late submission of the final assignment.

You think that special consideration will be given if the student provides a decent reason with some proof.

* Do not be too accommodating towards the test taker

You are required to imply:

- the due date is strict (the due is yesterday).
- a decent reason is necessary if the test taker asks for special consideration.
- that you are available today in the office and it’s best for you to receive the assignment today. (The student will say he/she cannot)
- that you are on a business trip abroad from tomorrow so it would be great if you can take a look at his/her assignment during the flight.

<Your schedule tomorrow>
- Will be in the office until 12pm but need to leave for the airport around that time.
- Flight departure: 4pm

You are required to:
- accept the student’s request if he/she wants to bring the assignment to you tomorrow while you are in the office or if he/she wants to send the assignment by email anytime by the end of the day (11:59 pm).
- imply to him/her that s/he needs to provide evidence explaining his/her circumstances, if it is not mentioned.
- imply to him/her that the evidence must be submitted to you as soon as possible (but you are not strict with the date and his/her ways of submission for this medical record).
- if the student cannot continue the conversation, tell him/her to visit your office at 12pm tomorrow.

The test taker has 5 minutes to prepare for this task at maximum.
Please start the recording when s/he is ready.

The test taker speaks first.
APPENDIX E: Questionnaire Items

Questions

(1)
- Did tiredness from doing six tasks at the same time change your performance?

- When you did the role play tasks, did doing six tasks with the same interlocutor change your performance?

(2)
Imagine that you had to address exactly the same given situations as the role plays.

Do you think your language use to handle the situations would be similar to your role play performances? Choose one of them below.

4. Very similar  3. Similar  2. Not so similar  1. Not similar at all

(3) How conscious are you (would you be) about the items below in communication in your university activities?


<table>
<thead>
<tr>
<th>Question item</th>
<th>Your answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) How formal/casual the language you use depending on the other person’s social status over you (e.g., when you speak to professor, vs peer student)</td>
<td></td>
</tr>
<tr>
<td>(b) How formal/casual the language you use depending on the other person’s familiarity to you (e.g., when the peer student is not so familiar in task 3, vs the same peer student in task 4 when you became familiar with the student).</td>
<td></td>
</tr>
<tr>
<td>(c) To use appropriate words, phrases, grammar to deliver what you really want to say for the situation</td>
<td></td>
</tr>
<tr>
<td>(d) To soften the way to request (e.g., indirectly mentioning) depending on what you ask for</td>
<td></td>
</tr>
<tr>
<td>(e) To correct what you said and re-state when you want to stress the information and/or soften what you are requesting</td>
<td></td>
</tr>
<tr>
<td>(f) To correct what you said and re-state when you made mistakes of language usage (vocabulary, grammar, sentence structure)</td>
<td></td>
</tr>
</tbody>
</table>
How do you assess your performances about each item today?

4. I could achieve this as well as I wanted to.
3. Sometimes I couldn’t do this.
2. I could not do this well even though I wanted to.
1. I did not do this.
0. Others (   )

<table>
<thead>
<tr>
<th>Question item</th>
<th>Your answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) How formal/casual the language you use depending on the other person’s social status over you (e.g., when you speak to professor, vs peer student)</td>
<td></td>
</tr>
<tr>
<td>(b) How formal/casual the language you use depending on the other person’s familiarity to you (e.g., when the peer student is not so familiar in task 3, vs the same peer student in task 4 when you became familiar with the student).</td>
<td></td>
</tr>
<tr>
<td>(c) To use appropriate words, phrases, grammar to deliver what you really want to say for the situation</td>
<td></td>
</tr>
<tr>
<td>(d) To soften the way to request (e.g., indirectly mentioning) depending on what you ask for</td>
<td></td>
</tr>
<tr>
<td>(e) To correct what you said and re-state when you want to stress the information and/or soften what you are requesting</td>
<td></td>
</tr>
<tr>
<td>(f) To correct what you said and re-state when you made mistakes of language usage (vocabulary, grammar, sentence structure)</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX F: Rating Rubrics

Social Actions to Achieve the Communicative Goal

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>The test taker’s performance shows understanding of the given or ongoing situation with clear, necessary and appropriate actions and information to prevent the problem and promote success in the situation (e.g., negotiate an alternative way with the administrator, proposing a meeting with the classmate, elaborate what the test taker says/tries to say) while the test taker handles the interlocutor’s unwillingness and resistance well.</td>
</tr>
</tbody>
</table>
| 3     | Actions (e.g., elaboration of explanations of the test taker’s situation) are taken to deal with the given situation although:  
  - introduction or closing is not fully tailored to the situation (e.g., redundant, too simple) and/or,  
  - some actions are redundant or not appropriate (e.g., initiating long self-introduction to the professor who may be in a hurry) for the situation. |
| 2     | - Actions are minimal in order to prevent the problem and promote success and/or,  
  - the test taker’s insufficient understanding of the situation is reflected in the test taker’s performance and/or,  
  - some actions (including introduction and closing) are not tailored for the situation and/or missing. |
| 1     | - The negative features characterizing score 2 are more noticeable and/or,  
  - insufficient actions are taken or insufficient information is provided to prevent the problem and promote success in the situation and/or,  
  - Copying (using the same wording) of part of the task prompt is noticeable. |

Notes for raters:  
* In voicemail situations, the test taker’s actions may not be diverse but it does not affect the evaluation negatively so long as test takers are clear enough to make the message receiver clearly understand what the message receiver can/should do for the student leaving the voice message.
# Facility with the Language

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4</strong></td>
<td>Clear and fluent throughout the speech/conversation and sound variation (e.g., intonation, stress) and speed control is explicitly seen. Repairs/ restatement, insertions may be used to elaborate and/or to stress or to mitigate the content, but these actions are well managed and fluency is not seriously flawed. Short pauses may be seen but not noticeably affect clarity and fluency.</td>
</tr>
</tbody>
</table>
| **3** | Generally clear, fluent although:  
  - sound variation or speed is generally flat/unnatural and/or,  
  - occasionally speech is paused (but resumed going). |
| **2** | Fluency is lost in some parts of the speech/conversation because of features such as pauses between words, pause between sentences, repairs, repetition, restatement, false starts, although clarity is maintained generally. |
| **1** | Clarity and fluency are noticeably flawed because of features such as  
  - unclear articulation and/or,  
  - insufficient control of repair (e.g., Repair/restatement is attempted but given up leaving parts/sentence(s) incomplete and/or the original part is replaced by another error/mistake) and/or,  
  - delivery being choppy, fragmented and minimal and/or,  
  - Combined and/or more frequent features characterizing score 2. |
## Language Use to Deliver the Intended Meanings

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>The test taker’s performance shows explicit control of linguistic resources (lexical choices, sentence structures) with few unclear parts to deliver the intended meanings throughout the speech. Employment of linguistics resources is natural.</td>
</tr>
<tr>
<td>3</td>
<td>The performance contains the following features although they do not seriously obscure the intended meaning: - unnatural word combination or lexical choice - missing /unnecessary components to structure the sentence / or incorrect word order (esp. in a wh-clause) or a different part of speech word is used (e.g., “data analyze” instead of “data analysis”, “consideration for healthy” instead of “consideration for health”).</td>
</tr>
<tr>
<td>2</td>
<td>Linguistics resources are insufficiently and/or inconsistently controlled with - unnatural part(s) and/or word combination(s) that change/obscure the intended meaning (the intended meaning should be different) and/or, - non-native like choice of linguistic resources being more noticeable and/or - the negative features characterizing score 3 being more noticeable.</td>
</tr>
<tr>
<td>1</td>
<td>The performance contains - errors and/or unintelligible parts that obscure the meanings and/or - part(s) where sentence structure is collapsed (clause boundary is unclear and/or jagged, most likely due to repairs, cancellation, and/or combined language errors or mistakes).</td>
</tr>
</tbody>
</table>
Language Use for Mitigation

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Lexical choices and/or combinations well mitigate imposition with devices including (bi-clausal, conditional: I wonder if/whether and/or modal verbs such as “would”, “could”, “might” in statement form). Even if these devices are not used and/or more direct expressions are combined, the test taker does not simply rely on direct expressions. The linguistic resources used for mitigation are naturally connected in the speech or in the ongoing flow of the conversation.</td>
</tr>
<tr>
<td>3</td>
<td>Imposition is mitigated generally although the test taker’s speech contains: - linguistic resources used for mitigation abruptly inserted or not always connected in the speech or in the ongoing flow of the conversation naturally and/or, - lexical choices/combinations not effectively mitigating and/or face-threatening.</td>
</tr>
<tr>
<td>2</td>
<td>The test taker is more likely to rely on limited and/or direct expressions (e.g., I need, I want,) and “can you (please)?” “could you (please)?” or simple structures considering the amount of speech and/or the test taker inconsistently uses complex structures for pragmatic meaning.</td>
</tr>
<tr>
<td>1</td>
<td>Mitigation devices and/or ways of mitigation are noticeably limited. The test taker is most likely to rely on simple structures and to use limited expressions.</td>
</tr>
</tbody>
</table>

Note for raters: (In monologue) the performance is not always long but the length of the performance does not matter as long as the imposition and face-threatening are mitigated in overall speech and the contents of the speech is clear enough for the message receiver to understand the student’s situation and what he/she can do for the student.
### Engagement in Interaction (for dialogue only)

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>A next turn clearly shows understanding of a previous turn response to the interlocutor (feedback to the interlocutor, clarification backchannel, acknowledgement tokens, questions) are well-tailored for the ongoing context throughout the interaction.</td>
</tr>
<tr>
<td>3</td>
<td>A next turn shows understanding of a previous turn in most parts of the interaction although the responses to the interlocutor (feedback to the interlocutor, clarification backchannel, acknowledgement tokens, questions) are generally simple in most part of the interaction.</td>
</tr>
<tr>
<td>2</td>
<td>The test taker’s response is generally small. Response to the interlocutor (evaluative feedback to the interlocutor, clarification backchannel, acknowledgement tokens, questions) is absent at least once or the test taker produces a turn that does not show understanding of a previous turn at least once.</td>
</tr>
</tbody>
</table>
| 1     | - The negative features described characterizing score 2 are more noticeable and/or,  
- the support from the interlocutor is noticeable (the test taker’s engagement in interaction is not active, which occasionally prompts the interlocutor to speak more and/or to initiate what the speaker is expected to say and/or to help the test taker to construct the conversation). |

**Note for raters:**
In relation to score 1, in some cases, an interlocutor may speak a lot even though the test taker is active/not passive. This case is not recognized as “the noticeable support from the interlocutor” that deserves score. Therefore, this does not affect the performance negatively.
**Turn Organization (for dialogue only)**

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
</tr>
</thead>
</table>
| 4     | Smooth natural turn-taking throughout the conversation as:  
- adjacency pairs are completed (e.g., question & answer, request & thanks) without awkward pauses and  
- the speaker (test taker) releases/receives/keeps conversation flowing in a way that will not confuse the interlocutor (as to whether the interlocutor should keep listening or take a turn to facilitate the speaker’s speech). Even if turns overlap, the speakers are not confused. |
| 3     | Generally smooth turn-taking although the performance may contain parts where:  
- adjacency pairs are not completed (e.g., question & answer, request & thanks) and/or,  
- the test taker’s turn may be a bit delayed. |
| 2     | The test taker puts awkward pause(s) between turns or confuses the interlocutor (as to whether the interlocutor should keep listening or take a turn to facilitate the speaker’s speech). |
| 1     | The negative features described for score 2 are more noticeable throughout the conversation and/or,  
- some turns are noticeably delayed or awkwardly initiated without response to the interlocutor’s comment. |

**Notes for raters:**
Please ignore the following when rating  
*Sometimes turns may be cut off by the interlocutor but it should not affect the evaluation of the test taker’s performance negatively.  
*In each task situation, the interlocutor resists and/or shows unwillingness to the test taker’s proposal. The test taker may pause for a moment after the resistance and/or unwillingness, but the pause after the interlocutor’s resistance and/or unwillingness does not affect the evaluation negatively.
**APPENDIX G: Transcription Conventions**

Adapted from (Atkinson & Heritage, 1984; Heritage, 1984) and partially modified for the current study

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
<td>A lengthened sound or syllable; more colons prolong the stretch</td>
</tr>
<tr>
<td>=</td>
<td>Latch; no gap or silence between turns</td>
</tr>
<tr>
<td>(( _ ))</td>
<td>Non-vocal action or the researcher’s description of an event</td>
</tr>
<tr>
<td>(word)</td>
<td>Empty parentheses: Something is being said but the utterance could not be transcribed</td>
</tr>
<tr>
<td></td>
<td>Words written within a single bracket: the researcher’s best guess at the uncertain utterances</td>
</tr>
<tr>
<td>↑</td>
<td>A sharp rise in pitch</td>
</tr>
<tr>
<td>word</td>
<td>Speaker emphasis</td>
</tr>
<tr>
<td>WORD</td>
<td>Noticeably louder than surrounding speech</td>
</tr>
<tr>
<td>?</td>
<td>Rising intonation, not necessarily a question</td>
</tr>
<tr>
<td>(.)</td>
<td>Very short pause less than two-tenths of a second</td>
</tr>
<tr>
<td>(3.5)</td>
<td>Silence measured in seconds and tenths of seconds</td>
</tr>
<tr>
<td>&gt; &lt;</td>
<td>Talk produced noticeably quicker than surrounding talk</td>
</tr>
<tr>
<td>[</td>
<td>Onset of overlapping talk</td>
</tr>
<tr>
<td>-</td>
<td>Cut-off or self-interruption</td>
</tr>
<tr>
<td>° °</td>
<td>Noticeably quieter than surrounding talk</td>
</tr>
</tbody>
</table>
Author/s:  
Ikeda, Naoki

Title:  
Measuring L2 oral pragmatic abilities for use in social contexts: development and validation of an assessment instrument for L2 pragmatics performance in university settings

Date:  
2017

Persistent Link:  
http://hdl.handle.net/11343/191879

File Description:  

Terms and Conditions:  
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.