RATER BIAS IN ASSESSING THE PRAGMATICS OF KFL LEARNERS USING FACETS ANALYSIS

SOO JUNG YOUN

University of Hawai'i at Mānoa

ABSTRACT

As interest in research on second language pragmatics increases, some pragmatics research has been done on Korean as a foreign language (KFL) learners. This research has focused on pedagogical aspects of Korean pragmatics and interlanguage pragmatics. However, very little research has been done on the pragmatics assessment of KFL learners, in terms of discussing appropriate test types and whether certain factors affect raters' assessments of KFL learners' pragmatics performance. The focus of this study is on investigating whether various factors, including test types, speech acts, groups of candidate, and test items, affect three raters' assessments of the pragmatic competence of KFL learners. For these purposes, this study analyzes interactions between test results and raters using the computer program FACETS (Linacre, 1996): the interactions between rater bias and test types, rater bias and speech acts, rater bias and item difficulty, and rater bias and examinee levels. This study uses three different pragmatics tests adapted from Hudson, Detmer, and Brown's (1995) pragmatics prototype tests: Open-written Discourse Completion Task, Language Lab, and Role-play. Within each of these three test types are three speech acts: refusal, apology, and request. The results of this research indicate that all three raters showed different degrees of severity in their ratings, depending on the test type and speech act. Additionally, each rater showed unique bias patterns within the interactions. I will discuss how certain speech acts and test types affect rater assessments, what kinds of systematic bias patterns the raters show across various factors, and what these research findings mean for KFL classrooms.

INTRODUCTION

Since Hymes (1972) proposed the communicative competence theory, this theory has greatly influenced the development of target objectives for language teaching and learning. Pointing out the limitations of Chomsky's (1965) distinction between competence and performance, Hymes proposed a broader notion of communicative competence, covering not only grammatical competence, but also contextual or sociolinguistic competence. Above all, Hymes' distinction between language knowledge and ability for language use, as well as his incorporation of sociolinguistic knowledge into the framework of communicative competence, have contributed to many of the discussions of language testing constructs (Canale & Swain, 1980; Canale, 1983; Bachman, 1990; Bachman & Palmer, 1996).

Adopting Hymes' notion of communicative competence, Canale and Swain (1980) defined communicative competence as "the underlying systems of knowledge and skill required for communication" and proposed three components of communicative competence: grammatical competence, sociolinguistic competence, and strategic competence. Based on the Canale and Swain model, Bachman and Palmer (1982) attempted to empirically validate components of communicative competence, and Bachman (1990) proposed a model of *communicative language ability* in which he included three components: language competence, strategic competence, and psychophysiological mechanisms. In his model, pragmatic competence was included as one of two main components of language competence with its two subcomponents: illocutionary competence (knowledge of the pragmatic conventions for carrying out appropriate language functions) and sociolinguistic competence (knowledge of sociolinguistic rules of appropriateness to a given context). Most recently, a revision of Bachman's model has been introduced (Bachman & Palmer, 1996). As seen in Bachman's model, pragmatic competence figures significantly in theories of communicative competence in L2 teaching and testing.

As linguistic pragmatics has origins from quite different philosophical, sociological, and linguistic traditions, pragmatics has been defined in various ways (Levinson, 1983). Among various definitions of pragmatics, Crystal's (1997) definition of pragmatics focused on communicative action in its sociocultural context:

> 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)

In addition to Crystal's definition, Kasper and Rose (2001, p. 2) pointed out that communicative action consists of not only using speech acts such as requests, refusals and apologies, but also joining in conversation, engaging in various discourse types and being involved in complex speech events. Leech divided pragmatics into two subcomponents: pragmalinguistics, and sociopragmatics. According to his distinction, pragmalinguistics is the linguistic end of pragmatics which refers to “the particular resources which a given language provides for conveying particular illocutions”; sociopragmatics is the sociological interface of pragmatics referring to the fact that the underlying participants’ social perceptions are relative to specific social conditions (Leech, 1983, p.11).

Interlanguage Pragmatics

Drawing on two different disciplines, second language acquisition (SLA) and pragmatics, interlanguage pragmatics (ILP) has been defined as “the study of nonnative speakers’ use and acquisition of linguistic action patterns in a second language (L2)” (Kasper & Blum-Kulka, 1993; Kasper, 1996). Early ILP studies mainly focused on learners’ demonstration of illocutionary force and different politeness perceptions. Especially, how nonnative speakers comprehend indirect speech acts in the target language has been focused on. Studies such as Bouton (1988) found that learners’ cultural background and type of implicature influenced learners’ comprehension of indirect speech acts, through comparing the differences among six groups of learners from different countries. However, as Kasper and Schmidt (1996) cautioned, the focus of ILP research has been within learners’ L2 use rather than developmental issues even though ILP is a subfield of SLA. There have been efforts to reinforce the connections between ILP and SLA by reexamining ILP research findings that answer SLA related questions, and by exploring cognitive and social-psychological theories (Kasper & Schmidt, 1996; Kasper & Rose, 2002). The studies of development of speech acts, especially focusing on English requests, are mainly focused on in the literature. Regarding the speech acts development, Kasper and Rose (2002) comprehensively discussed the proposed five-stages of L2 request development stages in terms of pragmalinguistics, based on the longitudinal studies of Ellis (1992) and Achiba (2002). They also noted that learners can gradually adjust their requests as they gain more proficiency level. When it comes to the sociopragmatics of L2 requests development, which shows learners’ underlying social perceptions depending on specific social situations, mixed findings were reported. There were also conflicting findings in the studies of other L2 speech acts development. About the possible explanations of these mixed findings, Kasper and Rose (2002) mentioned the effect of learners’ different learning contexts and individual learner differences.

Pragmatics in KFL Teaching

During the last three decades, there have been significant developments in KFL (Korean as a foreign language) education in US college settings. Teaching Korean honorifics has been one of the important concerns in KFL teaching because of its rich lexical and morphological variation depending on its meaning, and its significant role in understanding Korean language socio-cultural rules. In other words, Korean honorifics are essential resources for pragmatics' two subcomponents, pragmalinguistics and sociopragmatics. There has been a large literature on the Korean language's systematic and extensive honorifics (e.g., Kim-Park, 1995; Koh, 2002; Lee, 1996; Sohn, 1999; Strauss & Eun, 2005). Appropriate usage of Korean honorifics is bound by the speakers' knowledge of the social relationship with the addressee regarding variable factors, such as, age, social status, and kinship. Sohn noted "relative interpersonal relationships are elaborately encoded in various linguistic forms to the extent that speech acts cannot be performed without taking the notion of honorifics into account" (1999, p. 408). Therefore, KFL learners, especially those who don't have similar honorific systems in their L1, have difficulties engaging in various speech acts using appropriate Korean honorifics; Byon (2004a) emphasized the importance of teaching appropriate Korean honorifics in effective ways for improving KFL learners' pragmatic competence.

Regarding pragmatics studies on KFL learners, Byon (2002, 2004a, 2004b, 2005) embarked not only on ILP studies focusing on KFL learners' language use in various speech acts but also on teaching pragmatic competence in KFL classroom. Byon (2004b) investigated KFL learners' sociopragmatic features in requests through comparing request productions of three groups, KFL learners, Korean native speakers, and English native speakers, using discourse completion tasks (DCT). Byon analyzed semantic formulae for request supportive move (RSM) and for request head act (RHA) of request production, considering two social variables, power and distance. According to the findings of Byon's study, KFL learners' semantic formulae patterns were consistent with those of English native speakers, indicating the effect of L1 transfer. Also, KFL learners' use of RHA pattern showed less variation than those of Native Koreans in situations that involve different power relationship with the interlocutor. Byon's study implicates further investigation of ILP studies in KFL settings, which still needs more developments and interests.

Assessment of Pragmatic Competence

As Rose and Kasper (2001, p. 245) point out, research on the assessment of pragmatic competence has gained less attention compared with the significant amount of research on teaching of pragmatics. Hudson, Detmer, and Brown (1992) developed a framework for assessing cross-cultural pragmatics as a first phase of their project. Adapting Brown and Levinson's (1987) theory, they selected three variables to be included in the tests: the power of the speaker with regard to the hearer, the distance between the speaker and the hearer, and the obligation of the speaker to accomplish the acts. Later in 1995, in a second phase, they described developing test instruments with a discussion from both quantitative and qualitative points of view. As seen in Table 1, Hudson et al. developed the six prototype measures for assessing pragmatic competence targeting the L2 English learners: a multiple-choice discourse completion test (DCT), a open-ended DCT, an oral DCT, a role play, a self-assessment for the DCT, and a self-assessment for the role play (Hudson, Detmer, & Brown, 1995). To develop reliable test items of each of the six prototype measures, they developed several sets of test items, piloted these test items to both native speakers of English and non-native speakers, analyzed the test results qualitatively and quantitatively, and finalized the test items of each measure. Hudson et al. also developed a 1-5 rating scale, ranging from very unsatisfactory to completely appropriate, for native speakers who rated each examinee's response with the following criteria: ability to use the correct speech act, typical expressions, amount of speech in a given situation, formality level, directness level, and politeness level.

Table 1
Classification of Test Methods (Based on Hudson, Detmer, & Brown, 1995)

	Cued Response	**Free Response**
Paper and Pencil Measures	Items with multiple-choices following descriptions of situations	Items with open-ended responses following description of situations
Oral Measures	Listening laboratory taped items following descriptions of situations	Face-to-face structured oral interview
Self-assessment Measures	Self assessment of performance on each situation depicted in DCT	Direct observation and evaluation of the video-taped role play and interview

Following Hudson et al.’s (1992, 1995) projects, some researchers investigated the reliability and validity of the instruments that Hudson et al. developed in different target language teaching contexts, as described in Table 2 (Ahn, 2005; Brown, 2001; Hudson, 2001; Yamashita, 1996; Yoshitake, 1997). Other researchers developed their own test instruments to assess pragmatic competence (Roever, 2001; Tada, 2005). Yamashita (1996) used a Japanese version of Hudson et al’s six test instruments targeted at forty-seven American English speakers with Japanese as a second language (JSL) to examine differences among different test formats. Using various statistical procedures, the validity and reliability of the six types of measures were quantitatively investigated, and reasonably high reliability and validity for all six measures, excluding the multiple-choice DCT were reported. As seen in the third row of Table 2, Yoshitake (1997) qualitatively examined the twenty-five Japanese EFL learners’ written realizations and oral production data, elicited from the original six types of measures developed by Hudson et al. (1995). Also, this study compared its results with Hudson et al.’s (1995), in terms of the differences between Japanese EFL and ESL learners’ various speech acts realizations, since the data of Hudson et al. (1995)’s were collected from Japanese ESL learners. Yoshitake (1997) found that Japanese ESL learners showed a wider variety of strategies and grammatically more complex structures than Japanese EFL learners’. Brown (2001) compared the results of two previous studies which represent JSL and EFL contexts respectively, as described in Table 2. He reported that a factor analysis indicated there was a stronger method effect in the EFL study than in the JSL study. In other words, there might be more differences in the EFL group’s ability to handle oral tasks and their ability to handle paper-and-pencil tasks. Additionally, the six measures in the JSL study showed higher reliability and significant correlations compared with the ones in the EFL study.

Hudson (2001) examined three types of measures, language lab DCT, open-ended DCT, and role play, to assess pragmatic competence of twenty-five Japanese learners of English as a second language (ESL), as described in the fifth row of Table 2. Even though there was little variation among the participants and different speech acts, the results revealed that the role play performed differently from the other two measures, indicating a method effect between a role play and the DCT format. In his study, refusals seemed to be more difficult to perform for participants than requests and apologies.

Roever (2001) examined the development and validation of web-based tests of ESL and EFL learners' pragmalinguistic knowledge as described in the sixth row of Table 2. The tests are composed of assessing knowledge of implicatures and routines, using multiple choice items, and knowledge of three speech acts (refusal, request, and apology), using productive DCT items with rejoinders. Based on correlational analyses, it was found that the tests indeed assess learners' pragmalinguistic knowledge with reasonable accuracy. Also, there were negligible effects of computer familiarity to the test scores, which implicates the promising potential of web-based language tests in pragmatic assessment. Roever also reported that there are distinctive ILP characteristics of ESL and EFL learners' pragmalinguistic knowledge depending on various factors, because knowledge of routines was strongly influenced by exposure; knowledge of speech acts and implicatures was strongly influenced by test takers' proficiency.

Tada (2005) investigated Japanese EFL learners' pragmatic production and perception tests of three speech acts, refusal, request, and apology, using computerized video prompts, focusing on the relationship between perception and production, as seen in the seventh row of Table 2. Tada found that there was a stronger correlation between learners' proficiency and pragmatic production than between learners' proficiency and pragmatic perception in all three speech acts, which indicates pragmatic perception may develop more independently as learners' proficiency develops.

In addition to L2 English and Japanese pragmatic assessment, there has been one study of L2 Korean pragmatic assessment (Ahn, 2005). Ahn investigated the reliability and validity of five measures of pragmatics that Hudson et al. (1992, 1995) developed, language lab DCT, role play, role play self-assessment, and multiple choices DCT, for KFL learners using various statistics, as described in the last row of Table 2. It was found that the five pragmatics measures were reasonably reliable and valid, which indicates Hudson et al.'s test instruments are also applicable to KFL contexts. Also, Ahn reported that the level of examinees' language proficiency is closely related to role play self-assessment, language lab DCT, and open-ended DCT. Interestingly, the differences between heritage language learners and non-heritage language learners were more significant in open-ended DCT than in the other measures. Also, Ahn (2005) suggested incorporating the five measures into the KFL classroom, not only for assessing pragmatic competence but also for assessing learners' different competence.

Table 2
Summary of Previous Studies of Assessing Pragmatic Competence

Study	**Speech Act**	**Test Takers' Proficiency**	**L1**	**Context**	**Target Language**	**Test Instruments**	**Focus of Study**	**Statistical Instruments**
Yamashita, 1996	Refusal, Request, Apology	Beginning, Intermediate, Advanced (*N* = 47)	English	JSL	Japanese	SA, LL DCT, Open DCT, Role play, Role play SA, MC DCT	Investigation of reliability and validity of six measures of pragmatics in Japanese	Cronbach alpha coefficient, Pearson product-Moment correlation coefficients, Factor Analysis, MANOVA
Yoshitake, 1997	Refusal, Request, Apology	Participants who ranged from 423 to 577 on the TOEFL (*N* =25)	Japanese	EFL	English	SA, LL DCT, Open DCT, Role play, Role play SA, MC DCT	Qualitative analysis of study results, Comparison with the Hudson et al.'s (2005) result	N/A
Brown, 2001	Refusal, Request, Apology	Beginning, Intermediate, Advanced	English, Japanese	JSL, EFL	Japanese, English	SA, LL DCT, Open DCT, Role play, Role play SA, MC DCT	Comparison of Yamashita's (1996) study and Yoshitake's (1997) study	Cronbach alpha coefficient, K-R21, SEM, Correlation coefficients, Factor Analysis

Table 3 (continue)

Hudson, 2001	Refusal, Request, Apology	Intermediate, Advanced (*N* = 25)	Japanese	ESL	English	LL DCT, Open DCT, Role play	Investigation of three measures of pragmatics in English	ANOVA
Roever, 2001	Refusal, Request, Apology	Beginning, Intermediate, Advanced (*N* = 267)	German, Japanese, Chinese, Korean, Thai, Polish, and other languages*	ESL, EFL	English	Web-based Implicatures Test, Routines Test, Speech Acts Test	Development and validation of web-based test of pragmalinguistic knowledge (implicatures, routines, and speech acts)	K-R 21, SEM, FACETS Analysis, Factor Analysis, ANOVA
Tada, 2005	Refusal, Request, Apology	Beginning, Intermediate (*N* = 48)	Japanese	EFL	English	Production Test, Perception Test usingcomputerized audiovisual prompts	Investigation of the development of pragmatic production and perception of speech acts	MANOVA, ANOVA, Discriminant Function Analysis
Ahn, 2005	Refusal, Request, Apology	Beginning, Intermediate, Advanced (*N* = 53)	English	KFL	Korean	LL DCT, Open DCT, Role play, Role play SA, MC DCT	Investigation of reliability and validity of five measures of pragmatics in Korean	Cronbach alpha coefficient, Pearson product-Moment correlation coefficients, Factor Analysis, MANOVA

*Other languages include Arabic, Berber, Kurdish, Persian, Russian, Sinhala, Slovenian, Spanish, Swedish, Tamil, and Turkis

Item Response Theory and Multi-faceted Rasch Measurement

Although classical testing theories (CTTs) have played a significant role in language testing, some practical concerns in language testing that CTT could not solve have emerged. Among the numerous shortcomings of CTT, Hambleton and Swaminathan (1985) mentioned that test and item scores in CTTs are dependent upon the particular set of test items and upon the particular group of examinees who took the test. In other words, in CTTs, it is impossible to interpret and compare the test results of a particular group of examinees with the results from a more able or less able group of examinees, due to the different distribution of item scores. Item response theory (IRT) has been considered a promising potential for addressing CTTs' limitations (Bachman, 2004; Brown & Hudson, 2002; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; McNamara, 1996). Firstly, supplementing CTT, IRT compares the abilities of the examinees on a scale with the actual difficulty of the particular test item, which makes it possible to examine the contribution of items individually as they are added and removed. Secondly, IRT assumes that each examinee test performance on a particular test item can be predicted by defining both the examinee abilities and the item difficulty. Also, IRT facilitates computer adaptive testing, because IRT is the suitable for obtaining candidates' ability estimates based on candidates' performances on certain items and for selecting appropriate next test items of currently estimated candidates' ability levels from item banks. There are three frequently used models of IRT: the one-, two-, and three-parameter logistic models. Each model is defined mathematically involving three different parameters: item difficulty, discrimination, and guessing parameters.

This study will use multi-faceted Rasch measurement, which is an extension of the one-parameter Rasch model after the Danish mathematician Georg Rasch. Multi-faceted Rasch measurement, which can be executed by the computer program FACETS (Linacre, 1996), investigates the impact of various factors, or *facets*, including facets like task difficulty, candidate ability, or rater severity, on the rating process. According to McNamara (1996), each rating can be considered a result of the interaction of the three facets; based on these interactions, it is possible to predict estimates of probabilities of candidate responses under the various facets (ability, difficulty, and rater severity), and to evaluate the accuracy of this prediction. To put all the facets together into a single set of relationships, all estimates of the facets can be expressed on a single measurement scale, called a logit scale, which presents the relative status of elements within a facet. Therefore, we can express the difficulty of a certain item based on the likelihood of a certain candidate to earn a given score, or have a certain ability, from a rater of given severity on that item.

The computer program FACETS (Linacre, 1996) for multi-faceted Rasch measurement provides three sets of detailed results. Firstly, the FACETS summary that shows the relative status of all facets on the same scale in a single set of relationships is provided. Secondly, a thorough measurement report is presented for each facet. Through this measurement report, we can find detailed information about each facet, such as rater severity, rater consistency, candidate ability, and item difficulty. Thirdly, multi-faceted Rasch measurement provides an important feature known as bias analysis, which makes it possible to identify the rater particular bias patterns of harshness or leniency with regard to particular candidates or particular task types, by investigating the difference between expected and observed scores.

Multi-faceted Rasch measurement has been applied in various language assessment settings to investigate rater behaviors systematically, such as ESL speaking skills of immigrants (Lynch & McNamara, 1998), Japanese L2 writing (Kondo-Brown, 2002), English L2 oral discussion (Bonk & Ockey, 2003), a Japanese medical translation program (Kozaki, 2004), German as a foreign language learners' writing and speaking performance (Eckes, 2005), English L2 oral performance (Van Moere, 2006), and an online training program for English L2 writing assessment (Elder, Barkhuizen, Knoch, & Randow, 2007). Depending on the study's focus, some studies, such as Kondo-Brown (2002), investigated the measurement report of each facet and bias analysis as well. Kondo-Brown (2002) examined trained rater bias patterns across candidates and rating categories in detail using multi-faceted Rasch measurement in a Japanese L2 writing performance assessment context. In her study, raters showed consistency in their ratings; however, there were significant differences in overall severity among them. Based on bias analysis, she reported that raters revealed significant biased interactions with regard to the candidates and rating categories.

Other studies, such as Elder, Barkhuizen, Knoch, and Randow (2007), mainly examined the measurement report of each facet rather than bias analysis. Elder et al. (2007) examined rater reactions to the introduction of an online rater training program for English L2 writing assessment, and whether raters showed internal consistency in their scoring following online training using both quantitative (multi-faceted Rasch measurement) and qualitative (questionnaire) analyses. Although the study found that the online rater training program had minimal impact on improving intra-rater consistency and reducing the bias of each rater, it provided great potential for exploring the various rater interactions and different training modalities.

However, studies focusing on such detailed investigations of rater assessments of pragmatic competence have been rare. Therefore, the present study attempts to investigate not only raters' overall judgments, but also their systematic bias patterns across various factors in pragmatic competence assessment setting using bias analysis.

Purpose and Research Questions

The purpose of this study is to investigate whether various factors, including task types, speech acts, groups of candidate, and test items affect three raters' assessments of the pragmatic competence of Korean as a Foreign Language (KFL) learners using FACETS analysis; if so, this study will discuss what kinds of patterns exist among these various factors.

The following research questions will be investigated in this study:

1. What are the three raters' overall severities in their ratings?
2. Are the three raters consistent in their ratings?
3. How reliably do the three raters reveal different degrees of severity?
4. Are there any misfitting, or problematic, elements among the raters, task types, and speech acts?
5. Do any of the raters assess particular task types more harshly or more leniently than others? If so, what are the raters' sub-patterns of assessing task types?
6. Do the raters assess particular test items more harshly or more leniently than others? If so, what are the raters' sub-patterns of assessing items with different difficulties?
7. Do the raters assess particular speech acts more harshly or more leniently than others? If so, what are the raters' sub-patterns of assessing speech acts?
8. Do the raters assess particular ability of examinees more harshly or more leniently than others? If so, what are the raters' sub-patterns of assessing examinees with different abilities?

METHOD

Participants

The test takers in this study were twenty-four participants studying Korean as a foreign language (KFL) at the University of Hawai‘i at Mānoa (UHM) and Hanoi University of Foreign Studies (HUFS). The test takers were divided into two groups, heritage language learners (HLLs) and non-heritage language learners (NHLLs), following a definition of HLLs that foreign language educators in the U.S. generally adopt: "…a language student who is raised in a home where a non-English language is spoken, who speaks or at least understands the language, and who is to some degree bilingual in that language and in English" (Valdes, 2001, p. 38). As seen in Table 3, the test takers are composed of HLLs (N = 7) and NHLLs (N = 17). Following Valdes' definition of HLLs, even though a participant may have familiarity or affiliation with Korea, if they were not raised in a Korean-speaking home then they were categorized as a NHLL. The NHLL group's first languages were various: Vietnamese (N = 8), English (N = 6), Japanese (N = 2), and Spanish (N = 1). These participants were composed of 14 females and 10 males ranging in age from 18 to 58 years, with a mean age of 25. With regard to the level of study of the test takers, there were seven students from 300 level courses, eight students from the 400 level, and one student from the Korean flagship program of the UHM. The course numbering system in the Korean department is described as "courses numbered 300 or above are upper-division and count toward the non-introductory credit requirement" (University of Hawai‘i at Mānoa, 2006). According to the University of Hawai‘i at Mānoa Korean flagship program (2006), the program offers "students with advanced Korean language proficiency an opportunity to undertake additional intensive, task-based Korean language instruction." There were also four students from the third year and the fourth year, respectively, studying Korean as a foreign language at the HUFS, which has a four year Korean program focusing on translation, grammar, culture, and literature. All test takers' levels ranged from intermediate level to advanced level. Originally, beginning learners of Korean participated, but they could not complete all tests because of their lack of proficiency.

Table 3
Summary of Participants as Test Takers

Participant Groups	***N***	**%**
HLL	7	29%
NHLL	17	71%
NHLL's L1		
Vietnamese	8	47%
English	6	35%
Japanese	2	12%
Spanish	1	6%
Gender		
Female	14	58%
Male	10	42%
Level		
Intermediate	11	46%
Advanced	12	50%
High-advanced	1	4%
Age		
18-20	6	25%
21-30	14	58%
31-40	2	8%
41-50	1	4%
51-60	1	4%

Note. Total number of participants = 24

Three native Korean raters participated in rating all examinees' responses. They were two female raters and one male rater ranging in age from 27 to 32 years, with a mean age of 29.7. All three raters had MA degrees related to foreign language teaching and two of them had experience teaching Korean.

Materials

Adapting from the six measures for assessing pragmatic competence that Hudson et al. (1992, 1995) developed in English, three types of measurement were used in this study: an open-ended written DCT, a language lab DCT, and a role play. All these tests were translated into Korean and English supplements of all tests were prepared as well.

The open-ended written DCT, or the OPDCT (see Appendix A) was composed of eighteen different descriptions of situations that elicited either refusals, requests, or apologies. Each situation was controlled by different variables such as power (P), social distance (D), and degree of imposition (I) (Hudson, Detmer, & Brown, 1995, pp. 4-5). All test takers were asked to write in Korean what they would say in a given situation after they read each situation. Additionally, they were informed not to spend too much time on it, and not to discuss with others even though they were allowed to consult with a dictionary. Like the OPDCT, the language lab DCT, or the LLDCT (see Appendix B), also consisted of eighteen different descriptions of situations that elicited one of three different speech acts with three variables. After test takers listened to each situation, they were asked to respond orally. For the role play, or the RP (see Appendix C), six scenarios were provided and each scenario included three small consecutive situations in which a native Korean interlocutor and a test taker did a role play. Three speech acts and three variables were controlled for the eighteen situations as well.

The rating criteria that Hudson et al. (1995) developed were applied to all three test types. As shown in Table 4, which illustrates sample rating sheets, there were six aspects for the ratings, ability to use the correct speech act, ability to use of typical expressions, appropriate amount of speech and information given, level of formality including word choice, and phrasing, level of directness, and level of politeness (Hudson, Detmer, & Brown, 1995). Each aspect was rated on a five-point Likert scale from 1, very unsatisfactory, to 5, completely appropriate.

Table 4
Rating Sheet (Hudson, Detmer, & Brown, 1995)

SITUATION							
Response # ___________				**Response #** ____________			
Speech act	1-2-3-4-5			Speech act	1-2-3-4-5		
Expressions	1-2-3-4-5			Expressions	1-2-3-4-5		
Amount / info	1-2-3-4-5	-	+	Amount / info	1-2-3-4-5	-	+
Formality	1-2-3-4-5	-	+	Formality	1-2-3-4-5	-	+
Directness	1-2-3-4-5	-	+	Directness	1-2-3-4-5	-	+
Politeness	1-2-3-4-5	-	+	Politeness	1-2-3-4-5	-	+

Procedures

The twenty-four volunteer test takers took all three types of test. The researcher met with each test taker twice to complete all tests. It took approximately 30 minutes to complete each test, so it took total one and a half hours to finish all three tasks for each examinee. The purposes of this study were not explained to the participants until they completed all tests. Additionally, the tests were described as tasks or questionnaires to reduce the test anxiety of the participants. As Hudson (2001, p. 289) mentioned, to minimize the effect of one test on the following test, the order of the test administration was language lab DCT (LLDCT), written discourse completion task (OPDCT), and finally, role play (RP). Each test taker took the OPDCT home to finish, and the LLDCT and the RP were administered with a researcher present. All test administrations were conducted within a two week period. To prevent the effect of poor hand-written answers on the OPDCT raters' decisions, all hand-written responses in the OPDCT were typed as an MS Word document. All test takers' responses from the LLDCT and conversations from the RP were recorded as digital sound files.

For the rating procedures, a training manual was prepared for the three Korean native raters. All rating criteria were fully explained in the training manual, and the researcher also met each rater to explain the rating criteria. An audio CD, which contained all test takers' responses, was distributed to each of the raters, and it took about two weeks for them to complete all the ratings. To avoid any effects of test takers' backgrounds such as age or nationality on the raters' decisions, none of the test takers' information was provided to the raters.

Data Analysis

In this study, multi-faceted Rasch measurement (Linacre, 1989) analysis was conducted using the computer program FACETS, version 3.0 (Linacre, 1996). This study conducted FACETS program twice to examine characteristics of raters across task types, speech acts, examinees, and test items. Firstly, examinee, task, rater, and item were specified as four facets to investigate interactions between raters and task types, raters and examinees, and raters and test items. Secondly, examinee, speech act, rater, and item were specified as four facets to investigate interactions between raters and speech acts. As illustrated in Tables 5 and 6, all data of this study were transformed into a matrix of information to be executed by the FACETS program. In Table 5, the first set of four facets, examinees, task types, raters, and test items, are presented from the first to the fourth column respectively; the rest of columns indicate the examinee test results from the three raters on each test item within each of the three tasks. Similarly, in Table 6, the second set of four facets, examinees, speech acts, raters, and test items, are shown from the first to the fourth column respectively, and the examinee test results are represented in the rest of columns. Each examinee has nine rows of data, based on the three raters from the three task types and speech acts. Therefore, there are total 216 rows (24×9) in each data matrix, and each row has the facet information, including examinee, task type, speech act, rater, and item. As McNamara (1996, p. 132) noted, based on the information of data matrix, the predictability of elements within each facet can be found, and essentially the analysis continues to find any such consistent patterns in the data. Additionally, the FACETS analysis provides a detailed measurement report regarding each facet, including such features as severity measures and fit statistics for raters, and a bias analysis that presents systematic rater patterns of harshness or leniency across task types, speech acts, examinees, and test items.

Table 5
Data Matrix with Four Facets (examinee, task, rater, and item)

Examinee	**Task**	**Rater**	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**…**
1	1	1	1-18	29	27	18	27	22	28	24	28	30	…
1	1	2	1-18	25	24	19	27	25	23	26	26	20	…
1	1	3	1-18	26	24	19	24	18	19	20	23	19	…
1	2	1	19-36	25	29	23	22	28	30	19	30	28	…
1	2	2	19-36	22	18	19	19	21	28	17	23	13	…
1	2	3	19-36	22	17	16	13	21	24	10	16	19	…
1	3	1	37-54	27	30	24	29	28	29	28	28	29	…
1	3	2	37-54	21	28	24	21	25	27	25	19	25	…
1	3	3	37-54	22	27	15	20	27	24	23	23	24	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
24	3	3	37-54	21	23	21	23	22	6	18	22	28	…

Table 6
Data Matrix with Four Facets (examinee, speech act, rater, and item)

Examinee	**Speech Act**	**Rater**	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**…**
1	1	1	1-17	27	28	25	27	27	20	28	30	20	…
1	1	2	1-17	27	26	20	26	29	26	23	19	12	…
1	1	3	1-17	24	23	23	24	24	21	23	19	15	…
1	2	1	18-35	29	27	28	30	27	30	25	29	28	…
1	2	2	18-35	25	24	23	20	24	22	22	18	21	…
1	2	3	18-35	26	24	19	19	20	21	22	17	21	…
1	3	1	36-54	18	22	24	25	26	18	23	22	30	…
1	3	2	36-54	19	25	26	25	25	25	19	19	28	…
1	3	3	36-54	19	18	20	19	20	21	16	13	24	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…
24	3	3	36-54	18	24	19	20	24	17	12	22	12	…

RESULTS

In this results section, first, I will present the descriptive statistics for the three different task types and for the two different sub-groups of examinees, HLLs and NHLLs. Secondly, I will present two FACETS summaries from the two different sets of facets that show the relative status of all facets (examinee, task type, speech act, rater, and item) all on the same scale in a single set of relationships. Thirdly, I will present each facet detailed measurement report that shows examinee ability, rater severity, raters consistency, the difficulty of the three task types and speech acts, and test item difficulty. Fourthly, I will present the bias analysis that shows the systematic patterns of harshness or leniency between the three raters and various facets: examinee, task type, speech act, and item. Also, I will explain the technical terminology used in FACETS analysis as such items appear in the results.

Descriptive Statistics

A perfect score on each task is 30, based on five points for each of the six rating categories. Table 7 shows the descriptive statistics for the three task types, open-ended DCT (OPDCT), language lab DCT (LLDCT), and role play (RP), based on the scores from the three raters. The lowest and highest scores on each task are presented in the second and third rows, respectively. Note that the OPDCT has a perfect score 30 from all three raters while the LLDCT and RP do not. The means (*M*) for the three task types are displayed in the fourth row. The mean for the RP (*M* = 24.00) is the highest while the mean for the LLDCT (*M* = 22.20) is the lowest. The standard deviations (*S*) for the three task types, which indicate the dispersions for scores, are presented in the last row. Note that the LLDCT has the highest amount of dispersion (*S* = 4.20) among the three task types although the LLDCT has the lowest mean, and the RP has the lowest dispersion of the scores (*S* = 3.38).

Table 7
Descriptive Statistics for Each Task Type (*N* = 24)

Statistics	OPDCT	LLDCT	RP
Low	0.00	0.00	11.00
High	30.00	29.33	29.89
M	22.65	22.20	24.00
S	3.76	4.20	3.38

Table 8 presents the descriptive statistics for the three task types from each of the three raters. The order of statistics is the same as in Table 6. Note that the means for the RP are consistently the highest from the three raters, 27.31, 23.43, 21.27, respectively; the means for the LLDCT are consistently the lowest from the three raters, 26.31, 21.67, 18.63, respectively. Regarding the dispersion of the scores, the LLDCT has the highest dispersion of scores from both rater 1 (*S* = 3.63) and rater 3 (*S* = 4.54) while the OPDCT shows the highest dispersion from rater 2 (*S* = 4.60). Overall, the means for the three task types from rater 1 are higher than the ones of both rater 2 and rater 3.

Table 8
Descriptive Statistics from Each Rater on Each Task Type (*N* = 24)

Rater 1

Statistics	OPDCT	LLDCT	RP
Low	0	0	6
High	30	30	30
M	26.91	26.31	27.31
S	2.94	3.63	2.62

Rater 2

Statistics	OPDCT	LLDCT	RP
Low	0	0	6
High	30	30	30
M	22.09	21.67	23.43
S	4.60	4.42	3.34

Rater 3

Statistics	OPDCT	LLDCT	RP
Low	0	0	6
High	30	30	30
M	18.95	18.63	21.27
S	3.72	4.54	4.19

Tables 9 and 10 show the descriptive statistics for the three task types for the two sub-groups of examinees, HLLs and NHLLs. The descriptive statistics are the lowest score, the highest score, the mean, and the standard deviation, for each task type are presented from each rater (columns 2, 3, and 4) and from the three raters combined (column 5). Comparing Tables 9 and 10, HLLs group shows higher means for both the LLDCT (*M* = 24.08) and RP (*M* = 25.59) than the mean for the LLDCT (*M* = 21.43) and RP (*M* = 23.35) of the NHLLs group. However, the HLLs group shows a slightly lower mean for the OPDCT (*M* = 22.55) than the mean for the OPDCT (*M* = 22.69) of the NHLLs group from the three raters. Regarding the dispersion of the scores for each group, the NHLLs group has the higher amount of dispersion for all three task types from the three raters than ones of the HLLs group. This result shows that the HLLs group has better listening and speaking ability, manifested in the LLDCT and RP, than reading and writing ability, manifested in the OPDCT.

Table 9
Descriptive Statistics of HLLs Group (*N* = 7)

OPDCT				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	19	10	10	16.00
High	30	30	30	30.00
M	27.92	21.45	18.27	22.55
S	2.02	4.47	3.45	3.31
LLDCT				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	15	14	8	15.00
High	30	30	30	29.00
M	27.36	24.30	20.59	24.08
S	2.80	3.97	4.11	3.63
RP				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	19	14	8	16.00
High	30	30	30	30.00
M	28.01	25.84	22.93	25.59
S	2.07	2.47	3.45	2.66

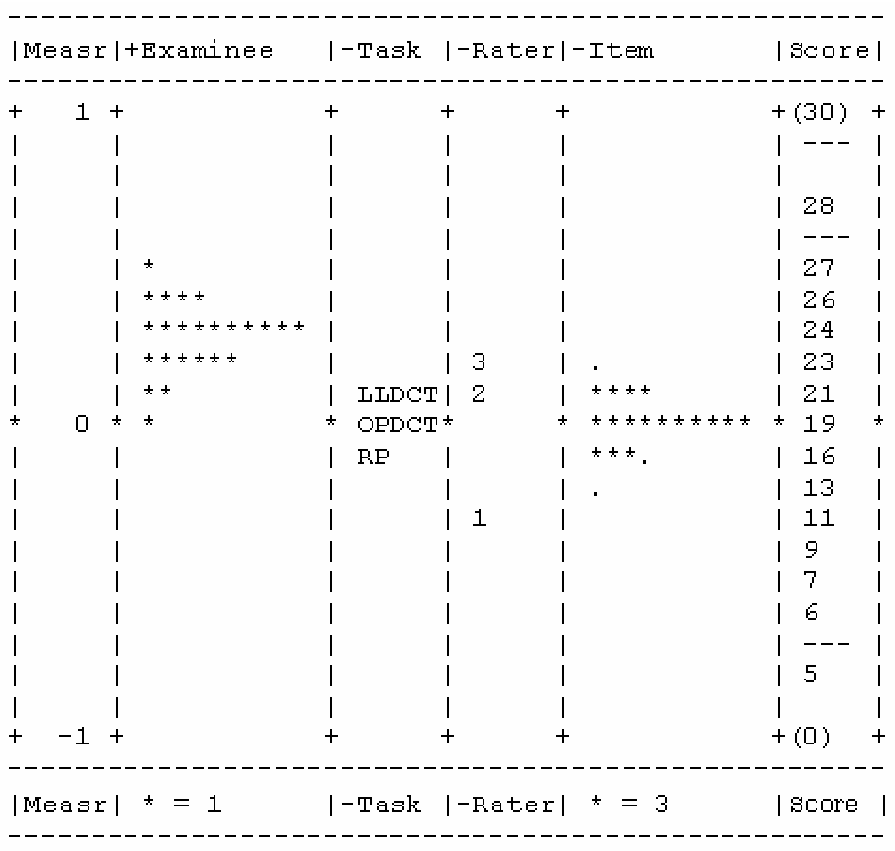
Table 10
Descriptive Statistics of NHLLs Group (*N* = 17)

OPDCT				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	0	0	0	0.00
High	30	30	29	28.89
M	26.50	22.35	19.23	22.69
S	3.32	4.66	3.84	3.94
LLDCT				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	0	0	0	0.00
High	30	30	28	29.22
M	25.88	20.59	17.83	21.43
S	3.97	4.61	4.72	4.43
RP				
Statistics	Rater 1	Rater 2	Rater 3	Total
Low	6	6	6	11.00
High	30	30	29	29.22
M	27.02	22.44	20.59	23.35
S	2.84	3.70	4.50	3.68

FACETS Summary

The FACETS computer program provides a summary that displays the relative status of all facets, in this case, examinees, task types, speech acts, raters, and items, in a single set of relationships. As mentioned in the data analysis section, I ran the FACETS program twice. Firstly, examinees, task types, raters, and items were specified as the first set of four facets to run FACETS, and secondly, examinees, speech acts, raters, and items were specified as the second set of four facets. The FACETS summaries of these two sets of four facets are presented in Figures 1 and 2.

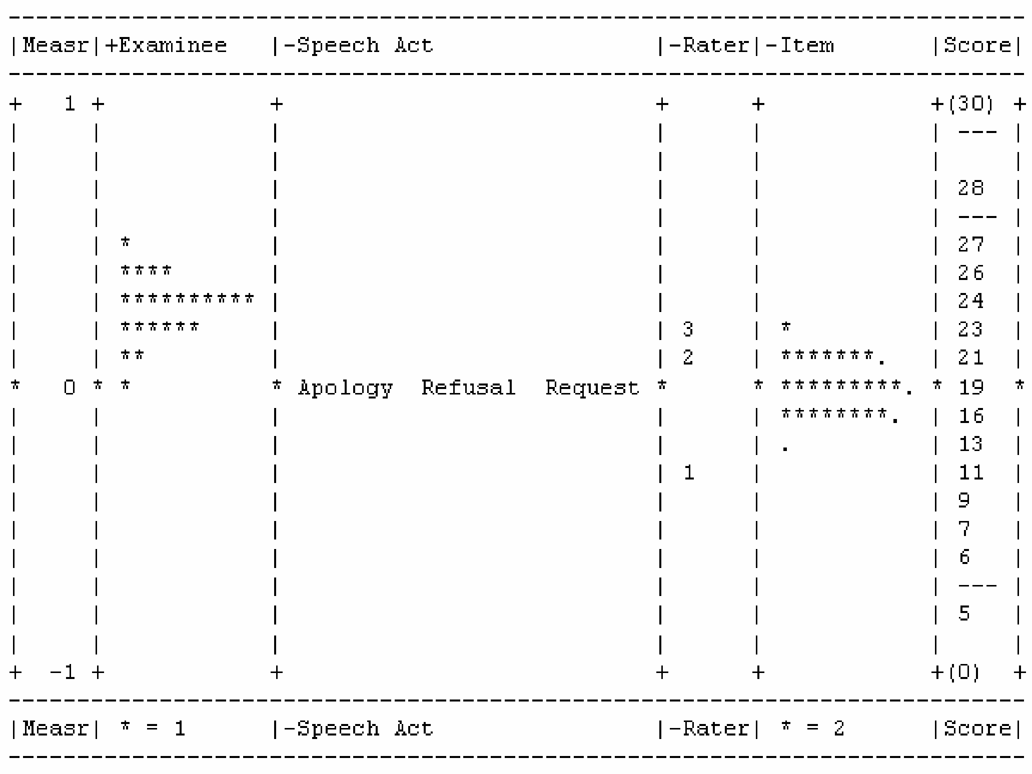
The information about the relative abilities of the examinees, the relative difficulties of the three task types and speech acts, the relative harshness of the three raters, and the relative difficulty of the items are shown in Figures 1 and 2 using a logit scale. The logit scale shown in the first column of Figure 1 is a true interval scale that reflects comparable distances for each of the facets. The second column shows a histogram of the ability range of the examinees, and one asterisk (*) in this column indicates one examinee. In the examinee ability estimates, an examinee with a zero logit would have a 50 percent chance of success on an item of average difficulty, which means that the higher on the logit scale, the more able. Therefore, it is shown that all examinees of this study have at least a 50 percent chance of succeeding on an average difficult item since all examinees scored above zero logit. The third column shows the difficulty estimates of each task type. In the difficulty estimates (columns 3), the items with a zero logit indicate the items of average difficulty. Also, the items with negative signs are easier than average, and those with positive signs are more difficult than average. Accordingly, the LLDCT was the most difficult task, and the RP was the easiest one. The fourth column shows the three raters' severity in the logit scale. The harshest rater is at the top and the most lenient rater is at the bottom, because the scale is set up such that the higher the logit value the harsher the rating. Among the three raters of this study, rater 3 was the harshest and rater 1 was the most lenient on the scoring. As shown in Figure 1, the rater severity (column 4) especially between rater 3 and rater 1 showed considerable variation, compared with the other facets. The fifth column shows the item difficulty, and an asterisk (*) in this column indicates three items. Similar with the third column, items expressed as negative logits are easier than average; those expressed as positive logits are more difficult than average. So, the items (column 5) shown in Figure 1 are distributed evenly around the zero logit, which indicates the average difficulty. Finally, the column to the far right indicates the spread of the raw test scores for the test items. Notice the distances between the intervals of the scores (column 6) are not equal, for example, there is a wider gap between score 28 and 30 than between the score 24 and 26, which indicates that the raw scores are not true interval scores.



```
-----------------------------------------------------
|Measr|+Examinee   |-Task |-Rater|-Item       |Score|
-----------------------------------------------------
+   1 +            +      +      +            +(30) +
|     |            |      |      |            | --- |
|     |            |      |      |            |     |
|     |            |      |      |            | 28  |
|     |            |      |      |            | --- |
|     | *          |      |      |            | 27  |
|     | ****       |      |      |            | 26  |
|     | ********** |      |      |            | 24  |
|     | ******     |      | 3    | .          | 23  |
|     | **         | LLDCT| 2    | ****       | 21  |
*   0 * *          * OPDCT*      * ********** * 19  *
|     |            | RP   |      | ***.       | 16  |
|     |            |      |      | .          | 13  |
|     |            |      | 1    |            | 11  |
|     |            |      |      |            | 9   |
|     |            |      |      |            | 7   |
|     |            |      |      |            | 6   |
|     |            |      |      |            | --- |
|     |            |      |      |            | 5   |
|     |            |      |      |            |     |
+  -1 +            +      +      +            +(0)  +
-----------------------------------------------------
|Measr| * = 1      |-Task |-Rater| * = 3      |Score |
-----------------------------------------------------
```

Figure 1. FACETS summary with four facets (examinee, task type, rater, and item)

Since Figure 2 is similar to Figure 1, the logit scale is also presented in the first column, and a histogram of the ability range of the examinees is shown in the second column. In the third column, the difficulty estimate of the three speech acts, apology, refusal, and request, are shown. The three speech acts showed similar difficulty. The fourth column shows the three raters' severity, and indicates that rater 3 was the harshest and rater 1 was the most lenient. The fifth column displays the even distribution of the three speech act test items around the average difficulty (the zero logit), and an asterisk (*) in this column indicates two items. Finally, the spread of raw test scores for the test items is shown in the column to the far right. Again, these raw scores have different intervals between each score.



```
--------------------------------------------------------------------------
|Measr|+Examinee   |-Speech Act                |-Rater|-Item       |Score|
--------------------------------------------------------------------------
+   1 +            +                           +      +            +(30) +
|     |            |                           |      |            | --- |
|     |            |                           |      |            |     |
|     |            |                           |      |            | 28  |
|     |            |                           |      |            | --- |
|     | *          |                           |      |            | 27  |
|     | ****       |                           |      |            | 26  |
|     | ********** |                           |      |            | 24  |
|     | ******     |                           | 3    | *          | 23  |
|     | **         |                           | 2    | *******.   | 21  |
*   0 * *          * Apology  Refusal  Request *      * *********. * 19  *
|     |            |                           |      | ********.  | 16  |
|     |            |                           |      | .          | 13  |
|     |            |                           | 1    |            | 11  |
|     |            |                           |      |            | 9   |
|     |            |                           |      |            | 7   |
|     |            |                           |      |            | 6   |
|     |            |                           |      |            | --- |
|     |            |                           |      |            | 5   |
|     |            |                           |      |            |     |
+  -1 +            +                           +      +            +(0)  +
--------------------------------------------------------------------------
|Measr| * = 1      |-Speech Act                |-Rater| * = 2      |Score|
--------------------------------------------------------------------------
```

Figure 2. FACETS Summary with four facets (examinee, speech act, rater, and item)

Measurement Report of Five Facets

Besides the FACETS summary, FACETS analysis provides detailed information about all five facets used in this study, examinee, rater, task type, speech act, and items, in separate tables. In each table for each facet, three crucial summary statistics, a logit value for each facet, standard error, and fit statistics, are included (Linacre, 1996). Firstly, the logit value of each facet represents what is intended to be measured by the researcher. For example, depending on the facet, the logit value can indicate the examinee abilities, rater severities, or test item difficulties. Secondly, the standard error means the level of error of the logit estimate. Lastly, fit statistics are crucial for the validity of the measure, because the fit statistics are an indication of the degree to which each element is observed in the way that is expected by the statistical model. Therefore, the fit statistics are a way to interpret the pattern of *residuals*, the gap between the expected and the observed score, which can be expressed as either a *mean square* or *t*. The range of *mean square* values signals the extent of variation in such values; a value greater than the mean plus twice the standard deviation would be considered as misfitting (McNamara, 1996, p. 172). In addition to these three statistics, FACETS analysis also provides indications of the degree of the differences among elements of each facet: reliability, separation index, and fixed (all same) chi-square. These terms will be explained in detail as they appear in the results.

Measurement report of examinee abilities. Detailed information about examinee ability measurement is shown in Table 11. All rows of Table 11 are arranged by examinee ability logits. The first column lists each examinee, and the second column indicates estimates of examinee abilities on a logit scale. The higher the logit, the more able the examinee. Therefore, examinee 23 with a logit of 0.50 was the most able examinee, and examinee 12 with a logit of 0.02 was the least able examinee. All examinee logit values are above the zero logit, which indicates all examinees had at least a 50 percent change of getting an item of average difficulty right. The third column shows that the standard error is small (0.02 logit) and equal among all examinees. Lastly, the fit statistics, expressed as a *mean square*, are presented in the fourth column. The fit statistics are a way of identifying problematic elements of each facet (in this case, examinee) through interpreting the pattern of *residuals*, the gap between the expected and the observed score. Also, fit values greater than the mean plus twice the standard deviation would be considered misfitting, or problematic (McNamara, 1996, p. 172). Applying these conventions to Table 11, two examinees (examinees 6 and 19, with infit mean square values of 1.8) are identified as misfitting because they are outside of the range between 0.4 and 1.6 (1.0± [0.3×2]), based on the mean (1.0) and the standard deviation (0.3) for the infit values. This indicates that the test results of both examinees 6 and 19 were not consistent but showed much variation from what was expected.

In addition to the summary statistics for examinee abilities, reliability, separation index, and fixed (all same) chi-square are found at the bottom of Table 11. Firstly, the reliability statistic indicates that the analysis reliably reveals the different degrees of ability or difficulty among the elements of each facet (in this case, examinee ability). The reliability in Table 11 is 0.96 for all examinees, indicating this analysis quite reliably divides the examinees into different levels of ability. Secondly, the separation index indicates the spread of each measure (in this case, examinee ability) in relation to the standard error. The examinee separation index in Table 11 is 5.18, indicating the variance among examinee abilities is about five times the standard error estimates. Thirdly, the fixed (all same) chi-square tests the fixed hypothesis “this set of elements can share the same measure after allowing for measurement error", and the significance indicates the probability that the fixed hypothesis is the case (Linacre, 1996). Therefore, in Table 11, the fixed hypothesis is "all examinees' ability can share the same ability measure." However, this fixed hypothesis must be rejected because the significance shows that the probability in this case is 0.00. Therefore, it can be concluded that all examinees have different levels of ability even after allowing for the standard error. In sum, based on the measurement report of examinees, FACETS analysis indicates that the test reliably separated the examinees into different levels of ability, and all the examinees scored above the average ability (the zero logit). However, there were two examinees that showed much variation in their test results outside of what the model predicted.

Table 11
Measurement report of twenty-four examinees

Examinee	Ability (logits)	Error	Infit (mean square)
23	0.50	0.02	1.1
20	0.43	0.02	1.2
22	0.40	0.02	1.3
5	0.39	0.02	0.8
16	0.36	0.02	0.8
13	0.34	0.02	0.9
6	0.33	0.02	1.8*
10	0.33	0.02	1.1
4	0.31	0.02	1.4
14	0.31	0.02	1.0
24	0.30	0.02	0.8
7	0.29	0.02	0.8
11	0.29	0.02	0.6
1	0.28	0.02	0.8
18	0.26	0.02	0.8
15	0.24	0.02	0.7
19	0.24	0.02	1.8*
17	0.22	0.02	0.8
2	0.20	0.02	1.0
9	0.18	0.02	0.6
8	0.15	0.02	1.5
3	0.14	0.02	1.3
21	0.09	0.02	0.9
12	0.02	0.02	1.1
M	0.27	0.02	1.0
S	0.11	0.00	0.3

Note. Reliability = 0.96; separation index = 5.18; fixed (all same) chi-square = 703.4; significance = 0.00

* = Misfitting examinee

Measurement report of rater severities and internal consistency. Detailed information about the three raters, in terms of the relative severity and consistency of their scoring, is found in Table 12. The first column lists the three raters, and the second column indicates an estimate of the rater severity in terms of the chance of getting a given rating with that rater. Also, the standard errors of these estimates are provided in the third column, which shows very small errors (0.01 logits). Rater 3 was the most severe and rater 1 was the least severe because the higher the logit value is the more severe the rater. The gap between the most severe rater and the least severe rater is 0.51 logits. The fourth column of Table 12 indicates the fit statistics of the three raters' judgments. Again, the fit statistics provide information for identifying the degree to which each element (in this case, each rater) is observed in the way that is expected by the statistical model. Applying the conventions of the fit statistics to Table 12, no raters were identified as misfitting, or problematic; all infit mean square values of the three raters were between 0.5 and 1.7 (1.1± [0.3×2]), based on the mean (1.1) and standard deviation (0.3) for the infit values.

As shown at the bottom of Table 12, the reliability of the measurement for the three raters is 1.00, indicating that the different degrees of severity among the three raters are reliable. The separation index is 27.51, indicating that there is much variation in rater severity logits in relation to the standard error, considering that the separation index of the examinee measurement report was 5.18. Finally, the fixed chi-square of 1883.0 was significant at a probability of 0.00. Therefore, the fixed hypothesis, in this case, that "the raters can share the same severity" can be rejected. In other words, all three raters have different degrees of severity even after considering the standard error. In sum, based on the measurement report of the three raters, in terms of rater severity and rating consistency, all three raters reliably showed different degrees of severity. Rater 1 (logit = -0.29) was consistently the most lenient among them, and rater 3 (logit = 0.22) was consistently the harshest among them. Also, no raters were found to be problematic in terms of their scoring performance.

Table 12
Measurement Report of Three Raters

	Severity (logits)	Error	Infit (mean square)
Rater 1	-0.29	0.01	1.5
Rater 2	0.07	0.01	1.0
Rater 3	0.22	0.01	0.9
M	0.00	0.01	1.1
S	0.21	0.00	0.3

Note. Reliability = 1.00; separation index = 27.51; fixed (all same) chi-square = 1883.0; significance = 0.00

Measurement report of difficulty of task types and speech acts. Tables 13 and 14 show the measurement report for the three task types and speech acts. The first columns of the two tables list each task type and speech act, and the second columns show difficulty logits, indicating the comparable difficulty estimates among the three task types and speech acts. Considering these difficulty logits of the three task types in more detail, it was found that the LLDCT (logit = 0.06) was the most difficult, the OPDCT (logit = 0.02) was the second difficult, and the RP (logit = -0.07) was the easiest task since items with negative sign in difficulty logit are easier than average, and those with positive sign are more difficult than average. When it comes to the difficulty of the three speech acts, apology (logit = 0.03) was the most difficult, and refusal and request were the second most difficult speech acts with the same difficulty logits (-0.02), as shown in Table 14. The degree of standard error of these estimates, which is small (0.01 logits), is reported in the third column. In the fourth column, the fit values, which provide information about whether any of the elements are misfitting, are shown. All fit values from Tables 13 and 14 are within the range of two standard deviations around the mean (1.0± [0.1×2=0.2]: 0.8 ~ 1.2).

Table 13
Measurement Report of Three Task Types

	Difficulty (logits)	Error	Infit (mean square)
OPDCT	0.02	0.01	1.1
LLDCT	0.06	0.01	1.1
RP	-0.07	0.01	0.9
M	0.00	0.01	1.0
S	0.06	0.00	0.1

Note. Reliability = 0.98; separation index = 7.72; fixed (all same) chi-square = 174.9; significance = 0.00

Table 14
Measurement Report of Three Speech Acts

	Difficulty (logits)	Error	Infit (mean square)
Refusal	-0.02	0.01	1.0
Apology	0.03	0.01	1.2
Request	-0.02	0.01	0.9
M	0.00	0.01	1.0
S	0.02	0.00	0.1

Note. Reliability = 0.91; separation index = 3.25; fixed (all same) chi-square = 34.6; significance = 0.00

The reliabilities reported in the two measurement reports are shown at the bottoms of Tables 13 and 14. These are 0.98 and 0.91, respectively, which indicates the three task types and speech acts reliably differ in their difficulties. The separation indexes of the task types and speech acts are 7.72 and 3.25, respectively, indicating the difficulties of the task types have nearly twice the variation than the speech acts in relation to the standard errors. Regarding the fixed chi-square, the fixed hypothesis of Tables 13 and 14 is that “the task types and speech acts can share the same difficulty.” This hypothesis must be rejected, for the task types and speech acts as the chi-squares of 174.9 and 34.6, respectively, are significant at a probability of 0.00. In other words, the levels of difficulty for the task types and speech acts are not equal even after allowing for the standard errors.

In sum, the FACETS analysis revealed that the three task types and speech acts reliably differ in their difficulties. Among the task types, the LLDCT was the most difficult, and the RP was the easiest task. Among the speech acts, apology was the most difficult, while refusal and request were the second most difficult. Notice that, as shown in the separation index, the task types showed more variation than the speech acts. Also, no task types or speech acts were identified as misfitting which also indicates that the differences between the expected and the observed scores were consistent.

Measurement report of test items. Table 15 presents the measurement report for all the test items from each of the three task types. All rows of Table 15 are arranged by item difficulty logit. The first column lists the 54 test items, based on 18 test items for each of the three task types. Therefore, the OPDCT items are from items 1 to 18; the LLDCT items are from items 19 to 36; and the RP items are from items 37 to 54. The second column shows the item difficulty logit values. The most difficult item was item 41 (logit = 0.21), which is an RP item, and the easiest item was item 28 (logit = -0.17), which is an LLDCT item, because the higher the logit the more difficult the test item. All test items are evenly distributed around the difficulty of the zero logits. The standard errors of these estimates are found in the third column. All item difficulties estimates have a standard error of 0.03 except for item 49 (0.04). Lastly, the fit values are provided in the fourth column of Table 15. Applying the standards of the fit values to Table 15, it is found that two items, which are items 19 and 41 (fit value =2.0), are outside the 0.2 and 1.8 (1.0± [0.4×2]) range, based on the mean (1.0) and the standard deviation (0.4) for the infit values. Therefore, these two items are identified as misfitting, indicating that these items have inconsistent patterns of *residuals*, the gap between the expected and the observed score.

In addition to the summary statistics for test items, reliability, separation index, and fixed (all same) chi-square are also shown at the bottom of Table 15. The reliability is 0.81, indicating the degree to which the FACET analysis indicates that the items reliably distinguished between different levels of difficulty. In this case, the items somewhat less reliably separate different levels of difficulty than other facets, considering that the reliability statistics in the previous measurement reports for the other facets, examinee, rater, task type, and speech act, were all above 0.90. In Table 15, the separation index is 2.09, indicating the difficulty variance among the test items is about twice the standard error. Lastly, the fixed chi-square of 297.6 is significant at a probability of 0.00, which indicates that the fixed hypothesis that “the test items can share the same difficulty estimate” must be rejected. In other words, all test items have different levels of difficulty even though the reliability of the separation of test items into different levels of difficulty is somewhat low. In sum, based on the measurement report for test items, the items somewhat less reliably reflect different degrees of difficulty than other facets. In addition, there were two misfitting test items: items 19 and 41.

Table 15

Measurement Report of Test Items

Items	Difficulty (logits)	Error	Infit (mean square)
41	0.21	0.03	2.0*
42	0.12	0.03	1.7
3	0.11	0.03	1.5
17	0.10	0.03	0.8
10	0.09	0.03	1.1
21	0.09	0.03	1.1
23	0.09	0.03	1.3
27	0.09	0.03	0.8
48	0.07	0.03	1.0
8	0.06	0.03	1.6
6	0.05	0.03	0.7
12	0.05	0.03	1.0
44	0.05	0.03	0.7
29	0.04	0.03	0.8
30	0.04	0.03	1.4
37	0.04	0.03	0.7
9	0.03	0.03	1.3
25	0.03	0.03	0.9
39	0.03	0.03	0.9
11	0.02	0.03	0.8
22	0.02	0.03	0.8
35	0.02	0.03	1.0
53	0.02	0.03	0.7
19	0.01	0.03	2.0*
14	0.00	0.03	0.9
31	0.00	0.03	0.9
45	0.00	0.03	0.8
52	0.00	0.03	0.4
20	-0.01	0.03	0.9
4	-0.02	0.03	0.9
5	-0.02	0.03	0.8
34	-0.02	0.03	1.3
47	-0.02	0.03	0.4
50	-0.02	0.03	1.0
26	-0.03	0.03	1.3
43	-0.03	0.03	1.3
54	-0.03	0.03	0.3
1	-0.04	0.03	1.0
13	-0.04	0.03	1.5
15	-0.04	0.03	1.0
18	-0.04	0.03	0.8
46	-0.04	0.03	0.8

16	-0.05	0.03	1.4
33	-0.05	0.03	0.9
40	-0.05	0.03	0.6
2	-0.08	0.03	0.9
38	-0.08	0.03	0.7
51	-0.08	0.03	0.6
7	-0.09	0.03	0.9
32	-0.09	0.03	1.4
36	-0.11	0.03	0.7
24	-0.12	0.03	1.7
49	-0.13	0.04	1.1
28	-0.17	0.03	1.1
M	0.00	0.03	1.0
S	0.07	0.00	0.4

Note. Reliability = 0.81; separation index = 2.09; fixed (all same) chi-square = 297.6; significance = 0.00
* = Misfitting item

Bias Analysis of Raters across Facets

FACETS analysis also allows us to categorize certain interactions that show systemic patterns between a particular rater and the other facets using bias analysis. The bias analysis further examines the residuals, the differences between the expected and the observed scores, to investigate additional sub-patterns of bias. For example, there might be a consistent relationship between a certain rater and a certain speech act if a that rater consistently scores more severely on that speech act than expected. In bias analysis, as a statistical summary, the *z*-score is used to represent the degree of the difference between what might have been predicted and what was actually observed. By convention, *z*-scores either above +2.00 or below -2.00 signal significant bias; a negative value indicates raters were more lenient than expected, and a positive value indicates they were harsher than expected. In the bias analysis of this study, four interactions, raters and task types, raters and items, raters and speech acts, and raters and examinees, were investigated.

Rater bias across the three task types. The results from the bias analysis of the interaction between the three raters and the three task types are found in Table 16. There were nine total interactions from the three raters and the three task types. The first and second columns show each rater and each task type, and the third and fourth columns represent the total expected scores and the total actually observed scores, respectively. These scores are totaled across all examinees (*N* = 24) and the 18 test items of each task; therefore, the average of the difference between the total observed and the total expected score (column 5) can be calculated by dividing the difference by 432 (24 candidates×18 items). The next two columns report a bias logit (column 6), which indicates the extent of residuals, and the estimate of the error (column 7). Also, the *z*-scores are shown in column 8. Applying the standards of the *z*-scores, the *z*-scores themselves were not significant, which means that all *z*-scores were within the range of between -2.00 and +2.00. Lastly, in column 9, the infit mean square, which shows the degrees of the consistency in this bias pattern, is reported. The mean and standard deviation for the infit mean square value are 1.1 and 0.3, respectively. Therefore, all nine interactions are all within the range of two standard deviations around the mean (1.1±(0.3×2)=0.5~1.7), and they showed a consistent bias pattern. Figure 3 graphically shows the results of the bias analysis between the three raters and the three task types in terms of the *z*-score. Although the *z*-scores themselves were not significant, it is found that the three raters showed quite different patterns on each task type. Rater 1 showed more leniencies on the OPDCT than on the LLDCT; rater 2 showed more harshness on the RP than on the LLDCT; and rater 3 showed more harshness on the OPDCT than on the RP.

Table 16
Bias Calibration Report: Interaction between Raters and Task types

Rater	Task Type	Observed Score	Expected Score	Obsered-expected average	Bias (logits)	Error	*z*-score	Infit (mean square)
1	OPDCT	11655	11565.7	0.21	-0.03	0.02	-1.51	1.6
1	LLDCT	11393	11467.5	-0.17	0.02	0.02	1.21	1.6
1	RP	11832	11846.9	-0.03	0.01	0.02	0.29	1.2
2	OPDCT	9536	9518.9	0.04	0.00	0.01	-0.19	1.2
2	LLDCT	9356	9303.9	0.12	-0.01	0.01	-0.58	1.0
2	RP	10117	10186.3	-0.16	0.01	0.01	0.85	0.8
3	OPDCT	8173	8279.5	-0.25	0.01	0.01	1.09	0.8
3	LLDCT	8047	8024.7	0.05	0.00	0.01	-0.23	1.0
3	RP	9186	9101.9	0.19	-0.01	0.01	-0.92	0.9

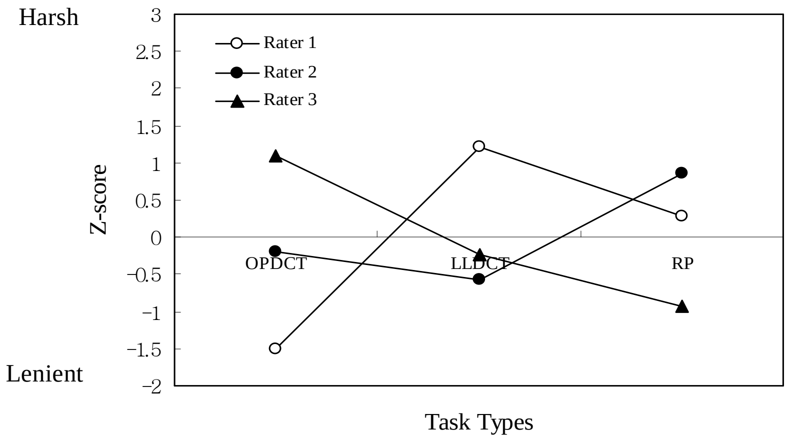


Figure 3. Bias analysis between raters and task types

Rater bias across test item difficulty. Besides the interaction between the three raters and the three task types, the interaction between the three raters and the test items of each task type was also investigated. As such, the bias analysis was focused on the relationship between the test item difficulty of each task type and the rater severity, to examine whether the raters score more harshly or leniently on certain items. Tables 17, 18, and 19 show the bias calibration report for the interaction between the three raters and the test items within each of the three task types, and all rows are sorted by the *z*-score. There were 162 total productions from the three raters, and the 18 test items within each of the three tasks (3×18 ×3 = 162). Tables 17, 18, and 19 list each rater (column 1), the total observed and the total expected score (columns 2 and 3), the average of the difference between the total observed and expected score (column 4), a bias logit (column 5), and the estimate of the error (column 6). Also, the z-scores, a crucial indicator for investigating the bias pattern, are also shown in column 7, followed by the infit mean square (column 8). In columns 9 and 10, items of each task type and these item difficulty logits are presented, respectively.

Numerous rater×item interactions that show significant bias are found in Tables 17, 18, and 19, although there were no significant overall biases between the raters and the three task types. In Table 17, which shows the interaction between the three raters and the OPDCT items, there are three interactions that show significant bias out of 54 interactions, either above +2.00 or below -2.00. No interaction among these three is identified as misfitting. However, two interactions (rater 1×item 8, rater 1×item 13) out of 54 interactions, are misfitting interactions because these are outside of the range between -0.9 and 2.3 (1.1± [0.3×2]), based on the mean (1.1) and the standard deviation (0.6) of infit mean square. In Table 18, which shows the interaction between the three raters and the LLDCT items, there are six interactions with significant bias out of 54 interactions. However, among these six interactions, one interaction (rater 1×item 19) should not be counted because this interaction is identified as misfitting with the infit value of 3.1. In addition to this interaction, there are two more interactions (rater 1×item 24, rater 1×item 30) that are misfitting because of the higher infit value. More significant *z*-scores are found in Table 19, which shows the interaction between the three raters and the RP items. There are ten total interactions that show significant bias, which have *z*-scores either above +2.00 or below -2.00. However, among these ten interactions, one interaction (rater 1×item 41) with the infit value of 2.0 is misfitting because the normal infit values range is between -0.2 and 1.8 (0.8±[0.5×2]), based on the mean (0.8) and the standard deviation (0.5) of the infit values. Therefore this interaction should not be counted. Also, there are two more mistiffing interactions (rater 1×item 50, rater 3×item 42) because of the slightly higher infit value which was above the normal range. Overall, misfitting interactions almost evenly appeared in the three task types, however interestingly, most of these misfitting interactions are from rater 1.

Table 17

Bias Calibration Report: Interaction between Three Raters and OPDCT Items

Rater	Observed Score	Expected Score	Observed-expected Average	Bias (logits)	Error	*z*-score	Infit (mean square)	Item	Item Difficulty Logits
1	681	649.3	1.32	-0.26	0.11	**-2.33**	1.1	5	-0.02
1	688	660.6	1.14	-0.29	0.13	**-2.25**	1.3	7	-0.09
2	537	501.5	1.48	-0.08	0.05	-1.63	1.9	8	0.06
3	449	410.6	1.60	-0.07	0.04	-1.60	0.5	17	0.10
3	440	402.2	1.57	-0.07	0.04	-1.57	1.0	3	0.11
2	578	548.7	1.22	-0.08	0.05	-1.45	1.6	15	-0.04
1	676	658.5	0.73	-0.15	0.10	-1.42	0.7	2	-0.08
2	538	509.7	1.18	-0.06	0.05	-1.31	1.1	12	0.05
2	595	570.6	1.02	-0.07	0.06	-1.27	1.1	7	-0.09
2	574	552.4	0.90	-0.06	0.05	-1.07	2.1	16	-0.05
1	650	638.4	0.48	-0.06	0.08	-0.81	1.8	9	0.03
1	638	626.0	0.50	-0.05	0.07	-0.77	0.8	10	0.09
1	644	634.4	0.40	-0.05	0.07	-0.65	1.7	6	0.05
1	659	651.5	0.31	-0.05	0.08	-0.57	1.0	18	-0.04
2	561	549.4	0.48	-0.03	0.05	-0.57	1.0	1	-0.04
3	428	415.4	0.53	-0.02	0.04	-0.53	1.1	10	0.09
1	649	641.9	0.29	-0.04	0.08	-0.50	1.0	11	0.02
1	659	653.4	0.23	-0.04	0.08	-0.44	0.7	16	-0.05
2	554	546.2	0.33	-0.02	0.05	-0.39	1.6	13	-0.04
3	475	466.0	0.38	-0.02	0.04	-0.39	0.6	14	0.00
1	651	645.7	0.22	-0.03	0.08	-0.38	2.3	14	0.00
3	479	472.4	0.27	-0.01	0.04	-0.29	0.5	4	-0.02
2	545	539.6	0.23	-0.01	0.05	-0.26	1.2	4	-0.02
3	482	476.1	0.25	-0.01	0.04	-0.26	0.6	5	-0.02
1	654	652.1	0.08	-0.01	0.08	-0.14	0.7	1	-0.04
2	528	525.6	0.10	-0.01	0.05	-0.12	1.2	11	0.02
1	625	624.0	0.04	0.00	0.06	-0.06	1.3	17	0.10
2	548	547.9	0.00	0.00	0.05	0.00	1.0	18	-0.04
1	651	651.8	-0.03	0.00	0.08	0.06	0.8	15	-0.04
2	507	508.9	-0.08	0.00	0.05	0.09	0.6	6	0.05
1	626	628.8	-0.11	0.01	0.07	0.18	3.0*	8	0.06
1	646	648.4	-0.10	0.01	0.08	0.19	3.3*	13	-0.04
2	561	565.3	-0.18	0.01	0.05	0.22	1.0	2	-0.08
3	476	481.4	-0.22	0.01	0.04	0.24	0.8	13	-0.04
2	512	517.6	-0.23	0.01	0.05	0.26	1.3	9	0.03
3	440	446.0	-0.25	0.01	0.04	0.26	1.1	9	0.03
3	428	435.7	-0.32	0.01	0.04	0.33	0.5	6	0.05
3	475	482.6	-0.32	0.01	0.04	0.34	0.6	18	-0.04
3	446	455.5	-0.40	0.02	0.04	0.41	0.4	11	0.02
3	423	436.6	-0.57	0.02	0.04	0.58	0.6	12	0.05
3	471	484.5	-0.56	0.03	0.04	0.60	1.1	1	-0.04

3	491	504.2	-0.55	0.03	0.05	0.60	0.8	2	-0.08
2	520	534.3	-0.59	0.03	0.05	0.69	0.8	14	0.00
1	636	648.0	-0.50	0.06	0.07	0.89	1.4	4	-0.02
2	458	479.7	-0.90	0.04	0.04	0.97	1.4	3	0.11
1	620	634.7	-0.61	0.06	0.06	0.99	1.3	12	0.05
1	602	618.1	-0.67	0.06	0.06	1.01	2.1	3	0.11
2	467	491.6	-1.02	0.05	0.04	1.11	1.2	10	0.09
3	461	488.2	-1.13	0.05	0.04	1.22	1.1	16	-0.05
3	455	483.5	-1.19	0.06	0.04	1.27	0.6	15	-0.04
3	395	427.7	-1.36	0.06	0.04	1.38	0.8	8	0.06
2	448	487.4	-1.64	0.08	0.04	1.77	0.6	17	0.10
2	505	542.6	-1.57	0.08	0.05	1.83	0.8	5	-0.02
3	459	510.8	-2.16	0.10	0.04	**2.39**	0.6	7	-0.09

Note. * = Misfitting

Table 18
Bias Calibration Report: Interaction between Three Raters and LLDCT Items

Rater	Observed Score	Expected Score	Observed-expected Average	Bias (logits)	Error	*z*-score	Infit (mean square)	Item	Item Difficulty Logits
2	605	559.6	1.89	-0.13	0.06	**-2.28**	0.7	36	-0.11
3	445	392.4	2.19	-0.09	0.04	**-2.16**	1.0	23	0.09
2	569	525.6	1.81	-0.11	0.05	**-2.05**	1.2	34	-0.02
3	428	392.0	1.50	-0.06	0.04	-1.48	0.9	27	0.09
2	559	528.6	1.27	-0.07	0.05	-1.45	1.3	26	-0.03
3	454	422.4	1.32	-0.06	0.04	-1.33	0.9	29	0.04
2	530	505.5	1.02	-0.05	0.05	-1.12	1.0	35	0.02
3	412	386.7	1.06	-0.04	0.04	-1.04	0.7	21	0.09
1	648	632.8	0.64	-0.08	0.07	-1.02	1.3	35	0.02
1	676	665.0	0.46	-0.10	0.10	-0.97	1.0	28	-0.17
3	458	436.2	0.91	-0.04	0.04	-0.93	1.9	19	0.01
1	645	631.7	0.55	-0.06	0.07	-0.88	1.4	22	0.02
1	649	636.3	0.53	-0.07	0.08	-0.87	1.6	31	0.00
1	665	656.3	0.36	-0.06	0.09	-0.70	0.4	36	-0.11
3	445	429.1	0.66	-0.03	0.04	-0.67	0.8	22	0.02
2	578	565.3	0.53	-0.03	0.05	-0.65	1.3	24	-0.12
2	532	518.4	0.57	-0.03	0.05	-0.64	0.7	20	-0.01
1	647	638.7	0.35	-0.04	0.07	-0.58	1.0	20	-0.01
1	655	647.2	0.32	-0.05	0.08	-0.58	1.2	33	-0.05
3	432	418.9	0.55	-0.02	0.04	-0.55	1.1	30	0.04
2	547	537.7	0.39	-0.02	0.05	-0.45	1.1	33	-0.05
3	435	425.1	0.41	-0.02	0.04	-0.42	1.0	25	0.03
2	518	509.3	0.36	-0.02	0.05	-0.40	1.2	19	0.01
2	562	554.7	0.30	-0.02	0.05	-0.36	1.8	32	-0.09
3	531	525.7	0.22	-0.01	0.05	-0.25	0.7	28	-0.17
1	619	615.8	0.14	-0.01	0.06	-0.20	0.9	27	0.09
3	494	491.0	0.13	-0.01	0.05	-0.14	1.0	32	-0.09
2	501	499.9	0.05	0.00	0.05	-0.05	0.7	25	0.03
3	439	440.6	-0.07	0.00	0.04	0.07	0.8	31	0.00

3	502	504.2	-0.09	0.00	0.05	0.10	1.3	24	-0.12
2	465	471.7	-0.28	0.01	0.04	0.29	0.8	23	0.09
2	487	494.0	-0.29	0.01	0.05	0.32	1.1	30	0.04
1	619	625.2	-0.26	0.03	0.06	0.40	2.8*	30	0.04
2	502	513.1	-0.46	0.02	0.05	0.51	0.7	31	0.00
1	634	641.9	-0.33	0.04	0.07	0.56	1.3	34	-0.02
1	601	611.2	-0.42	0.04	0.06	0.62	2.1	21	0.09
2	451	466.2	-0.63	0.03	0.04	0.67	0.9	21	0.09
1	619	630.1	-0.46	0.05	0.06	0.73	1.0	25	0.03
1	633	643.3	-0.43	0.05	0.07	0.74	1.7	26	-0.03
3	453	470.1	-0.71	0.03	0.04	0.75	0.5	33	-0.05
1	617	629.0	-0.50	0.05	0.06	0.78	0.6	29	0.04
1	644	654.3	-0.43	0.06	0.07	0.81	1.6	32	-0.09
1	648	658.5	-0.44	0.06	0.07	0.86	3.4*	24	-0.12
2	566	582.3	-0.68	0.04	0.05	0.87	1.4	28	-0.17
3	439	459.2	-0.84	0.04	0.04	0.88	1.0	26	-0.03
2	478	497.6	-0.82	0.04	0.04	0.89	0.7	29	0.04
3	425	446.9	-0.91	0.04	0.04	0.94	1.1	20	-0.01
2	474	503.3	-1.22	0.06	0.04	1.34	0.6	22	0.02
3	420	455.5	-1.48	0.06	0.04	1.54	1.2	34	-0.02
3	392	431.7	-1.66	0.07	0.04	1.68	0.7	35	0.02
2	432	471.3	-1.64	0.07	0.04	1.73	0.3	27	0.09
1	604	634.5	-1.27	0.12	0.06	**2.05**	3.1*	19	0.01
3	443	497.1	-2.25	0.10	0.04	**2.45**	0.5	36	-0.11
1	570	615.9	-1.91	0.14	0.05	**2.77**	1.6	23	0.09

Note. * = Misfitting

Table 19

Bias Calibration Report: Interaction between Three Raters and RP Items

Rater	Observed	Expected	Observed-	Bias	Error	*z*-score	Infit	Item	Item
	score	score	expected	(logits)			(mean		difficulty
			average				square)		logits
1	696	667.2	1.20	-0.40	0.16	**-2.53**	0.6	40	-0.05
1	698	675.4	0.94	-0.37	0.17	**-2.24**	0.3	49	-0.13
3	556	507.5	2.02	-0.11	0.05	**-2.22**	0.5	45	0.00
3	449	395.8	2.22	-0.09	0.04	**-2.19**	1.8	41	0.21
2	615	579.3	1.49	-0.11	0.06	-1.89	1.5	43	-0.03
3	567	528.1	1.62	-0.09	0.05	-1.84	0.2	46	-0.04
1	682	663.2	0.78	-0.18	0.11	-1.61	0.9	47	-0.02
2	638	611.4	1.11	-0.11	0.07	-1.58	0.6	49	-0.13
3	505	474.3	1.28	-0.06	0.05	-1.36	0.7	48	0.07
3	546	517.0	1.21	-0.07	0.05	-1.35	1.0	50	-0.02
1	670	654.0	0.67	-0.12	0.09	-1.25	0.5	37	0.04
1	678	663.9	0.59	-0.13	0.11	-1.22	1.1	43	-0.03
1	683	670.3	0.53	-0.14	0.12	-1.19	0.5	38	-0.08

2	543	520.3	0.95	-0.05	0.05	-1.07	1.1	42	0.12
1	653	639.6	0.56	-0.07	0.08	-0.94	1.8	42	0.12
1	679	669.9	0.38	-0.09	0.11	-0.85	0.5	51	-0.08
3	509	490.1	0.79	-0.04	0.05	-0.85	0.4	37	0.04
2	491	474.7	0.68	-0.03	0.05	-0.72	1.1	41	0.21
3	558	543.3	0.61	-0.04	0.05	-0.72	0.6	51	-0.08
3	539	524.7	0.59	-0.03	0.05	-0.67	0.2	54	-0.03
1	670	662.5	0.31	-0.06	0.09	-0.64	2.1*	50	-0.02
3	524	510.8	0.55	-0.03	0.05	-0.61	0.3	52	0.00
1	659	651.7	0.31	-0.05	0.08	-0.56	0.3	44	0.05
2	557	548.3	0.36	-0.02	0.05	-0.43	0.5	44	0.05
2	565	557.7	0.30	-0.02	0.05	-0.37	0.6	39	0.03
2	568	561.5	0.27	-0.02	0.05	-0.33	0.7	53	0.02
3	502	494.7	0.30	-0.02	0.05	-0.33	0.9	39	0.03
2	570	570.6	-0.03	0.00	0.05	0.03	0.5	52	0.00
2	580	581.5	-0.06	0.00	0.05	0.08	0.2	54	-0.03
1	658	659.6	-0.06	0.01	0.08	0.13	0.4	45	0.00
2	538	541.1	-0.13	0.01	0.05	0.15	0.6	48	0.07
1	655	657.0	-0.08	0.01	0.08	0.16	0.4	53	0.02
2	594	596.9	-0.12	0.01	0.06	0.16	0.9	38	-0.08
3	495	499.4	-0.19	0.01	0.05	0.20	0.8	53	0.02
3	526	533.5	-0.31	0.02	0.05	0.36	0.3	40	-0.05
2	570	577.4	-0.31	0.02	0.05	0.39	0.3	47	-0.02
3	535	544.8	-0.41	0.02	0.05	0.48	0.6	38	-0.08
3	508	519.4	-0.48	0.02	0.05	0.54	0.5	47	-0.02
3	467	483.1	-0.67	0.03	0.04	0.72	1.0	44	0.05
2	570	584.2	-0.59	0.04	0.05	0.77	0.6	46	-0.04
1	648	660.6	-0.52	0.08	0.07	1.05	0.2	52	0.00
1	652	664.7	-0.53	0.09	0.08	1.11	0.4	54	-0.03
1	641	655.5	-0.60	0.08	0.07	1.15	1.3	39	0.03
2	567	588.3	-0.89	0.06	0.05	1.16	0.9	40	-0.05
2	572	595.8	-0.99	0.07	0.05	1.33	0.6	51	-0.08
3	413	449.2	-1.51	0.06	0.04	1.55	2.1*	42	0.12
2	519	554.0	-1.46	0.08	0.05	1.74	0.9	37	0.04
2	539	575.5	-1.52	0.09	0.05	1.92	0.5	50	-0.02
1	621	648.7	-1.15	0.13	0.06	**2.05**	1.6	48	0.07
1	641	665.7	-1.03	0.15	0.07	**2.17**	1.2	46	-0.04
3	472	521.9	-2.08	0.10	0.04	**2.34**	1.0	43	-0.03
2	521	568.0	-1.96	0.11	0.05	**2.41**	0.7	45	0.00
3	515	564.2	-2.05	0.12	0.05	**2.50**	1.1	49	-0.13
1	548	617.5	-2.90	0.21	0.05	**4.18**	2.0*	41	0.21

Note. * = Misfitting

Figures 4, 5, and 6 illustrate the relationship between the three rater severities and the test items difficulties using scatterplots. In each figure, each *z*-score/item difficulty logit set is plotted using a black dot, and the item difficulty logits are also plotted as a dark black line. The interesting phenomena in these interactions are not only shown in the many significant *z*-score values, either above +2.00 or below -2.00, but are also revealed in the unique relationships between each rater and the test items. In the interactions between the items of the three task types and the three raters, shown in Figures 4, 5, and 6, each rater showed distinctive bias patterns within items of the three speech acts.

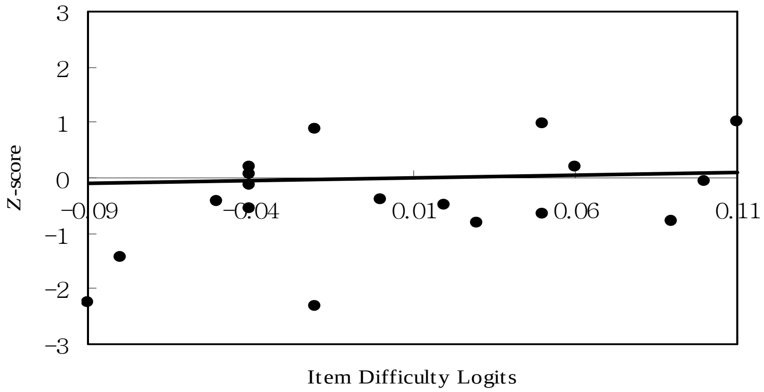


Figure 4.1. Interaction between rater 1 and the test items of OPDCT

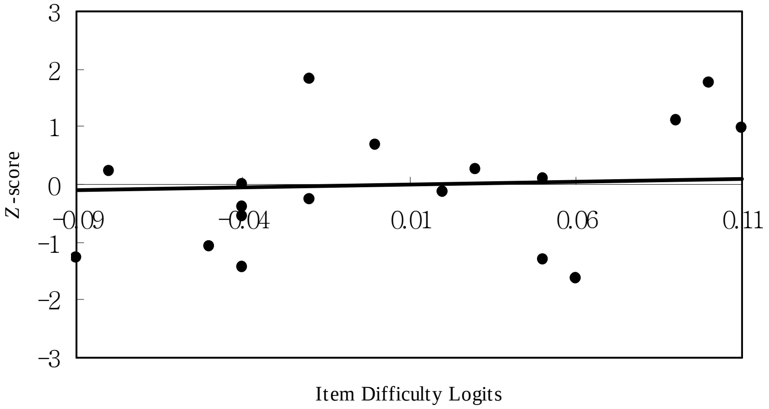


Figure 4.2. Interaction between rater 2 and the test items of OPDCT

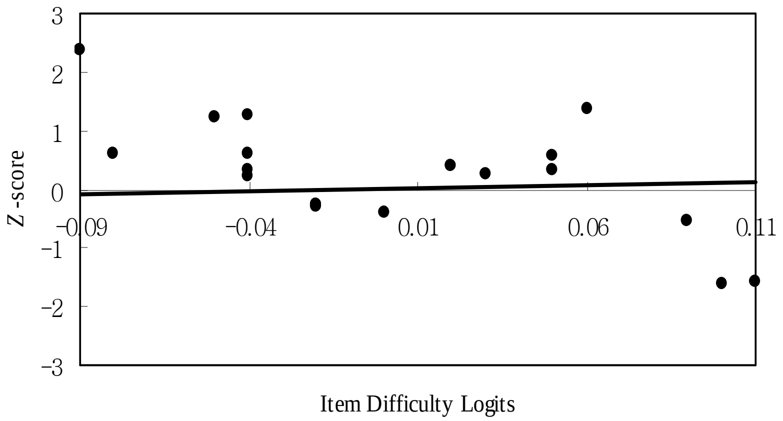


Figure 4.3. Interaction between rater 3 and the test items of OPDCT

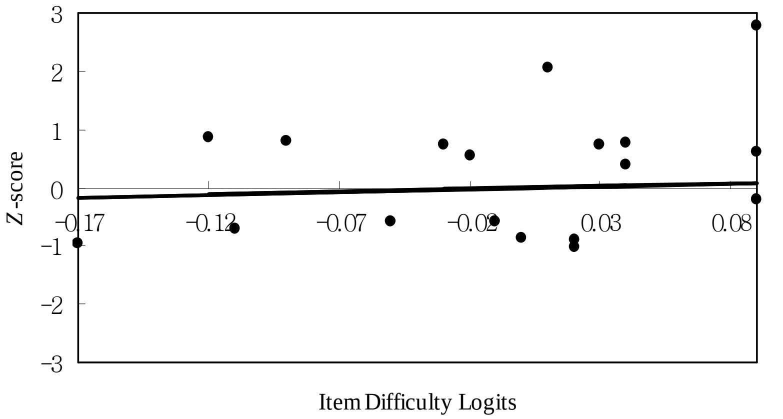


Figure 5.1. Interaction between rater 1 and the test items of LLDCT

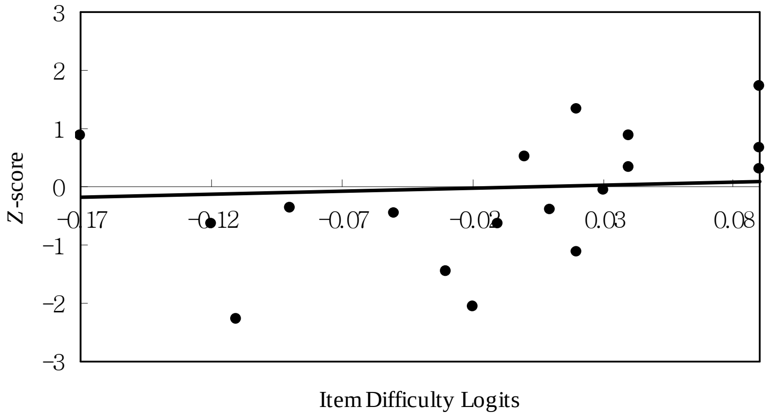


Figure 5.2. Interaction between rater 2 and the test items of LLDCT

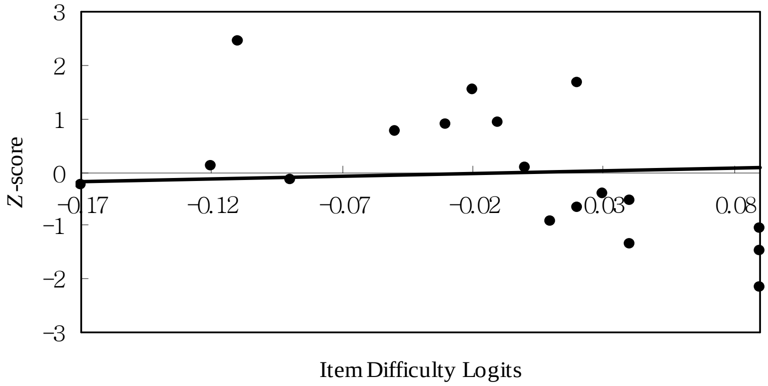


Figure 5.3. Interaction between rater 3 and the test items of LLDCT

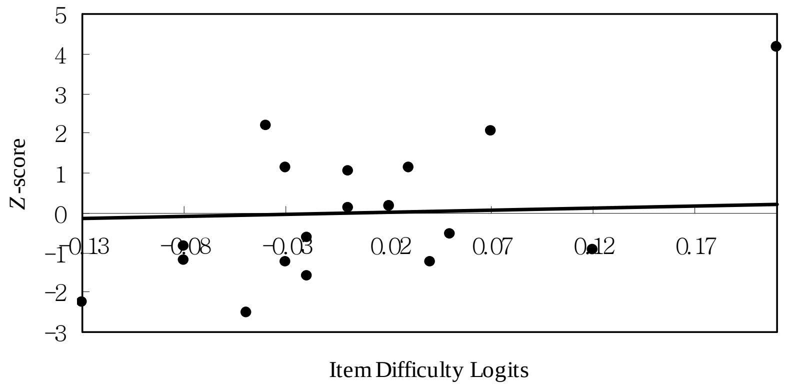


Figure 6.2. Interaction between rater 2 and the test items of RP

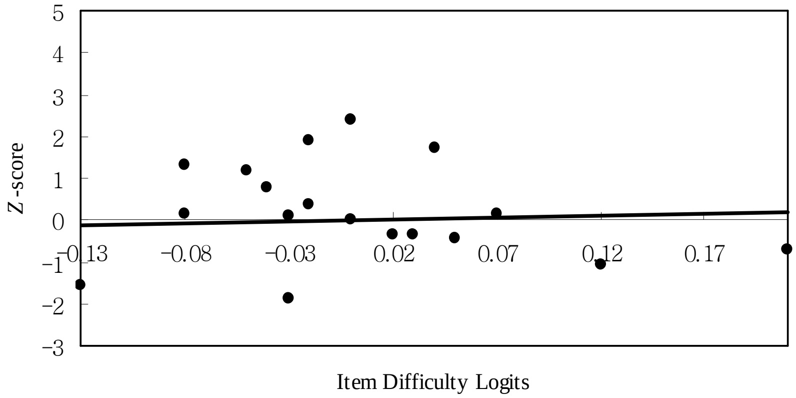


Figure 6.3. Interaction between rater 3 and the test items of RP

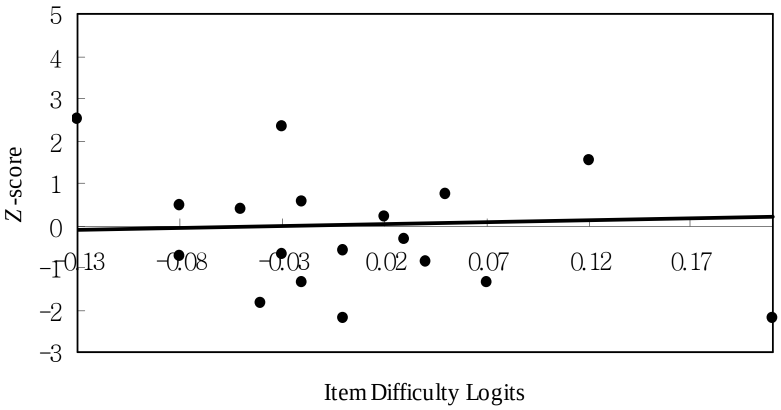


Figure 6.1. Interaction between rater 1 and the test items of RP

Firstly, in the OPDCT as shown in Figures 4.1, 4.2 and 4.3, rater 1 and rater 3 showed the opposite bias pattern on the easiest item (item 7, request). Rater 1 was more lenient on the easiest item of the OPDCT than expected with a *z*-score of -2.33; rater 3 was harsher on the same item than expected with a *z*-score of 2.39. Secondly, in the LLDCT as shown in Figures 5.1, 5.2 and 5.3, rater 1 and rater 3 again showed the opposite bias pattern on the most difficult item (item 23, apology). Rater 1 was harsher on the most difficult item of the LLDCT than expected with a *z*-score of 2.77; rater 3 was more lenient on the same item than expected with a *z*-score of -2.16. Lastly, in the RP as shown in Figures 6.1, 6.2 and 6.3, rater 1 and rater 3 again showed the opposite bias patterns on both the easiest item (item 49, apology) and the most difficult item (item 41, apology). Rater 1 was more lenient on the easiest item of the RP than expected with a *z*-score -2.24; rater 3 was harsher on the same item than expected with a *z*-score of 2.50. Also, rater 1 was harsher on the most difficult item of the RP than expected with a *z*-score of 4.18; rater 3 was more lenient on the same item than expected with a *z*-score of -2.19. In summary, especially, rater 1 and rater 3 showed the opposite bias patterns on the easiest items of the OPDCT and RP and the most difficult items of the LLDCT and RP as well. Also, interestingly, these bias interactions for the LLDCT and RP were all shown in apology although the bias interaction on the OPDCT was from request.

Focusing on the significant *z*-score values in detail, the summary of the frequency of these values within the three task types’ test items is provided in Table 20. In the first column, the test items are categorized as the three task types, and in the second and third columns, the numbers of significant interactions from each rater, which are either above a *z*-score of 2.0 or below -2.0, are shown. The last column indicates the total frequency of significantly biased interactions within the three task types. The test items on the RP showed the most frequent biased interactions (53%), compared with the test items of OPDCT (16%) and LLDCT (32%).

Table 20
Frequency of Significantly Biased Interactions between Raters and Test Item

Task Type		Harsher than expected (greater than z-score 2.0)			More lenient than expected (below z-score -2.0)			Total number of significantly biased interactions (% of biased interactions)
		R1	R2	R3	R1	R2	R3	
OPDCT	Number of items	0	0	1	2	0	0	3 (16%)
LLDCT	Number of items	2	0	1	0	2	1	6 (32%)
RP	Number of items	3	1	2	2	0	2	10 (53%)

Rater bias across speech acts. Table 21 shows the bias calibration report for the raters and the speech acts. There are nine total interactions from the three raters and the three speech acts: refusal, apology, and request. The first two columns represent each rater and each speech act. The next three columns (columns 3, 4, and 5) indicate the total observed and the total expected scores, and the average of the difference between these scores. Also listed are bias logits (column 6), error of the bias estimates (column 7), the converted *z*-scores (column 8), and the fit statistics (column 9). The rows of Table 21 are sorted by *z*-score values and there are numerous significant *z*-scores either below -2.00 or above 2.00, indicating that the raters either scored the certain speech act more harshly or leniently than expected. There are six total interactions that show the significant bias out of nine interactions. However, one interaction (rater 1×apology) should not be counted because this interaction is slightly misfitting with an infit value of 1.8, considering that the normal range of the infit value is between 0.5 and 1.7 (1.1± [0.3×2]), based on the mean (1.1) and the standard deviation (0.3) for the infit values.

Figure 7 illustrates the relationship between the three raters and the three speech acts based on the *z*-score. The three raters show quite different bias patterns in terms of their severity on each of the three speech acts, as shown in Figure 7. In refusals, rater 2 and rater 3 showed the opposite bias pattern. Rater 2 was more lenient on refusals than expected, with a *z*-score of -2.36; rater 3 was harsher on refusal than expected, with a *z*-score of 2.28. In apology, only rater 1 showed a significant bias, harsher than expected with a *z*-score of 2.14. In requests, rater 2 and rater 3 again showed the opposite bias pattern. Rater 2 was harsher on requests than expected, with a *z*-score of 3.40; rater 3 was more lenient on requests than expected, with a *z*-score of -2.02.

Table 21
Bias Calibration Report: Interaction between Raters and Speech Acts

Rater	Speech Acts	Observed Score	Expected Score	Observed -expected average	Bias (logits)	Error	*z*-score	Infit (mean square)
2	Refusal	9434	9236.6	0.48	-0.03	0.01	**-2.36**	1.1
3	Request	8941	8740.1	0.44	-0.02	0.01	**-2.02**	0.8
1	Request	12304	12195.8	0.24	-0.03	0.02	-1.78	1.3
2	Apology	9844	9732.3	0.26	-0.02	0.01	-1.29	1.0
1	Refusal	11043	11029.2	0.03	0.00	0.02	-0.25	1.4
3	Apology	8555	8544.8	0.02	0.00	0.01	-0.11	1.1
1	Apology	11533	11655.0	-0.28	0.04	0.02	**2.14**	1.8*
3	Refusal	7910	8121.2	-0.52	0.02	0.01	**2.28**	0.9
2	Request	9731	10040.2	-0.68	0.04	0.01	**3.40**	0.9

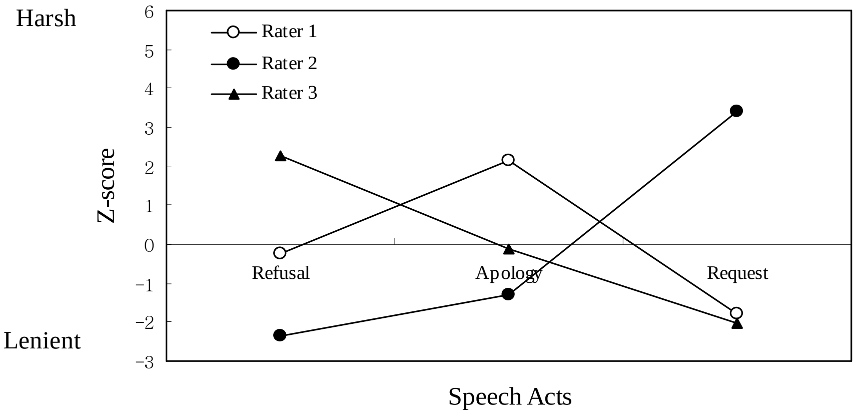


Figure 7. Bias analysis between raters and speech acts

Rater bias across examinee abilities. Besides the three task types, the test items, and the three speech acts, the relationship between the rater severities and the examinee abilities was also investigated. There were 72 total productions from three raters and twenty-four candidates (3×24 = 72). Tables 22, 23, and 24 show the bias calibration report of the interaction between the raters and the candidate abilities. In each of Tables 22, 23, and 24, the first two columns show examinees (column 1) and their ability logits (column 2). The next three columns show the total observed scores (column 3), the total expected scores (column 4), and the average of the difference between these scores (column 5). Also, bias logits (column 6), the estimate of the error (column 7), the converted *z*-scores (column 8), and the infit values (column 9) are listed. The rows of these tables are sorted by the *z*-score values.

Numerous significant *z*-score values are found in Tables 22, 23, and 24. In Table 22, which shows the interaction between rater 1 and examinee abilities, there are eight interactions that have significant *z*-scores out of 24 interactions. None of these interactions are misfitting because the infit values are within the normal range between 0.0 and 2.8 (1.4±[0.7×2]), based on the mean (1.4) and the standard deviation (0.7) for the infit values. There is one interaction (rater 1 ×examinee 19) identified as misfitting because of the high infit value of 3.1. In Table 23, which shows the interaction between rater 2 and examinee abilities, there are four interactions that show significant bias. However, among these interaction, one interaction (rater 2×examinee 6) should not be counted because this is slightly misfitting with the infit value of 1.8, considering the normal infit value range is between 0.4 and 1.6 (1.0± [0.3×2]), based on the mean (1.0) and the standard deviation (0.3) for the infit values. Lastly, two significant *z*-scores are found in Table 24, which shows the interaction between rater 3 and examinee abilities. Also, none of these interactions are misfitting because the infit values are within the normal range between 0.3 and 1.5 (0.9±[0.3×2]), based on the mean (0.9) and the standard deviation (0.3) for the infit values.

To display the bias patterns between the rater severities and the examinee abilities, each set of an examinee ability logits and *z*-scores are plotted as a dark black dot in a scatterplot, and the examinee abilities logit values are plotted as a line as seen in Figures 8.1, 8.2, and 8.3. Interestingly, the unique bias patterns between the three raters and the examinee abilities were found. Especially, rater 2 and rater 3 showed the opposite bias pattern in their severities on the most able examinee performance. Rater 2 was harsher than expected on the most able examinee (ability logit = 0.5), with a *z*-score of -2.80 shown in Figure 8.2; rater 3 was more lenient than expected on the same examinee, with a *z*-score of 3.76 shown in Figure 8.3. On the other hand, Rater 1 showed significant bias patterns across various examinee abilities either harsher or more lenient than expected.

Table 22
Bias Calibration Report: Interaction between Rater 1 and Examinee Abilities

Examnee	Ability (logits)	Observed Score	Expected Score	Observed-expected average	Bias (logits)	Error	*z*-score	Infit (mean square)
23	0.50	1565	1525.3	0.74	-0.29	0.1	**-2.76**	2.7
15	0.24	1501	1447.8	0.99	-0.15	0.06	**-2.52**	0.5
18	0.26	1506	1454.2	0.96	-0.16	0.06	**-2.51**	0.9
13	0.34	1522	1483.5	0.71	-0.14	0.07	**-2.11**	1.3
16	0.36	1524	1490.4	0.62	-0.13	0.07	-1.90	1.1
9	0.18	1460	1417.3	0.79	-0.09	0.05	-1.85	0.9
14	0.31	1504	1473.3	0.57	-0.10	0.06	-1.61	1.3
11	0.29	1498	1468.6	0.54	-0.09	0.06	-1.51	0.7
17	0.22	1464	1439.1	0.46	-0.06	0.05	-1.16	1.2
7	0.29	1486	1466.1	0.37	-0.06	0.06	-1.01	1.2
21	0.09	1387	1366.3	0.38	-0.03	0.04	-0.81	0.8
5	0.39	1511	1499.8	0.21	-0.04	0.06	-0.67	1.0
2	0.20	1431	1426.6	0.08	-0.01	0.05	-0.20	0.7
20	0.43	1507	1510.7	-0.07	0.01	0.06	0.23	1.1
24	0.30	1459	1471.1	-0.22	0.03	0.05	0.63	0.8
10	0.33	1468	1482.0	-0.26	0.04	0.05	0.77	1.1
8	0.15	1373	1395.3	-0.41	0.04	0.04	0.94	2.5

22	0.40	1482	1501.6	-0.36	0.06	0.05	1.18	2.6
1	0.28	1434	1462.0	-0.52	0.06	0.05	1.41	1.1
19	0.24	1404	1437.6	-0.62	0.07	0.04	1.59	3.1*
12	0.02	1238	1314.0	-1.41	0.09	0.03	**2.71**	1.5
6	0.33	1430	1483.3	-0.99	0.13	0.05	**2.91**	2.2
3	0.14	1314	1390.8	-1.42	0.11	0.04	**3.14**	1.7
4	0.31	1412	1473.5	-1.14	0.14	0.04	**3.21**	1.5

Table 23

Bias Calibration Report: Interaction between Rater 2 and Examinee Abilities

Examnee	Ability	Observed	Expected	Observed-	Bias	Error	*z*-score	Infit
	(logits)	Score	Score	expected	(logits)			(mean
				average				square)
20	0.43	1426	1348.2	1.44	-0.13	0.04	**-2.91**	1.3
23	0.50	1460	1390.8	1.28	-0.14	0.05	**-2.80**	1.1
6	0.33	1351	1273.4	1.44	-0.10	0.04	**-2.63**	1.8*
24	0.30	1280	1242.3	0.70	-0.04	0.03	-1.25	0.5
4	0.31	1277	1248.4	0.53	-0.03	0.03	-0.95	1.1
22	0.40	1343	1322.6	0.38	-0.03	0.04	-0.74	1.1
12	0.02	946	929.3	0.31	-0.01	0.03	-0.46	0.7
19	0.24	1183	1173.9	0.17	-0.01	0.03	-0.29	1.5
3	0.14	1075	1071.1	0.07	0.00	0.03	-0.11	1.3
16	0.36	1291	1292.0	-0.02	0.00	0.03	0.04	0.9
8	0.15	1081	1084.0	-0.05	0.00	0.03	0.09	1.4
10	0.33	1267	1270.0	-0.06	0.00	0.03	0.10	1.1
17	0.22	1164	1167.1	-0.06	0.00	0.03	0.10	1.0
21	0.09	1017	1021.1	-0.08	0.00	0.03	0.12	1.1
5	0.39	1311	1317.7	-0.12	0.01	0.04	0.24	1.0
1	0.28	1211	1220.2	-0.17	0.01	0.03	0.30	0.8
13	0.34	1259	1273.8	-0.27	0.02	0.03	0.51	0.9
11	0.29	1218	1236.2	-0.34	0.02	0.03	0.60	0.8
18	0.26	1181	1201.6	-0.38	0.02	0.03	0.66	0.9
2	0.20	1106	1140.0	-0.63	0.03	0.03	1.04	0.9
14	0.31	1210	1248.0	-0.70	0.04	0.03	1.26	0.8
15	0.24	1143	1186.8	-0.81	0.04	0.03	1.38	0.9
7	0.29	1178	1230.1	-0.96	0.05	0.03	1.70	0.7
9	0.18	1031	1120.4	-1.66	0.08	0.03	**2.70**	0.5

Table 24

Bias Calibration Report: Interaction between Rater 3 and Examinee Abilities

Examnee	Ability (logits)	Observed Score	Expected Score	Observed-expected average	Bias (logits)	Error	*z*-score	Infit (mean square)
3	0.14	966	893.1	1.35	-0.06	0.03	-1.99	0.8
12	0.02	792	732.7	1.10	-0.04	0.03	-1.58	1.0
9	0.18	997	950.3	0.86	-0.04	0.03	-1.30	0.5
1	0.28	1107	1069.8	0.69	-0.03	0.03	-1.10	0.6
4	0.31	1137	1104.2	0.61	-0.03	0.03	-0.98	1.4
7	0.29	1114	1081.8	0.60	-0.03	0.03	-0.95	0.7
2	0.20	1003	973.5	0.55	-0.02	0.03	-0.83	1.1
19	0.24	1040	1015.5	0.45	-0.02	0.03	-0.70	1.3
8	0.15	934	908.8	0.47	-0.02	0.03	-0.69	1.2
10	0.33	1148	1131.0	0.31	-0.02	0.03	-0.52	1.1
14	0.31	1111	1103.7	0.14	-0.01	0.03	-0.22	1.0
22	0.40	1197	1197.8	-0.01	0.00	0.03	0.02	1.0
5	0.39	1187	1191.4	-0.08	0.00	0.03	0.14	0.6
15	0.24	1020	1029.4	-0.17	0.01	0.03	0.27	0.5
11	0.29	1078	1089.2	-0.21	0.01	0.03	0.33	0.5
21	0.09	818	834.6	-0.31	0.01	0.03	0.45	0.8
17	0.22	984	1005.8	-0.40	0.02	0.03	0.62	0.4
13	0.34	1112	1135.7	-0.44	0.02	0.03	0.72	0.8
6	0.33	1111	1135.3	-0.45	0.02	0.03	0.74	1.4
24	0.30	1071	1096.7	-0.48	0.02	0.03	0.77	0.9
18	0.26	1016	1047.2	-0.58	0.03	0.03	0.91	0.7
16	0.36	1126	1158.6	-0.60	0.03	0.03	1.01	0.6
20	0.43	1157	1231.2	-1.37	0.08	0.03	**2.42**	1.1
23	0.50	1180	1289.0	-2.02	0.12	0.03	**3.76**	0.7

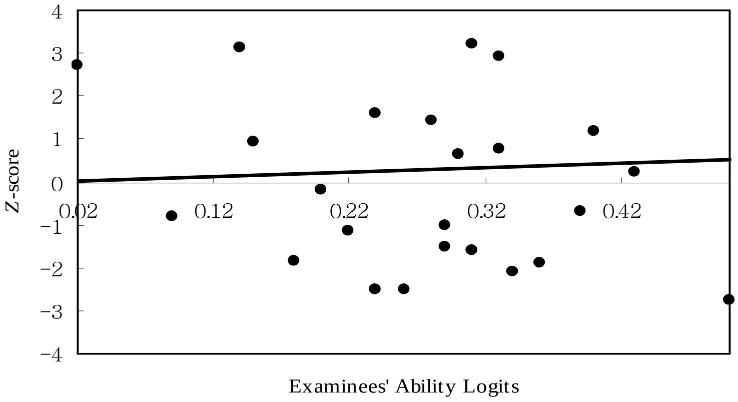


Figure 8.1. Bias analysis between rater 1 and examinee abilities

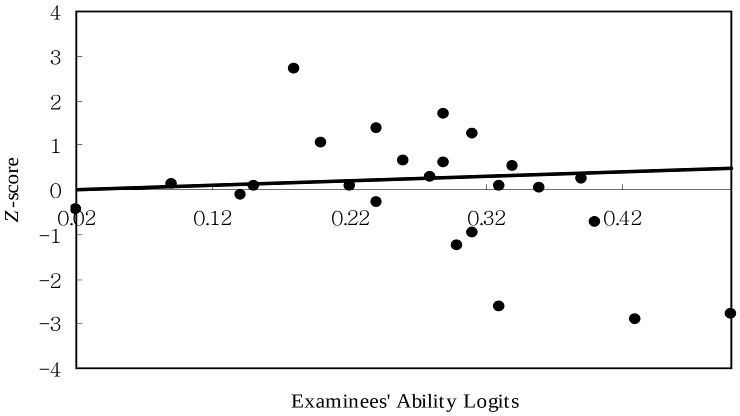


Figure 8.2. Bias analysis between rater 2 and examinee abilities

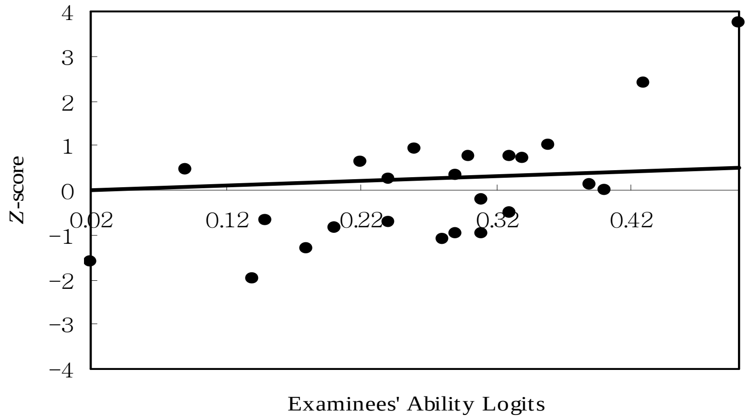


Figure 8.3. Bias analysis between rater 3 and examinee abilities

DISCUSSION

In this section, I will discuss the eight research questions of this study based on the results of the data analysis.

Research question 1: What are the three raters' overall severities in their rating? In the measurement report of the rater severities, the three raters' relative severity logits were reported. The higher on the logit value the more severe on the scoring. The most severe rater was identified as rater 3 (0.22 logit), the second severe rater was rater 2, and the least severe rater was identified as rater 1 (-0.29 logit). It also shows a high reliability of 1.00 which indicates that this measurement reliably reports different degrees of rater severity, even though the difference between the logits of the most severe and the least severe raters was somewhat small (0.51 logit).

Research question 2: Are the three raters consistent in their rating? The fit statistics provided by FACETS analysis provide an important indication for examining the raters' judgments in terms of the consistency. By convention, any fit values greater than the range of two standard deviations around the mean are considered misfitting. As such, if the fit value of each rater falls outside of the normal range, which is two standard deviations around the mean $(M \pm 2S)$, then this indicates that the rater is inconsistent with much greater variation in ratings than expected. In this study, all fit values of the three raters were within the expected range. Therefore, no raters in this study were identified as misfitting, indicating that the three raters showed quite consistent rating performance.

Research question 3: How reliably do the three raters reveal different degrees of severity? FACETS analysis provides a reliability coefficient for each facet, which represents how reliably the analysis distinguishes among the elements. In other words, in the case of raters, the reliability indicates how reliably the raters were separated into different degrees of severity. According to the measurement report of the three raters, the reliability coefficient was 1.00. Therefore, it was found that the analysis quite reliably reveals the different degrees of severity among the three raters.

Research question 4: Are there any misfitting, or problematic, elements among the examinees, raters, task types, speech acts, and test items? In addition to identifying the misfitting raters, other facets such as, examinees, task types, speech acts, and test items were also examined to investigate whether each facet has misfitting elements. There were two examinees identified as slightly misfitting, which means these examinee test results were not consistent but showed much variation from the predicted pattern. Also, one test item from the LLDCT and one from the RP were also slightly misfitting because of the slightly high infit values. However, no task types and no speech acts were identified as misfitting.

Research question 5: Do any of the raters assess particular task types harsher or more leniently than others? If so, what are the raters' sub-patterns of assessing task types? Through conducting the bias analysis, the bias patterns between the three raters and the three task types were examined. The bias analysis provides *z*-scores, which represent the degree of differences between the expected scores and the observed scores. Also, the *z*-scores, either above +2.00 or below -2.00, indicate significance of bias. It was found that all raters showed different general patterns on scoring task types even though there were no significant *z*-scores. The following summary of bias patterns between the three raters and the three task types can be made in terms of their *z*-scores:

1. Rater 1 was harsher on the LLDCT and more lenient on the OPDCT.
2. Rater 2 was harsher on the RP and more lenient on the LLDCT.
3. Rater 3 was harsher on the OPDCT and more lenient on the RP.

Research question 6: Do the raters assess particular test items harsher or more leniently than others? If so, what are the raters' sub-patterns of assessing speech acts? To investigate the interaction between the three raters and the test items in detail, the relationship between rater severities and item difficulty logit was analyzed. To do so, the bias patterns between each rater and each test item within each of the three task types were identified. Contrary to the fact that there was no significant bias in the general interaction between the three raters and the three task types, numerous significant bias values, which are either *z*-scores above +2.00 or below -2.00, were found in the interaction between the rater severities and the item difficulties. Out of 54 interactions, three significant biases (6%) from the OPDCT items, five significant biases (9%) from the LLDCT items, and nine significant biases (17%) from the RP items were identified, excluding misfitting items among the significantly biased interactions. These significant biases more often appeared in the RP than in the OPDCT and LLDCT. In addition, unique bias patterns between the raters and the test items were also identified. The following summary can be made for the interaction between the raters and the test items:

1. Regarding the easiest item of the OPDCT and RP, rater 1 and rater 3 showed the opposite bias pattern. Rater 1 was more lenient on the easiest item than expected; however, rater 3 was harsher on the same item than expected.
2. Regarding the most difficult item of the LLDCT and RP, rater 1 and rater 3 again showed the opposite bias pattern. Rater 1 was harsher on the most difficult item

than expected; however rater 2 was more lenient on the same item than expected.

Research question 7: Do the raters assess particular speech acts harsher or more leniently than others? In the interaction between the rater severities and the speech acts, the three raters showed significant bias patterns on rating the speech acts. Out of nine total interactions, five significant biases (56%) were identified, excluding one misfitting interaction. No raters showed the same degree of harshness on the same speech act. Especially, compared with the fact that the raters showed no significant bias on the three task types, it is notable that all three raters showed significant bias on the speech acts. Different speech acts, refusal, apology, and request, appear to be more influential on the rater bias than different task types. The following summary can be made based on the interaction between the three raters and the three speech acts:

1. In refusal, rater 2 and rater 3 showed the opposite bias pattern. Rater 2 was more lenient on refusal than expected, whereas rater 3 was harsher on refusal than expected.
2. In apology, only rater 1 showed the significant bias, harsher on apology than expected.
3. In request, rater 2 and rater 3 again showed the opposite bias pattern. This time, rater 2 was harsher on request than expected, whereas rater 3 was more lenient on request than expected.

Research question 8: Do the raters assess particular abilities of examinees harsher or more leniently than others? Lastly, the interaction between the rater severities and the examinee abilities was investigated, and numerous significant bias patterns were found from all three raters. Out of 24 interactions, eight significant biases from rater 1 (33%), three significant biases from rater 2 (13%), and two significant biases from rater 3 (8%) were identified. There significant biases more often appeared in the interaction between rater 1 and examinee abilities than in the other two rater interactions. Also, the significant bias patterns especially with the most able examinees were found. Especially, rater 2 and rater 3 showed the opposite bias pattern. Rater 2 was harsher than expected on the most able examinee; however, rater 3 was more lenient than expected on the same examinee. The following summary can be made based on the interaction between the three raters and the examinee abilities.

1. Rater 1 was harsher than expected on the least able examinee, but more lenient than

expected on the most able examinee.

2. Rater 2 was more lenient than expected on the most able examinee.
3. Rater 3 was harsher than expected on the most able examinee.

CONCLUSION

This study investigated whether various factors, including examinees, raters, task types, speech acts, and test items, influence the rater patterns are systematic sources of bias in the assessment of pragmatic competence. Although there was a limited range in examinee's language levels, interesting results were found. Overall, the three raters, who had a similar educational and professional background, showed different degrees of severities on their rating. Also, they were quite consistent, and they showed reliably different degrees of severity in their scoring, even though they had a minimal rater training. However, even though the three raters showed consistency in their rating, when it comes to the interactions between the raters and the other facets, which are examinees, task types, speech acts, and items, very interesting and distinctive bias patterns within each interaction were found. The three raters showed more variation on scoring different speech acts than on scoring different task types. Rater 1 and rater 3 showed the opposite bias patterns across the test items; rater 2 and rater 3 showed the opposite bias pattern across the speech acts and test items.

Limitations of This Study

One of the main limitations of this study is the limited range of the participants as examinees. Due to the time-consuming administration of the three task types, open-ended DCT, language lab DCT, and role play, only twenty-four examinees were able to finish all three task types during the limited time period. Also, most of the examinees in this study ranged from intermediate to advanced level, and it was found that all of these participants had 50 percent chance of succeeding on items of average difficulty, based on FACETS analysis. Even though beginning KFL learners might have difficulty in accomplishing all the test items, it would have been better to have a wide range of examinees' levels.

The second limitation is the relatively small numbers of task types and speech acts. For the sake of this study's scope, this study only examined three measurements, open-ended written DCT, language lab DCT, and role play, out of the six measures that Hudson et al. (1995) developed for assessing pragmatic competence and this study chose three speech acts: refusal, request, and apology. However, there are various measurements that can be employed, and there are more speech acts that can be investigated.

Implications of This Study

The findings of this study should encourage further investigation of not only assessing pragmatic competence in the KFL teaching context, but also of ILP studies on KFL learners in general. Despite significant developments in KFL education in US college settings during the last three decades, studies on pragmatics in KFL teaching still need much more attention. There has been attention paid to investigating the sociopragmatic and pragmalinguistic features of speech acts, and assessment of pragmatic competence of KFL learners, such as Byon (2002, 2004b, 2005) and Ahn (2005), though there are still few of studies and, to date there are no longitudinal studies on speech act development of KFL learners. Furthermore, considering that previous ILP studies of KFL learners mainly employed open-ended written DCTs as an analytic resource, using various analytic resources for investigating authentic productions of learners, such as conversation analysis (CA) should be encouraged. Especially, since using appropriate Korean honorifics is crucial for engaging in various speech acts in Korean (Byon, 2004a), micro analyses of Korean honorifics usages of KFL learners, drawing on CA, can provide essential insights into the teaching pragmatics for KFL learners.

This study used multi-faceted Rasch measurement, which has great potential that classical testing theories cannot provide, including the comparison of various factor characteristics on the same logit scale and bias analysis that identify raters’ particular patterns of harshness or leniency with relation to certain test items or certain examinees. Therefore, multi-faceted Rasch measurement should be further employed in various Korean L2 performance-based language assessment settings, including KFL learner’s writing and oral discussion, to investigate raters’ judgments. Especially, as unreliability of rater judgments has been pointed to as one of the essential issues in performance-based assessment (Brown, Hudson, Norris, & Bonk, 2002, McNamara, 1996; Norris, Brown, Hudson, & Yoshioka, 1998; Shohamy, 1995), multi-faceted Rasch measurement has great potential for dealing with this issue. For example, Weigle (1998) used multi-faceted Rasch measurement to investigate rater training's effectiveness, and found that rater training was crucial for making raters more consistent and for lowering the level of rater variability in their overall severity.

Also, this study's findings reinforce the importance of having numerous raters in performance-based assessment. McNamara (1996) mentioned that raters may demonstrate particular patterns of harshness or leniency regarding certain candidates or task types, that each rater may interpret the rating scale differently, and that they may vary in rating consistency (pp. 123-124). In this study, there were significant bias patterns among the three raters with the same examinees' test results depending on various factors even though the three raters showed overall consistency in their scoring. These findings were consistent with what McNamara (1996) noted. Therefore, I recommend further investigations of rater bias on performance-based assessment. In addition to investigating rater scoring judgments quantitatively, future qualitative research should further examine how different variables within each speech act, such as power, distance, and imposition, influence the examinees' pragmatic competence and the raters' decisions.

Regarding this study's pedagogical implications for KFL teaching, several important issues should be addressed. Firstly, heritage language learners showed an imbalance between written competence and spoken competence in this study. This phenomenon is also reported in Ahn's (2005) study. Therefore, this issue should be further investigated and considered in KFL pragmatic teaching. Secondly, incorporating the six measures that Hudson et al. (1992, 1995) developed into the existing curriculum of KFL classrooms would encourage teachers to improve KFL learners' pragmatic competence. Above all, suitable curriculum and materials for teaching pragmatics for KFL learners must be developed, based on KFL learner needs analysis or other relevant research findings.

Pragmatic competence figures significantly in the communicative competence model (Bachman, 1990). However, despite this, there has been less attention paid to assessing pragmatic competence than to assessing the other aspects of language manifested in communicative competence theory in second and foreign language teaching contexts. The current study attempted to fill this gap, and, hopefully, this study will encourage other researchers to further investigate the various important issues involved in teaching pragmatics in KFL.

ACKNOWLEDGEMENTS

It has been such a privilege to be helped and supported by leading scholars throughout my program of study. I would like to express sincere gratitude for my thesis committee members. First of all, I am very grateful to my advisor and committee chair Professor James Dean Brown for his critical feedback, guidance, and encouragement. I would also like to extend my sincere gratitude to my thesis committee, Professor Thom Hudson and Professor Gabriele Kasper, for their sage advice, valuable feedback, and ongoing support.

I would also like to express many thanks to all participants in my study, including the test takers who sincerely and patiently completed all three tests, and the three raters who patiently and thoughtfully helped throughout the rating process. A special thank you also to the Korean teachers at the University of Hawaiʻi and Hanoi University of Foreign Studies in Vietnam who gave permission to visit their classes and conduct this research.

REFERENCES

Achiba, M. (2002). *Learning to request in a second language: Child interlanguage pragmatics*. Clevedon: Multilingual Matters.

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

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*, 449-465.

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*, 89-110.

Bouton, L. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes*, *17*, 183-196.

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

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

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*, 653-675.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University.

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

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

Byon, A. S. (2002). Pragmalinguistic features of KFL learners in the speech act of request.

Korean linguistics, *11*, 151-182.

Byon, A. S. (2004a). Learning linguistic politeness. *Applied Language Learning, 14*, 37-62. Retrieved January 7, 2007, from http://www.dliflc.edu/academics/academic_materials/all/ALLissues/all14/14onecomplete.pdf

Byon, A. S. (2004b). Sociopragmatic analysis of Korean request: Pedagogical settings. *Journal of Pragmatics*, *36*, 1673-1704.

Byon, A. S. (2005). Teaching refusals in Korean. *The Korean Language in America*, *10*, 1-18.

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

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 1-27). New York: Longman.

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

Crystal, D. (1997). *A dictionary of linguistics and phonetics*. Oxford: Basil Blackwell.

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

Elder, C., Barkhuizen, G., Knock, U., & Randow, J. (2007). Evaluating rater response to an online training program for L2 writing assessment. *Language Testing*, *24*, 37-64.

Ellis, R. (1992). Learning to communicate in the classroom: A study of two learners' requests. *Studies in Second Language Acquisition*, *14*, 1-23.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer·Nijhoff Publishing.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing*, *13*, 53-61.

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

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

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

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: selected readings* (pp. 269-293). Harmondsworth: Penguin.

Kasper, G. (1996). Introduction: Interlanguage pragmatics in SLA. *Studies in Second Language Acquisition*, *18*, 145-148.

Kasper, G., & Blum-Kulka, S. (1993). Interlanguage pragmatics: An introduction. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 3-17). New York: Oxford University.

Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, *13*, 215-247.

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

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

Kim-Park, J. (1995). *Linguistic variation and territorial functioning: A study of the Korean honorific system*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, Pennsylvania.

Koh, H. E. (2002). *A cross-cultural study of address terms in Korean and English*. Unpublished doctoral dissertation, University of Hawai‘i, Honolulu, Hawai‘i.

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

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*, 1-27.

Lee, C. (1996). *Variation in use of Korean honorific verbal endings: An interactional sociolinguistic study*. Unpublished doctoral dissertation, Boston University, Boston, MA.

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

Levinson, S. C.(1983). *Pragmatics.* Cambridge: Cambridge University.

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

Linacre, J. M. (1996). *Facets, version no. 3.0.* Chicago: MESA.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-faceted Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*, 158-180.

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

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report #18). Honolulu: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Roever, C. (2001). *A web-based test of interlanguage pragmalinguistic knowledge: Speech acts, routines, and implicatures*. Unpublished doctoral dissertation, University of Hawai‘i, Honolulu, Hawai‘i.

Rose, K. R., & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching.* Cambridge: Cambridge University.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

Sohn, H. (1999). *The Korean language*. Cambridge: Cambridge University.

Strauss, S., & Eun, J. O. (2005). Indexicality and honorific speech level choice in Korean, *Linguistics*, *43*, 611-651.

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

University of of Hawaii‘i at Mānoa (2006). The 2006-2007 general and graduate information catalog of UH. Retrieved October, 8, 2006, from http://www.catalog.hawaii.edu/courses/departments/kor.htm

University of Hawaii‘i at Mānoa Korean flagship program (2006). Retrieved October, 8, 2006 from http://www2.hawaii.edu/~flagship/index.html

Valdes, G. (2001). Heritage language students: Profiles and possibilities. In J. K. Peyton, D. A. Ranard, & S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 37-77). McHenry: The Center for Applied Linguistics and Delta Systems.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*, 411-440.

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

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

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

APPENDIX A

Korean version Open-ended DCT (Based on Hudson et al. 1995)

TASK1: Open-ended Discourse Completion Task

Directions:
Read each of the situations. After reading each situation, write (**in KOREAN**) what you would say in each situation in a normal conversation. You don't have to spend too much time to think about each situation. Feel free to express your natural response on each situation. If you are not sure about each situation in Korean, please refer to an English supplement.

상황1. (+P, -D, -I)
당신은 큰 집을 가지고 있습니다. 그래서 남는 방은 다른 사람에게 세를 주고 있습니다. 당신은 집세를 받기 위해 다른 사람의 방에 왔습니다. 집세를 건네 받으려고 하는 순간 당신은 실수로 책상 위의 빈 꽃병을 넘어뜨렸습니다. 꽃병은 깨지지 않았습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황2. (-P, +D, +I)
당신은 보석 수리를 하는 작은 가게에서 일하고 있습니다. 고객께서 가게에 물건을 찾으러 오셨습니다. 이 물건은 선물로 할 골동품의 가치를 지닌 시계입니다. 사실 당신은 오늘까지 수리를 마친다고 약속했지만 아직 수리를 끝내지 못했습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황3. (-P, +D, +I)
당신은 어느 작은 회사에 취직하고 싶어합니다. 그런데 인사부장님이 너무 바빠서 오후 1시부터 4시까지밖에 면접을 볼 수가 없습니다. 하지만 당신은 지금 오후에 다른 일을 하고 있기 때문에 오전에 면접을 보고 싶어합니다. 당신은 아침에 지원 서류를 내기 위해 회사에 방문을 했을 때 인사부장님을 만났습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황4. (+P, -D, -I)
당신은 전국 스키 클럽의 회원입니다. 매달 당신의 스키 클럽에서는 스키 여행을 떠납니다. 이번 달 스키 여행 계획을 위해 당신은 지금 모임에 와 있습니다. 클럽 회장이 당신 옆에 앉아 있는데 당신에게 펜을 빌려달라고 부탁합니다. 하지만 당신에게는 펜이 하나밖에 없고 이 펜으로 메모를 해야 하므로 빌려 줄 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황5. (-P, -D, -I)
당신은 대기업의 작은 부서에서 일하고 있습니다. 당신은 지금 부서 회의에 참석하고 있는데 메모를 하기 위해 펜을 빌려야 합니다. 당신 옆에 앉아 계신 부장님한테 남는 펜이 있을지도 모릅니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황6. (+P, +D, +I)
당신은 어떤 회사의 부장입니다. 지금 당신은 사원을 뽑기 위해 면접을 하고 있습니다. 면접 서류를 올려놓은 책상으로 가던 중 당신은 실수로 면접 볼 사람이 들고 온 종이 가방을

밟아버렸습니다. 이때 당신은 종이 가방 안의 무엇이 부서지는 소리를 들었습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황7. (+P, +D, -I)
당신은 친구에게 줄 생일 선물을 사기 위해 상점에서 쇼핑을 하던 중 상자에 담긴 물건을 좀더 자세히 보고 싶어합니다. 이때 점원이 다가옵니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황8. (+P, -D, +I)
당신은 큰 집을 가지고 있습니다. 그래서 남는 방은 다른 사람들에게 세를 주고 있습니다. 당신 집에 살고 있는 각 방의 사람들은 매주 집안일을 나누어서 하고 있습니다. 그런데 이 집의 한 사람이 여행을 가기 때문에 당신은 그 사람의 이번 주 집안일을 대신 부탁 받았습니다. 하지만 당신은 이번 주에 바빠서 비는 시간이 전혀 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황9. (-P, +D, -I)
당신은 작은 가게에서 일하고 있습니다. 당신이 뒤쪽 방에서 중요한 일 때문에 전화를 받던 중 손님이 가게로 들어오셨음을 알리는 벨 소리를 들었습니다. 당신은 전화를 급히 끊고 기다리고 계시는 손님에게로 갔습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황10. (-P, +D, -I)
당신은 어느 회사에 취직하고 싶어합니다. 응시 원서를 받기 위해 사무실에 들렀고 의자에 앉아계신 인사부장님을 만났습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황11 (+P, -D, -I)
당신은 전국 독서 클럽의 지부 회장입니다. 이 클럽에서는 매달 새 책을 읽고 이에 관해 토론을 하는 모임을 가집니다. 당신은 지금 이 모임에 와있고 옆에 앉은 다른 회원과 이야기를 하고 있습니다. 이때 당신은 클럽 회원인 이수진씨의 전화번호가 필요하고 옆에 앉은 회원이 마침 이수진씨의 전화번호를 알고 있는 것 같습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황 12 (+P, -D, +I)
당신은 큰 규모를 가진 학교의 선생님입니다. 학교에서 주임선생님과 마주쳤습니다. 주임 선생님께서 부탁하시기를, 다른 모든 선생님들에게 오늘 밤 전화를 걸어 내일 회의가 있다는 사실을 전하라고 합니다. 하지만 이 일은 너무 시간이 걸리고 게다가 오늘 밤에 당신 친구가 집에 놀러 오기 때문에 당신은 이 일을 할 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황 13 (+P, +D, -I)
당신은 지금 여행자 수표를 사기 위해 작은 은행에 와있습니다. 이때 당신은 수표를 은행 직원으로부터 건네 받다가 실수로 책상 위에 있는 작은 장식품을 넘어뜨렸습니다. 이것은 깨지지 않았습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황14 (-P, -D, -I)
당신은 서점에서 일하고 있습니다. 오늘 정오부터 일하기로 일정이 잡혀있습니다. 당신의 상사는 오전 근무를 하고 당신은 그 후에 교대로 일합니다. 그런데 당신은 오늘 12시가 넘어 가게에 도착했습니다. 당신의 상사를 만났습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황15 (+P, -D, -I)
당신과 동료 몇 명은 특별한 프로젝트를 함께하고 있고 당신은 이 프로젝트의 리더입니다. 당신은 복도에서 프로젝트를 함께하고 있는 동료 직원을 만났습니다. 그런데 이 동료 직원이 말하기를 오늘 오후 모임에 이수진씨를 만나면 그녀에게 메시지를 전해달라는 부탁을 당신에게 합니다. 하지만 당신은 이수진씨가 모임을 취소해서 그녀를 못 만나기 때문에 그녀에게 메시지를 전달할 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황16 (+P, +D, -I)
당신은 백화점 안을 걷고 있습니다. 당신이 마침 진열장 옆을 지나가는데 직원이 당신에게 이번에 새롭게 출시된 상품의 선전 비디오를 잠깐 동안 보지 않겠냐고 부탁합니다. 하지만 당신은 지금 친구와의 점심 약속에 가는 길이기 때문에 비디오를 볼 시간이 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황17 (-P, -D, +I)
당신은 큰 집의 방을 빌려서 살고 있습니다. 이 큰 집의 집주인 역시 함께 이 집에 살고 있습니다. 당신은 일주일에 2시간씩 잔디를 깎는 일을 맡고 있습니다. 그런데 이번 주에 당신은 여행을 가기 때문에 집주인이 당신 대신에 잔디를 깎아 주길 원하고 있습니다. 집주인을 만났습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황18 (-P, +D, -I)
당신은 백화점에서 점원으로 일하고 있습니다. 손님께서는 물건을 사고 돈을 지불하셨고 당신은 3000원을 잔돈으로 손님께 돌려 드려야 합니다. 이 손님은 3000원을 1000원짜리 말고 500원짜리 잔돈으로 받길 원합니다. 하지만 당신은 여분의 500원짜리 동전이 없기 때문에 손님의 부탁을 들어줄 수 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

English Supplement (Open DCT)

Situation1: You live in a large house. You hold the lease to the house and rent out the other rooms. You are in the room of one of your housemates collecting the rent. You reach to take the rent check when you accidentally knock over a small, empty vase on the desk. It doesn't break.

Situation2: You work in a small shop that repairs jewelry. A valued customer comes into the shop to pick up an antique watch that you know is to be a present. It is not ready yet, even though you promised it would be.

Situation3: You are applying for a new job in a small company and want to make an appointment for an interview. You know the manager is very busy and only schedule interviews in the afternoon from one to four o'clock. However, you currently work in the afternoon. You want to schedule an interview in the morning. You go into the office this morning to turn in your application form when you see the manager.

Situation4: You are a member of the local chapter of a national ski club. Every month the club goes on a ski trip. You are in a club meeting now helping to plan this month's trip. The club president is sitting next to you and asks to borrow a pen. You cannot lend your pen because you only have one and need it to take notes yourself.

Situation5: You work in a small department of a large office. You are in a department meeting now. You need to borrow a pen in order to take some notes. The head of your department is sitting next to you and might have an extra pen.

Situation6: You are an office manager and are interviewing to fill a position that is open. You are interviewing someone now. You walk over to the filing cabinet to get the applicant's application when you accidentally step on a small shopping bag belonging to the applicant. You hear a distinct crunching. You are certain you have broken whatever is in the small bag.

Situation7: You are shopping for your friend's birthday and see something in a display case. You want to look at it more closely. A salesclerk comes over to you.

Situation8: You live in a large house. You hold the lease to the house and rent out the other rooms. Each person in the house is responsible for a few hours of chores every week. One of your housemates asks if you can do extra chores this week because your housemate is going out of town. You cannot do your housemate's chores this week because you are very busy at work this week and do not have any extra time.

Situation9: You work in a small shop. You are working in the back room when you hear the bell that tells you there is a customer in the front room. You are on the phone making an important business call. You finish the call as quickly as you can and go out to help the waiting customer.

Situation10: You want to apply for a job in a small office. You want to get an application form. You go to the office and see the office manager sitting behind a desk.

Situation11: You are the president of the local chapter of a national book club. The club reads and discusses a new book every month. You are at this month's meeting, talking with a member of the book club. You need to get the phone number of Sujin Lee, another member of the club. You think this person has Sujin’s number.

Situation12: You are a teacher at a large school. You see the lead teacher on campus. The lead teacher asks you to call all of the other teachers tonight and tell them that there will be a meeting tomorrow. You cannot do it because you know that it will take hours and you have friends coming over to your house tonight.

Situation13: You are in a small bank buying traveler’s checks. You move to take the checks when you accidentally knock over a small ceramic figure on the clerk’s desk. It doesn’t break.

Situation14: You work in a bookstore. You are scheduled to start work at noon today. You will take over for your supervisor who is working the morning shift. You go to work and arrive at the bookstore a few minutes after noon. You see your supervisor.

Situation15: You and a few of your co-workers are working on a special project. You have been appointed the project leader. You are walking in the hallway when another co-worker also working on the project asks you to give a message to Sujin when you see her at a meeting you and Sujin have scheduled this afternoon. You cannot deliver the message because you will not be seeing her. Sujin has canceled the meeting.

Situation16: You are walking through a department store. As you walk past a display, a salesclerk asks you to watch a short video demonstration for a new product. You cannot stop because you are on your way to meet someone for lunch.

Situation17: You rent a room in a large house. The person who holds the lease lives in the house as well. You are responsible for mowing the lawn every week, a job that takes you about two hours to do. You want the lease-holder to mow the lawn for you this week because you are going out of town. You are in the living room when the lease-holder walks in.

Situation18: You work as a sales clerk in a department store. A customer is paying an item and should get 3000 won back in change. The customer asks that the 3000 won be given in 500 won, not 1000 won bills. You cannot give the change because you do not have enough quarters to spare.

APPENDIX B
Korean version Language Lab DCT (Based on Hudson et al. 1995)

상황1. (+P, -D, -I)
당신은 큰 회사의 작은 부서에서 일하고 있습니다. 이 회사에서 당신은 수년간 일했고 지금은 작은 부서의 부장입니다. 당신은 지금 동료의 사무실에서 회의를 하고 있습니다. 그런데 당신은 실수로 책상 위에 있는 액자를 넘어뜨리고 맙니다. 액자는 깨어지지 않았습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황2. (-P, +D, +I)
당신은 지금 회사에 취직하고 싶어합니다. 당신은 지원서를 내기 위해 새로운 회사의 사무실에 들렀고 당신은 인사부장님을 만나서 잠깐 동안 이야기를 하고 있습니다. 당신이 인사부장님께 지원서를 건네드리려는 순간 실수로 책상 위의 꽃병을 넘어뜨려서 꽃병 안의 물을 종이 위에 쏟아버렸습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황3. (-P, +D, +I)
당신은 작은 은행에서 등록금 대출을 신청하려고 합니다. 당신은 지금 대출 담당 직원과 만나고 있습니다. 이 대출 담당 직원 한 명이 이 은행의 모든 대출 관련 서류를 담당하고 있습니다. 이 직원이 말하기를 지금 많은 서류들이 밀려있어서 당신의 신청 서류를 처리하려면 2주는 걸릴 것이라고 합니다. 하지만 당신은 등록금을 마감 기간까지 내기 위해 빨리 서류가 검토되었으면 합니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황4. (-P, -D, -I)
당신은 전국 스키 동호회의 회원입니다. 매달 당신의 동호회에서는 스키 여행을 떠납니다. 당신은 지금 다음달의 여행을 결정하기 위해 동호회 회장과 이야기를 하고 있습니다. 당신은 회장에게 메모를 하기 위한 종이를 빌리고 싶어합니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황5. (-P, +D, +I)
당신은 큰 집의 방하나에 세 들어 살고 있습니다. 집주인도 함께 이 집에 살고 있습니다. 이 집의 각 사람들은 집안일을 각자 나누어 맡아 하고 있습니다. 당신은 청소기를 돌리는 일을 맡고 있습니다. 오늘 아침 당신은 집주인의 청소기를 사용하던 중 실수로 이것을 떨어뜨리고 말았습니다. 그리고 청소기는 더 이상 작동을 하지 않습니다. 이때 집주인이 당신 곁으로 다가옵니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황6. (+P, -D, -I)
당신은 지금 비행기를 타고 여행을 하고 있습니다. 그리고 지금 당신은 저녁 식사를 할 것입니다. 승무원이 당신 식사를 준비합니다. 그리고 당신은 냅킨이 필요합니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황7. (+P, +D, +I)
지난 주 당신은 회사의 차에 문제가 생겨 정비공에게 맡겼었습니다. 이 정비공은 내일 아침까지 수리를 마치겠다고 약속했습니다. 당신은 내일 오후에 출장을 가고 이 때문에 차가 필요합니다.

당신은 정비공에게 다시 가서 내일 아침까지 수리를 마쳐달라는 부탁을 합니다. 하지만 정비공은 지금 너무 할 일이 많아서 수리를 마치는데 하루가 더 걸리 것 같다고 말합니다. 하지만 당신은 출장을 늦출 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황8. (-P, -D, -I)
당신은 해외 출장을 마치고 공항의 세관을 통과하려고 합니다. 당신의 차례가 왔고 세관 직원이 당신에게 서류를 보여달라고 말합니다. 하지만 당신은 이 서류를 어디에 넣어두었는지 몰라서 한참을 가방 안을 뒤적이고 난 후 서류를 찾았고 이 서류를 기다리고 있는 직원에게 줍니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황9. (-P, +D, -I)
당신은 레스토랑에서 일하고 있습니다. 당신은 방금 손님의 주문을 받았습니다. 당신은 다른 테이블에 있는 손님을 위해 메뉴판이 아직 필요합니다. 하지만 손님께서 아직 메뉴판을 손에 쥐고 계십니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황10. (+P, -D, +I)
당신은 전국 캠핑 동호회의 회장입니다. 매달 동호회에서는 캠핑 여행을 떠나고 당신은 이 여행의 계획을 맡고 있습니다. 지난 주 당신은 이번 달의 여행 계획을 짜기 위해 다른 회원과 만나기로 약속했었습니다. 하지만 요즘 당신은 조금 바빠 회원과의 약속 시간을 오늘 아침 7:30으로 바꿨습니다. 그런데 오늘 아침 당신은 심한 교통 마비 때문에 약속 장소에 아침 9시에 도착했습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황11. (+P, -D, -I)
당신은 큰 집을 소유하고 있고 남는 방은 다른 사람에게 세를 주고 있습니다. 당신의 세탁기가 고장이나 오늘은 토요일이고 오늘 아침에 고치는 사람이 오기로 되어있습니다. 하지만 당신은 공항에 당신 부모님을 모시러 가야 하기 때문에 곧 집을 떠나야 합니다. 당신은 지금 부엌에 있고 옆방 사람이 다가옵니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황12. (-P, +D, +I)
당신은 작은 복사 가게에서 일하고 있습니다. 어느 늦은 오후에 소중한 고객 한 분이 내일 아침까지 새로운 광고 전단지 1500부를 부탁합니다. 이 일을 하기 위해 당신은 밤 늦게까지 일해야 합니다. 하지만 당신은 오늘 많이 피곤해서 밤늦게까지 일할 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황13. (-P, -D, +I)
당신은 전국 캠핑 동호회의 회원입니다. 매달 동호회에서는 캠핑 여행을 떠나고 회장이 매달 몇 시간에 걸쳐 여행 계획을 짭니다. 당신은 이번 달 여행에서 회장과 이야기를 나누고 있습니다. 회장이 몇 주 동안 해외에 가있기 때문에 당신에게 다음 달의 캠핑 계획을 짜달라고 부탁합니다. 하지만 당신은 직장 일로 바빠서 여행 계획을 짤 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황14. (-P, -D, -I)
당신은 작은 학교에서 학생을 가르치고 있습니다. 당신은 오늘 오후2시에 주임선생님과의

약속이 있습니다. 그런데 당신은 주임선생님과의 약속에 몇 분 정도 늦었습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황15. (+P, -D, -I)
당신은 큰집을 가지고 있고 남는 다른 방은 다른 사람에게 세를 주고 있습니다. 그런데 당신 옆방 사람이 당신에게 와서 중요한 이야기는 아니지만 잠깐 동안 이야기를 나눌 수 있냐고 물어봅니다. 하지만 당신은 지금 밖으로 나가는 길이라 이야기를 나눌 수 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황16. (+P, +D, -I)
당신은 지금 점심 시간에 잠시 회사밖에 나와있습니다. 이때 잠시 작은 상점에 들러 친구의 생일 선물을 둘러봅니다. 마침 마음에 드는 물건을 찾았고 당신은 이 물건을 샀습니다. 당신이 가게를 나가려고 할 때 점원이 당신에게 펜을 잠시 빌려달라고 합니다. 하지만 당신은 빨리 회사로 돌아가봐야 하기 때문에 펜을 빌려 줄 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황17. (+P, +D, +I)
당신은 어느 회사의 인사부장이고 지금 새로운 직원을 뽑고 있습니다. 지원자들이 지원서의 내용이 꽤 많아서 작성하는데 시간이 많이 걸립니다. 그리고 지원서는 타이프로 쳐서 작성을 해야 합니다. 한 지원자가 작성한 지원 서류를 내고 갑니다. 하지만 이 지원서는 희미한 잉크 때문에 잘 보이지 않기 때문에 다시 작성 되어야 할 것 같습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

상황18. (-P, +D, -I)
당신은 작은 상점에서 일하고 있습니다. 한 손님이 가게로 들어와 만 원짜리를 천 원짜리로 바꾸어달라고 부탁합니다. 하지만 지금 천 원짜리가 없기 때문에 당신은 바꾸어 줄 수가 없습니다.

당신은 이 상황에서 어떻게 말하겠습니까?

English Supplement (Language lab DCT)

Situation 1
You work in a small department of a large office. You have worked here for a number of years and are the head of the department. You are in the office of another member of the department in a meeting. You accidentally knock over a framed picture on the desk. It doesn’t break.

Situation 2
You are applying for a job in a company. You go into the office to turn in your application form to the manager. You talk to the manager for a few minutes. When you move to give the manager your form, you accidentally knock over a vase on the desk and spill water over a pile of papers.

Situation 3
You are applying for a student loan at a small bank. You are now meeting with the loan officer. The loan officer is the only person who reviews the applications at this bank. The loan officer tells you that there are many other applications and that it should take two weeks to review your application. However, you want the loan to be processed as soon as possible in order to pay your tuition by the

deadline.

Situation 4
You are a member of the local chapter of a national ski club. Every month the club goes on a ski trip. You are in a meeting with the club president, helping plan this month's. You want to borrow some paper in order to take some notes.

Situation 5
You rent a room in a large house. The person who holds the lease lives in the house as well. Each person in the house is responsible for a few hours of chores every week. Your chore is to vacuum the house. This morning when you were using the leaseholder's vacuum you accidentally dropped it and now it does not work. You are now in the living room and the leaseholder walks in.

Situation 6
You are on an airplane. It is dinner time. The flight attendant sets your food on your tray. You need a napkin.

Situation 7
Last week you had trouble with your company car and took it to a company mechanic. The mechanic promised to have it ready tomorrow morning. You are going on a business trip tomorrow afternoon and need the car. You stop by the repair shop to make sure the repairs will be finished in time. Now the mechanic tells you the shop is very busy and asks if you can wait an extra day for your car. You cannot delay your trip.

Situation 8
You are in the airport going through customs after a trip to a foreign country. It is your turn, but when the customs officer asks you for your papers, you realize you do not know where they are. You look in your bag for a little while, find them, and give them to the waiting officer.

Situation 9
You work in a restaurant. You have just taken a customer's order and are ready to leave the table. The customer is still holding the menu and you need it for another table.

Situation 10
You are the president of the local chapter of a national camping club. Every month the club goes on a camping trip and you are responsible for organizing it. Last week you were supposed to meet with another member of the club to plan this month's trip. You had to reschedule because you were too busy. The rescheduled meeting was for 7:30 this morning, but you got caught in heavy traffic and just now arrive at the club headquarters. It is 9:00 a.m.

Situation 11
You live in a large house. You hold the lease to the house and rent out the other rooms. The washing machine is broken. It is Saturday and the repair person is scheduled to fix it this morning. However, you will not be home because you have to pick up your parents at the airport. You are in the kitchen when a house-mate walks in.

Situation 12
You work in a small printing shop. It is late afternoon an a valued customer comes into ask if you

can print 1500 copies of a new advertisement by tomorrow morning. To do this you would have to work into the night. You are tired after a long day and cannot stay late.

Situation 13
You are a member of the local chapter of a national camping club. Every month the club goes on a camping trip. The president of the club is responsible for organizing the trips, a job that takes a number of hours. You are on this month's trip talking to the president of the club. The president is going to be out of town for a week and asks you to plan the next trip. You cannot plan the trip because you are going to be very busy with work.

Situation 14
You teach in a small school. You have a meeting with the lead teacher at two o'clock today. When you show up the meeting it is a few minutes after two.

Situation 15
You live in a large house. You hold the lease to the house and rent out the other rooms. You are in the living room when one of your housemates asks to talk to you. Your housemate explains that it will only take a few minutes and is not important. However, you cannot talk now because you are on your way out.

Situation 16
You are on your lunch hour. You go into a small shop to look for a present for you friend's birthday. You find something you like and but it. As you are ready to leave the clerk asks to borrow your pen. You cannot lend your pen because you have to hurry back to work.

Situation 17
You are the personnel officer in an office that is now hiring new employees. The application form is quite long and takes most applicants several hours to complete. The form must be typed. An applicant comes in and gives you a completed form. However, it has been typed with a very faint ribbon. The application needs to be retyped.

Situation 18
You work in a small store. A customer comes into the store and asks you change for a ten dollar bill. You cannot give the change because you don't have it in the register.

APPENDIX C
Korean version Role play (Based on Hudson et al. 1995)

Role play #1 자동차 수리점에서

당신: 자동차 수리를 부탁한 손님
어디서: 자동차 수리점
누구와: 자동차 정비공
장면: 당신은 고장 난 자동차를 저번 주에 자동차 수리점에 맡겼습니다. 당신은 오늘 수리점에 차를 가지러 왔고 정비공을 만났습니다. (이때 정비공은 마침 점심 식사를 하고 있습니다.)

1. 사실 자동차는 모레까지 수리가 마쳐질 예정입니다. 하지만 당신은 내일 아침까지 수리가 마쳐달라는 부탁을 정비공에게 하러 왔습니다. (+P, +D, +I)
2. 정비공은 당신에게 커피를 권하지만 당신은 마시고 싶어하지 않습니다. (+P, +D, -I)
3. 정비공이 다른 정비공에게 물어보러 간 사이 당신은 실수로 이 정비공이 마시고 있던 커피를 책상 위의 서류에 쏟고 말았습니다. (+P, +D, +I)

#1 At the car garage

You go to a mechanic, who is having lunch now, to pick up company van, which had some trouble last week.

1. You want to ask the mechanic to have the van ready by early tomorrow morning, which is one day earlier than it is supposed to be ready.
2. The mechanic offers you some coffee but you don't want any.
3. After the mechanic leaves to talk with another mechanic about the van, you accidentally knock over his coffee and it spills over some papers on the desk.

Role play #2 선물 가게에서

당신: 손님.
어디서: 선물 가게
누구와: 가게 점원과
장면: 당신은 다음주까지가 유효 기간인 상품권을 가지고 선물 가게로 들어왔습니다. 당신은 이 상품권으로 이번 주 토요일이 생일인 친구의 생일 선물을 사려고 합니다.

1. 당신은 예쁜 꽃병이 들어있는 상자 안을 들여다보고 있고, 이 꽃병을 자세히 들여다 보고 싶어 점원에게 부탁합니다. (+P, +D, -I)
2. 돈을 지불하려고 당신은 상품권을 지갑에서 꺼냅니다. 이때 당신은 이 상품권이 많이 구겨졌다는 것을 알아챕니다.(+P, +D, -I)
3. 점원은 이번 주 토요일에 열리는 10주년 기념 세일 행사에 당신을 초대합니다만 당신은 갈 생각이 없습니다. (+P, +D, -I)

# 2 At the gift shop

You walk into a gift shop to use a gift certificate before it expires next week. You want to buy a birthday present for your friend with this certificate which you will give to your friend at the birthday party this Saturday. You see a nice vase in a case and want to get a closer look at it.

1. You ask the salesperson walking by to take it out for you.
2. Now when you take out the gift certificate and hand it to the salesperson, you noticed that the gift certificate is very dirty.
3. You are invited to the “10th anniversary sale” on Saturday, but you will not go.

Role play #3 집에서

당신: 큰 집을 소유하고 있고 남는 방은 다른 사람에게 세를 주고 있습니다.
언제: 저녁
어디서: 집에서
누구와: 당신 집에 살고 있는 세입자
장면: 당신은 당신의 세입자와 저녁쯤에 만나기로 했습니다. 당신은 외출했다가 지금 막 집으로 들어왔습니다.

1. 당신은 세입자와 만나기로 한 약속 시간에 조금 늦었습니다.(+P, -D, -I)
2. 세입자는 이번 주 토요일에 파티를 열고 싶다고 말합니다. 하지만 이번 주 토요일에는 페인트 칠 약속이 이미 정해져 있다고 옆방 사람에게 말합니다. (+P, -D, +I)
3. 당신은 다음주 토요일에 새 카페트를 깔 예정이기 때문에 세입자 방의 모든 가구들을 잠시 옮겨 달라고 부탁합니다. (+P, -D, +I)

#3 At your house

You hold a lease to a big house and rent out the other rooms. One of your housemate has asked to see you this evening.

1. You are late for the appointment to talk with him
2. When you meet him, he will want to hold a party on the coming Saturday. You explain that you have already scheduled painters to come this Saturday.
3. You also ask him to move all his furniture out of his bedroom to put in new carpeting next Saturday.

Role play #4 복사기 옆에서

당신: 어느 회사의 직원이며 지금은 회사의 특별 프로젝트를 맡은 리더입니다.
어디서: 회사 안의 복사기 옆
누구와: 같은 프로젝트를 맡은 동료 직원
장면: 당신이 복사를 하고 있을 때 동료 직원이 당신에게 다가왔습니다.

1. 동료 직원은 프로젝트를 같이 진행하고 있는 다른 동료 직원인 김철수 씨에게 메모를 전해 달라고 당신에게 부탁합니다. 하지만 당신은 김철수 씨가 오늘 몸이 안 좋아서 오늘 모임을 취소한 사실을 전합니다. (+P, -D, -I)
2. 당신은 복사를 끝내고 스태플러가 필요합니다만 지금 당신은 가지고 있지 않습니다. 그런데 지금 동료 직원은 스태플러를 가지고 있습니다. (+P, -D, -I)
3. 당신은 실수로 동료 직원의 스태플러를 떨어뜨리고 맙니다. 그런데 이 스태플러가 고장이 났습니다. (+P, -D, +I)

#4 At work by the photocopier

You are the project leader working on a special project with your staff. When you are at the photocopier, one of your staff members comes in.

1. He asks you to give a memo to Chulsoo Kim, another staff of yours, this evening. However, she is sick and her meeting with you has been canceled.
2. You need to staple some materials you've just copied, but you don't have a stapler. You notice that the staff member has a stapler.
3. You accidentally drop his stapler and it brakes.

Role play #5 사진 동호회에서

당신: 전국 연합 사진 동호회의 회원
어디서: 사진 동호회의 사무실
누구와: 동호회의 회장과
장면: 당신에게는 카메라 수리를 할 줄 아는 친구가 있습니다. 당신은 사진 동호회 회장의 고장 난 카메라를 비교적 저렴한 요금으로 이 친구에게 카메라 수리를 부탁하기로 했습니다. 그래서 당신은 동호회 회장의 카메라를 받기 위해 회장과 만나기로 했습니다.

1. 당신은 동아리 회장과의 약속 시간에 조금 늦었습니다. (-P, -D, -I)
2. 당신은 동아리 회장의 전화번호를 잊어버려서 회장에게 다시 전화번호를 물어봐야 합니다. (-P, -D, -I)
3. 동아리 회장은 당신에게 다음 달의 동호회 모임 일정을 정하기 위해 한 시간 더 이야기를 하자고 합니다. 하지만 당신은 조금 있다가 다시 회사로 들어가봐야 합니다. (-P, -D, +I)

#5 Photography club

You are a member of the local chapter of a national photography club. You have arranged to meet the club president to pick up his camera. Your friend, who repairs cameras at home for low prices, will do the repairs.

1. You are a little late for the appointment.
2. You have lost the club president's phone number, so you need to ask him to give it to you again.
3. You are asked to stay for about an hour and help him plan next month's meeting. But you are in a hurry because you have to get back to work after lunch.

Role play #6 교수님의 방에서

당신: 어느 대학교의 대학원생
어디서: 담당 교수님의 방에서
누구와: 당신의 담당 교수님과
장면: 당신은 다음 학기 과목을 상담하기 위해 담당 교수님 방에 와 있습니다. 교수님에게 이런저런 과목에 관해 물어 보던 중.

1. 당신은 교수님에게 여러 과목들에 대한 설명을 듣고 싶어서 각 과목에 대한 자세한 설명을 부탁합니다. (-P, -D, +I)
2. 교수님은 다음 학기 수업의 조교가 필요하다고 당신에게 말하며 당신에게 조교를 하지 않겠냐고 말합니다. 하지만 당신은 이번 학기 논문 발표로 많이 바빠서 조교를 못할 것 같습니다. (-P, -D, +I)
3. 교수님 방을 나오면서 당신은 실수로 테이블에 놓여져 있던 유리 장식품을 떨어뜨립니다. 다행이 깨어지지 않았습니다. (-P, -D, -I)

#6 In your academic advisor’s room

You are a graduate student. You meet your academic advisor to plan the next semester’s courses.

1. You want to hear the description of each course in detail. So you ask your advisor to explain each course.
2. Your advisor asks you whether you are interested in working as a teaching assistant for one of his courses next semester. But you might be busy next semester to prepare for your thesis.
3. You accidentally knock over a small figure on the advisor’s desk. It doesn’t break.