Developing a Test of Pragmatics of Japanese as a Foreign Language

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Masayuki Itomitsu

Graduate Program in East Asian Languages and Literatures

The Ohio State University

2009

Dissertation Committee:

Mari Noda, Advisor

Mineharu Nakayama

Charles J. Quinn, Jr.
Abstract

This dissertation reports development and validation studies of a Web-based standardized test of Japanese as a foreign language (JFL), designed to measure learners’ off-line grammatical and pragmatic knowledge in multiple-choice format. Targeting Japanese majors in the U.S. universities and colleges, the test is designed to explore possible relationships between JFL learners’ grammatical knowledge and domains of pragmatic knowledge. Pragmatic knowledge is defined as knowledge of “[linguistic] forms necessary for productively conveying speech intentions, and assigning pragmatic meaning to interlocutor utterance” (Röver 2005), and is operationalized as the ability to associate linguistic features in specific contexts with appropriate speech acts, routines, and speech styles. Grammatical knowledge includes well-formedness of isolated sentences at the morpho-syntactic level, independent of contextual information. Research questions address reliability of the test, the relationship between JFL learners’ grammatical and pragmatic knowledge (Bardovi-Harlig and Dörnyei 1998), as well as the relationship between the overall scores of this multiple-choice test and the ACTFL Proficiency Guidelines Levels for speaking.

The data from 110 participants of JFL learners in four different universities revealed that the test was overall reliable. The total scores of the test positively correlated
with length of study/exposure, levels of instruction, and proficiency ratings based on the ACTFL Guidelines. As for the relationship among grammatical and domains of pragmatic knowledge, the study found complex interrelationships: correlation analyses showed that the four subsections (Grammar, Routines, Speech Acts, and Speech Styles) are significantly correlated overall, but the strengths of correlation varied depending on the levels of instruction and proficiency levels. The Routines section was somewhat separate and was not highly correlated with other sections at upper levels. The data also suggest that, as far as JFL is concerned, the knowledge of speech styles is in closer relationship with the knowledge of grammar and speech acts, compared with studies on learners of English as a Second/Foreign Language. Exploratory factor analysis, however, found unexpected patterns of factor loadings, not matching the internal structure of the test, suggesting that the hypothesized components of language ability may not be empirically distinguishable. Construct validity of the test is defensible only for the overall score being a valid indication of communicative language ability in general. The extent to which each section scores are valid indications of corresponding hypothesized separate components of the language ability needs to be further investigated in future studies. Pedagogical implications for testing knowledge of natives’ prototypical communication patterns, limitations of the study, and future development of the test are discussed.
Dedication

Dedicated to my mother, Nobuko Itomitsu

with affection and admiration
Acknowledgments

First and foremost, I would like to thank my advisor, Professor Mari Noda, for her intellectual support, guidance, and encouragement that made this dissertation possible. It has been a privilege to work with her on a number of projects, including this dissertation project. I would also like to thank my committee members, Professor Charles Quinn Jr. and Professor Mineharu “JJ” Nakayama for their valuable comments and suggestions.

I am grateful to the following individuals who participated in the pilot stage of the project and provided valuable comments that improved the final product greatly: Hiroko Baker, Seth Goss, Teppei Kiyosue, Mari Noda, Kaori Suzuki, and Ai Terada of The Ohio State University; Sanae Eda, Rie Maruyama, and Mishizu Ohtake of the University of Kansas; Ginger Marcus and Kaori Nakata of Washington University in St. Louis; James-Henry Holland of Hobart and William Smith Colleges; and Patricia Wetzel of Portland State University. I would also like to thank Paul Kotheimer and Yuko Kuwai of The Ohio State University for their help with recording the audio portion of the test.

I am indebted to the students of Japanese at The Ohio State University, the University of Cincinnati, the University of Kansas, and Washington University in St. Louis, who took time from their busy schedule to participate in the operationalized test. Special thanks go to Noriko Fujioka-Ito of University of Cincinnati, Sanae Eda of the
University of Kansas, and Ginger Marcus of Washington University in St. Louis who helped me with the data collection process.

Last but not least, I would like to express my sincere appreciation to my family in Okinawa for their understanding and encouragement, and to Sheldon Safko, my partner, without whose long-term support, encouragement and patience I would not be where I am right now.
Vita

March 1970 ................................. Born — Okinawa, Japan

1992 ................................. B.A. English Language and Literature, University of the Ryukyus, Okinawa, Japan

1994 ................................. B.S. English, Northeast Missouri State University (Truman State University), Kirksville, MO

1996 ................................. M.A. Japanese, The Ohio State University, Columbus, OH


2002 – 2006 ................................. Graduate Research/Teaching Associate, The Ohio State University

2006 – 2007 ................................. Assistant Professor, Defense Language Institute, Monterey, CA

2008 – 2009 ................................. Lecturer, Foreign Language Center, The Ohio State University

2009 – present ................................. Assistant Professor, Linfield College, McMinnville, OR

Publications


Fields of Study

Major Field: East Asian Languages and Literatures
Area of Specialization: Japanese Language Pedagogy
Note on Japanese Transcriptions

Different romanization systems as well as *kanji/kana* are used throughout this document. In citations and discussion about the citations, Japanese examples are transcribed in a way consistent with the original documents, unless otherwise noted. Otherwise, Japanese orthography (*kanji/kana*) is used in the main text.

List of Abbreviations and Terminology

ACTFL = American Council on Teaching of Foreign language  
CALT = Computer-Assisted Language Teaching  
CCSARP = Cross-Cultural Speech Act Realization Project  
CLA = Communicative Language Ability  
DCT = Discourse Completion Task  
EFL = English as a Foreign Language  
ESL = English as a Second Language  
GR = Grammar  
ID = Item Discrimination  
IF = Item Facility  
IFID = Illocution Force Identification Device  
ILR = Interagency Language Roundtable  
JFL = Japanese as a Foreign Language  
JSL = Japanese as a Second Language  
OPI = Oral Proficiency Interview  
RT = Routines  
SA = Speech Acts  
SS = Speech Styles

Kana = Japanese syllabery systems (hiragana and katakana)  
Kanji = Chinese characters  
Keigo = polite language (styles)
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>Note on Japanese Transcriptions</td>
<td>ix</td>
</tr>
<tr>
<td>List of Abbreviations and Terminology</td>
<td>ix</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xv</td>
</tr>
<tr>
<td><strong>Chapter 1: Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1. Background: Rationales for the Study</td>
<td>1</td>
</tr>
<tr>
<td>1.2. The Study</td>
<td>5</td>
</tr>
<tr>
<td>1.3. Organization of the Dissertation</td>
<td>6</td>
</tr>
<tr>
<td><strong>Chapter 2: Literature Review</strong></td>
<td>7</td>
</tr>
<tr>
<td>2.1. Definition Issues: Grammatical and Pragmatic knowledge in Communicative Language Ability</td>
<td>7</td>
</tr>
<tr>
<td>2.1.1. Hymes (1974)</td>
<td>9</td>
</tr>
<tr>
<td>2.1.2. Canale and Swain (1980)</td>
<td>10</td>
</tr>
<tr>
<td>2.1.3. Bachman and Palmer (1996)</td>
<td>11</td>
</tr>
<tr>
<td>2.1.5. Unitary vs. Multi-componential Views of Communicative Language Ability</td>
<td>13</td>
</tr>
<tr>
<td>2.1.6. Summary</td>
<td>14</td>
</tr>
<tr>
<td>2.2. Components of Communicative Language Ability</td>
<td>15</td>
</tr>
<tr>
<td>2.2.1. Relationship between Grammatical and Pragmatic Competence</td>
<td>15</td>
</tr>
<tr>
<td>2.2.2. Components of Pragmatic Knowledge</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2.1. Interlanguage Pragmatics: Pragmalinguistics and Sociopragmatics</td>
<td>21</td>
</tr>
<tr>
<td>2.2.2.2. Speech Acts</td>
<td>24</td>
</tr>
<tr>
<td>2.2.2.3. Routines</td>
<td>37</td>
</tr>
<tr>
<td>2.2.2.4. Speech Styles</td>
<td>41</td>
</tr>
<tr>
<td>2.3. Tests of Interlanguage Pragmatics</td>
<td>43</td>
</tr>
<tr>
<td>2.3.1. Tests of Sociopragmatics</td>
<td>45</td>
</tr>
<tr>
<td>2.3.2. Tests of Pragmalinguistics</td>
<td>49</td>
</tr>
</tbody>
</table>
2.3.3. Relationship among Components of Pragmatic Knowledge ................................................................. 51
2.3.4. Test of JSL/JFL Pragmalinguistics ................................................................................................. 55
2.4. Summary .................................................................................................................................................. 63

Chapter 3: The Study ........................................................................................................................................... 65
3.1. Purpose .................................................................................................................................................... 65
3.2. Construct ................................................................................................................................................ 66
3.3. Research Questions ............................................................................................................................. 68
3.4. Instrument Design ............................................................................................................................... 71
3.5. Content Specifications ......................................................................................................................... 72
3.6. Format Specifications ........................................................................................................................... 76
3.7. Pilot Testing ............................................................................................................................................ 80

Chapter 4: Methods .......................................................................................................................................... 90
4.1. Participants ............................................................................................................................................ 90
4.2. Materials ............................................................................................................................................... 92
4.2.1. The Questionnaire (“About You”) .................................................................................................. 92
4.2.2. The Operationalized Test (“Test of Pragmatics”) ........................................................................ 95
4.2.3 Proficiency Rating Data ................................................................................................................ 96

Chapter 5: Results ........................................................................................................................................... 99
5.1. General Test Characteristics ............................................................................................................... 99
5.2. Reliability of the Test and Sections ..................................................................................................... 100
5.3. Relationship between the Test Scores and Length of Study/Exposure ............................................. 106
5.4. Relationship between the Test Scores and Levels of Instruction ...................................................... 114
5.5. Relationship between the Test Scores and Proficiency Ratings ....................................................... 122
5.6. Relationship among the Section Scores ............................................................................................ 130

Chapter 6: Discussions and Conclusion ........................................................................................................ 141
6.1. Summary of the Findings and General Discussions ........................................................................ 141
6.1.1 Test Reliability .................................................................................................................................. 142
6.1.2 Relationships between the Test Scores and Length of Study/Exposure, Levels of Instruction .................................................................................................................. 144
6.1.3 Relationship between the Test Scores and Proficiency ratings ..................................................... 146
6.1.4 Relationship among Components of Communicative Language Ability .................................. 148
6.2. Limitations of the Study ..................................................................................................................... 152
6.3. Pedagogical Implications .................................................................................................................... 155
6.4. Suggestions for Future Research ....................................................................................................... 162
6.4.1. Test Improvement .......................................................................................................................... 162
6.4.2. Future Research Projects................................................................. 164
6.5. Conclusion ......................................................................................... 168

References.................................................................................................. 170

Appendix A: Content Specifications for the Grammar Section.................. 181
Appendix B: Content Specifications For The Routines Section................ 187
Appendix C: Content Specifications For The Speech Acts Section........... 191
Appendix D: Content Specifications For The Speech Styles Section .......... 197
Appendix E: Solicitation E-Mail Message ................................................ 198
Appendix F: The Operationalized Test (Sample Test Items and Deleted Items) .... 199
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1: Pilot Test Results Summary (Grammar, GR)</td>
<td>84</td>
</tr>
<tr>
<td>Table 3.2: Pilot Test Results Summary (Routines, RT)</td>
<td>85</td>
</tr>
<tr>
<td>Table 3.3: Pilot Test Results Summary (Speech Acts, SA)</td>
<td>86</td>
</tr>
<tr>
<td>Table 3.4: Pilot Test Results Summary (Speech Styles, SS)</td>
<td>87</td>
</tr>
<tr>
<td>Table 3.5: Average Scores of the Expert Ratings in Sections</td>
<td>89</td>
</tr>
<tr>
<td>Table 4.1: Participants during Phase One (Fall 2008)</td>
<td>91</td>
</tr>
<tr>
<td>Table 4.2: Participants during Phase Two (Spring 2009)</td>
<td>91</td>
</tr>
<tr>
<td>Table 4.3: Summary of Learners’ Background Information</td>
<td>94</td>
</tr>
<tr>
<td>Table 4.4: Participants with Proficiency Rating (Phase One)</td>
<td>97</td>
</tr>
<tr>
<td>Table 4.5: Participants with Proficiency Rating (Phase Two)</td>
<td>97</td>
</tr>
<tr>
<td>Table 5.1: Descriptive Statistics of the Test and Sections</td>
<td>100</td>
</tr>
<tr>
<td>Table 5.2: Cronbach’s Alpha for the Test and Sections</td>
<td>101</td>
</tr>
<tr>
<td>Table 5.3: Item Statistics for the Grammar Section</td>
<td>103</td>
</tr>
<tr>
<td>Table 5.4: Item Statistics for the Routines Section</td>
<td>104</td>
</tr>
<tr>
<td>Table 5.5: Item Statistics for the Speech Acts Section</td>
<td>105</td>
</tr>
<tr>
<td>Table 5.6: Item Statistics for the Speech Styles Section</td>
<td>106</td>
</tr>
<tr>
<td>Table 5.7: Descriptive Statistics of Total Scores by Length of Study/Exposure</td>
<td>108</td>
</tr>
<tr>
<td>Table 5.8: Descriptive Statistics of Section Scores by Length of Study/Exposure</td>
<td>110</td>
</tr>
<tr>
<td>Table 5.9: Learners in Different Levels of Instruction from the Four Universities</td>
<td>115</td>
</tr>
<tr>
<td>Table 5.10: Descriptive Statistics of Total Scores by Levels of Instruction</td>
<td>115</td>
</tr>
<tr>
<td>Table 5.11: Descriptive Statistics of Section Scores by Levels of Instruction</td>
<td>118</td>
</tr>
<tr>
<td>Table 5.12: Descriptive Statistics of Total Score by Proficiency Rating</td>
<td>123</td>
</tr>
<tr>
<td>Table 5.13: Descriptive Statistics of Section Scores by Proficiency Rating</td>
<td>125</td>
</tr>
<tr>
<td>Table 5.14: Conversion Scheme for Proficiency Rating</td>
<td>129</td>
</tr>
<tr>
<td>Table 5.15: Correlation between Proficiency Rating and Section Scores (N=50)</td>
<td>130</td>
</tr>
<tr>
<td>Table 5.16: Correlations of Section Scores among Low-Proficiency Students (Intermediate-Mid and below), N=32</td>
<td>131</td>
</tr>
<tr>
<td>Table 5.17: Correlations of Section Scores among High-Proficiency Students (Intermediate-High and above), N=18</td>
<td>131</td>
</tr>
<tr>
<td>Table 5.18: Correlation among Section Scores (all subjects, N=110)</td>
<td>132</td>
</tr>
<tr>
<td>Table 5.19: Correlation among Section Scores (2nd Level students, N=41)</td>
<td>132</td>
</tr>
</tbody>
</table>
Table 5.20: Correlation among Section Scores (3rd Level students, N=35) .......... 133
Table 5.21: Correlation among Section Scores (4th Level students, N=14) .......... 133
Table 5.22: Correlation among Section Scores (5th Level students, N=10) .......... 133
Table 5.23: Factors with eigenvalues > 1 .............................................................. 135
Table 5.24: Factor Loadings, Promax rotation with Kaiser normalization .......... 138
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 3.1: Item Format Example</td>
<td>78</td>
</tr>
<tr>
<td>Figure 3.2: Pilot Test Screen Capture</td>
<td>82</td>
</tr>
<tr>
<td>Figure 5.1: Box-Plot of Total Scores by Length of Study/Exposure</td>
<td>108</td>
</tr>
<tr>
<td>Figure 5.2: Box-Plots of Section Scores by Length of Study/Exposure</td>
<td>111</td>
</tr>
<tr>
<td>Figure 5.3: Mean Plots of the Section Scores by Length of Study/Exposure</td>
<td>113</td>
</tr>
<tr>
<td>Figure 5.4: Box-Plot of Total Scores by Levels of Instruction</td>
<td>116</td>
</tr>
<tr>
<td>Figure 5.5: Box-Plots of Section Scores by Levels of Instruction</td>
<td>119</td>
</tr>
<tr>
<td>Figure 5.6: Mean Plots of the Section Scores by Levels of Instruction</td>
<td>121</td>
</tr>
<tr>
<td>Figure 5.7: Box-Plots of Total Scores by Proficiency Rating</td>
<td>123</td>
</tr>
<tr>
<td>Figure 5.8: Box-plots of Section Scores by Proficiency Rating</td>
<td>126</td>
</tr>
<tr>
<td>Figure 5.9: Mean Plots of Section Scores by Proficiency Rating</td>
<td>128</td>
</tr>
<tr>
<td>Figure 5.10: Scree Plot</td>
<td>136</td>
</tr>
<tr>
<td>Figure 6.1: Role-play Rating Criteria by Grabowski (2008: 161)</td>
<td>165</td>
</tr>
<tr>
<td>Figure 6.2: Interview Rating Criteria by Bachman and Palmer (1982: 456-7)</td>
<td>166</td>
</tr>
</tbody>
</table>
1.1. Background: Rationales for the Study

In foreign language education and in language testing, pragmatic competence — the ability to use language appropriately in the given context — has been identified as an important component of communicative language ability\(^1\) (Canale and Swain 1980; Bachman 1990). The National Standards in Foreign Language Education Project (Student Standards Task Force of the National Standards in Foreign Language Education Project 1996: 11) asserts that “formerly, most teaching in foreign language classrooms concentrated on the how (grammar) to say what (vocabulary). While these components of language remain crucial, the current organizing principle of language study is communication, which also highlights the why, to whom, and the when (the sociolinguistic and cultural aspects of language)” (original emphasis). Recent publications and conferences on pragmatic competence in formal instructional settings (Rose and Kasper 2001; Yoshimi and Wang 2006) reflect an increasing interest among

\(^1\) Knowledge is thought of as a domain of information structure in memory, and competence and ability as a capacity to use knowledge in a meaningful way. See Bachman and Palmer (1996: 66-7).
researchers and educators in fostering pragmatic competence in face-to-face communication.

Pragmatic knowledge is deemed important in teaching and learning of Japanese as a foreign language (JFL) in the U.S. The Framework for Introductory Japanese Language Curricula in American High Schools and Colleges, first published in 1993, states: “A language is not a mere collection of words, phrases, […] but a complex structure of behaviors dictated by conventions, including how speech-sounds are combined to form words, words to form phrases and sentences, and sentences to form discourse that is both socially appropriate and culturally meaningful” (Unger 2001: 351, added emphasis). Subsequently, there have been increasing efforts to teach and measure development of pragmatic ability and/or awareness in JFL in the U.S. (Aoshima, Ishihara, and Akikawa 2008; Cohen and Ishihara 2005; Ishihara 2007; Ishihara, Aoshima, and Akikawa 2008; Rose and Kasper 2001).

Assessment of pragmatic competence in second language testing, however, is still in its infancy (Röver 2005; McNamara and Röver 2006), despite the fact that many influential models of communicative language ability have included pragmatic competence since 1980s (Bachman 1990; Canale and Swain 1980; Savignon 1983).

---

2 This dissertation focuses on pragmatic knowledge necessary for face-to-face, verbal interaction in learning of non-primary language (i.e., learning a language other than the native tongue). Communication through the written language is beyond the scope of this dissertation study.

3 Following conventions in the literature, JFL/EFL refer to learning Japanese or English “at home” (i.e., in learners’ home country, where the target language is not the official language), while JSL/ESL refer to learning Japanese or English in the target countries.

4 This document was produced in conjunction with the College Board and Educational Testing Service, which developed a standardized achievement test (SAT II Subject Test in Japanese) (Unger, 2001: 338).
There are only a few studies on standardized tests of pragmatic knowledge and skills for English as a Second Language (ESL) (Hudson, Detmer, and Brown 1995). For Japanese, there is only one comprehensive study (Yamashita 1996) that reports on measurements of pragmatic knowledge and abilities in Japanese as a second language (JSL) following Hudson et al. (1995). If nurturing pragmatic knowledge is among a program’s educational goals, that program’s curriculum should have the means to measure what pragmatic knowledge the learners have gained throughout the curriculum. Kasper and Rose (2001: 9) assert: “Especially in instructional contexts where formal testing is regularly performed, curricular innovations that comprise pragmatics as a learning objective will be ineffective as long as pragmatic ability is not included as a regular and important component of language tests.”

Many widely recognized large-scale standardized tests of JSL and JFL, however, primarily test grammar, vocabulary, and reading (Eda, Itomitsu, and Noda 2008), without explicit specifications for pragmatic knowledge as part of the test construct. The Oral Proficiency Interview (OPI) developed by American Council on the Teaching of Foreign Languages (ACTFL) — perhaps the best-known testing procedure to measure “learners’ functional competency” (Breiner-Sanders et al. 2000) — views language proficiency as a unitary ability, without addressing pragmatic competence per se (Bachman 1990; Makino 1991). Many language programs utilize performance tests (role-plays, oral interview tests, etc.) to test what learners can demonstrate in performing a task (Aoshima, Ishihara, and Akikawa 2008; Ishihara, Aoshima, and Akikawa 2008; Noda et al. 2002). While such classroom-based assessments are certainly useful and
important, they may also be time-consuming, labor-intensive, and costly. Small-scale interviews and role-plays may allow intra-program comparisons, but probably not inter-program comparisons unless the test is well standardized and the test reliability is well established. Recent advances in computer-assisted language testing (CALT) (Chapelle and Douglas 2006) are promising, as they can play a complementary role in assessing pragmatic competence in a standardized manner, since they allow multimedia presentation of the test items over the Internet at relatively low-cost. The potential of CALT is certainly worth exploring.

In defining pragmatic competence, the relationship between pragmatic competence and other attributes — assuming that communicative language ability is multicomponential — is another research topic that deserves more attention (Kasper and Rose 2003). For ESL, there have been some studies on the relationship between grammatical and pragmatic knowledge and general proficiency (Takahashi and Beebe 1987) as well as the relationship between pragmatic awareness with grammatical awareness in terms of learning environment (Bardovi-Harlig and Dörnyei 1998; Niezgoda and Röver 2001). There are few studies, however, on pragmatic knowledge and other such attributes as general proficiency levels or knowledge in JFL, except for Takamiya (2006). Whether these discrete components of communicative language ability can or should be assessed independently or together remains one of the more significant research questions in language testing (Purpura 2008).
1.2. The Study

The study reported in this dissertation is intended as a step toward a more comprehensive assessment of communicative language ability. More specifically, this dissertation reports the development and validation studies of a Web-based, discrete-point, standardized test of JFL, designed to measure learners’ pragmatic knowledge “off-line” (processing without simultaneously engaging in language production, speech interaction, or communication). Targeting Japanese majors in the U.S., the test is designed to examine potential relationships between JFL learners’ grammatical knowledge and domains of pragmatic knowledge — speech acts, routines, and speech styles. Such a test, combined with other test instruments and techniques such as role-play, etc., can function as an important and practical tool for monitoring development of second language pragmatic competence, and for contributing to second language acquisition research in general (Bachman and Cohen 1998; Cohen 2004).

This dissertation focuses on reliability and validation of the Web-based test. The study also investigates the relationship between JFL learners’ grammatical and pragmatic knowledge, the relationship among the three subcomponents of the pragmatic knowledge, and the relationship between the test score and JFL learners’ proficiency ratings obtained by interview tests developed by the ACTFL (Breiner-Sanders et al. 2000).
1.3. Organization of the Dissertation

This dissertation is organized as follows. Chapter 2 presents a review of literature on studies on communicative language ability, interlanguage pragmatics, and the testing of pragmatic knowledge, with a special focus on JFL. Chapter 3 explains this study in detail, including its purpose, theoretical construct, operationalization, research questions, the test design (format and content specifications), as well as the pilot test conducted prior to the operationalized test. Chapter 4 explains methods (relating to participants, materials, and procedures) used for data collection through the operationalized test. Chapter 5 provides statistic analyses of the collected data. Finally, Chapter 6 discusses the results and considers limitations of the study as well as pedagogical implications and possible directions for future research.
Chapter 2: Literature Review

This chapter provides review of previous studies that are relevant to the development of a test of pragmatics of JFL. Three major topics have been identified: (1) definition issues surrounding grammatical and pragmatic knowledge in communicative competence and communicative language ability, (2) major studies in interlanguage pragmatics, and (3) testing interlanguage pragmatics. They are reviewed separately in the following sections.

2.1. Definition Issues: Grammatical and Pragmatic knowledge in Communicative Language Ability

Grammar, and pragmatics especially, have been notoriously hard to define. There exist numerous definitions from various perspectives in the literature (Levinson 1983; Crystal 1987; Mey 1993; Purpura 2004; Ariel 2008). This difficulty is probably due to the complexity of what affords meaning: the intertwined relationship among grammar, semantics, and pragmatics, as well as the complexity of language use in culture in general. There seems to be a general consensus, however, to treat grammatical knowledge as a
“code” component, and treat it separately from pragmatic knowledge, which is often treated as a “use”, “user” or “inference” component (Ariel 2008; Röver 2005).

Mey (1993: 42) defines pragmatics as “the study of the conditions of human language uses as there are determined by the context of society.” Similarly, Levinson (1983: 24) includes “Pragmatics is the study of the ability of language users to pair sentences with the contexts in which they would be appropriate” as one of the favored definitions in the field, while pointing out the difficulty of defining “context” and “appropriateness,” as well as the potential overlap with sociolinguistics.

Different definitions and perspectives also exist for the notion of communicative competence or communicative language ability – a notion that has proved to be one of the most influential theoretical developments in language education as it helped redefine the objectives of foreign language instruction and target language proficiency (Lee 2006). One of the instances of this wide range of perspectives was witnessed at the Language Proficiency Assessment (LPA) Symposium, held in 1981, which failed to reach a consensus on a working definition of communicative fluency, given the diversity of perspectives, theories and research findings (Rivera 1984). Whether or not grammatical competence is not a good predictor of communicative competence is also a contentious issue in the literature (Canale and Swain 1980; Liu 2006).

There seems to be an emerging agreement that the communicative language ability is multicomponential, and it includes more than grammatical knowledge (Purpura 2008). Röver (2005: 5) states “definitions of communicative competence minimally tend to include a code component, describing a language learner’s procedural and declarative
knowledge of the rules of syntax, semantics, morphology, and phonology [; and] a *use* component, describing a language learner’s knowledge of the social norms governing language use and the assignment of linguistic options to speech intentions for production and comprehension” (emphasis added). It is this distinction that serves as the basic guideline for this dissertation project.

In the following section, four major contributions to the development of communicative language ability in the context of language testing are reviewed: Hymes (1974), Canale and Swain (1980), Bachman and Palmer (1996), and Purpura (2004). See Purpura (2008) for a more comprehensive review on historical development of the concept of communicative language ability from the 1960s.

2.1.1 Hymes (1974)

The term *communicative competence* was coined by Dell Hymes in 1966, as a critical reaction to the then-popular and dominant Chomskian perspective on grammar (Purpura 2004). Hymes (1974: 92-3) argues that “Linguistic theory must extend the notion of competence to include more than the grammatical [, as] the grammatical mechanisms [Chomsky] considers can never account for appropriateness. Appropriateness involves […] a kind of competence regarding situations and relations of sentences to them.” Hymes proposed four aspects of competence: “(a) systematic potential – whether and to what extent something is not yet realized, and, in a sense, not yet known; it is to this that Chomsky in effect reduces competence; (b) appropriateness – whether and to what extent something is in some context suitable, effective, or the like;
(c) occurrence – whether and to what extent something is done; (d) feasibility – whether and to what extent something is possible, given the means of implementation available” (ibid, p.95). In this model, grammatical knowledge belongs to systemic potential, while pragmatic knowledge belongs to appropriateness.

2.1.2. Canale and Swain (1980)

Canale and Swain (1980) bring Hymes’ notion of communicative competence into the context of language teaching and testing in the very first volume of the journal Applied Linguistics. In this seminal article, they interpret the four aspects of Hymes’ communicative competence as grammatical, sociocultural, probabilistic, and psycholinguistic aspects of communicative competence, respectively. They further propose a three-component theoretical framework for communicative competence: (1) grammatical knowledge (including knowledge of lexical items and of rules of morphology, syntax, sentence-grammar semantics, and phonology; (2) sociocultural competence (consisting of sociocultural rules of use and rules of discourse), and (3) strategic competence (verbal and non-verbal strategies to compensate for breakdowns in communication and/or insufficient competence (p.29-31). Discourse competence is presented as a fourth category, distinct from sociocultural competence, in Canale (1983). They also speculate, “within each of the three components of communicative competence […], there will be a subcomponent of probability rules of occurrence […i.e.,] the knowledge of relative frequencies of occurrence” (p. 31). In their model, pragmatic knowledge — “the extent to which certain propositions and communicative functions are
appropriate within a given sociocultural context” (p.30) — belongs to sociocultural competence, which is distinct from the grammatical competence.

Van Ek (1986) follows Canale and Swain’s direction, and his definition of communicative competence includes (1) linguistic competence, (2) sociolinguistic competence (awareness of ways in which the choice of language-forms is constrained by such conditions as setting, relationship between communication partners, etc.), (3) discourse competence, (4) strategic competence, (5) socio-cultural competence (awareness of the sociocultural context in which the language is used by native speakers and of ways in which this context affects the choice and the communicative effect of particular language forms), (6) and social competence (the ability to use social strategies appropriate to the achievement of one’s communicative goals). His work has been further developed in van Ek and Trim (1991), which has been widely regarded as a model for the notional-functional approach to foreign language.

2.1.3 Bachman and Palmer (1996)

The model of communicative language ability in Bachman (1990), revised in Bachman and Palmer (1996), is considered as the state-of-the-art in the field of language testing (Purpura, 2008). Bachman and Palmer view communicative language ability as a dynamic system, in which strategic competence (with goal setting, assessment, and planning) is supplied, through affective schemata (affective or emotional correlates of

---

5 His distinction between sociolinguistic and socio-cultural competences does not seem to be clear-cut. His sociolinguistic competence may include knowledge of linguistic forms associated with certain contextual features, and socio-cultural competence includes knowledge of appropriate choices of strategies. This issue is addressed again in Section 2.3.
topical knowledge), with topical knowledge (or world knowledge), language knowledge, and personal characteristics. In this view, language knowledge consists of organizational knowledge (grammatical and textual knowledge) and pragmatic knowledge (functional knowledge and sociolinguistic knowledge). They define pragmatic knowledge as knowledge of “how utterances or sentences and texts are related to the communicative goals of the language user and to the features of the language user setting” (p.68).


Purpura (2004) also defines grammatical knowledge for the purpose of language testing. His model of grammatical knowledge (p. 91) is concerned with accuracy and meaningfulness, and consists of grammatical form and meaning. More specifically, grammatical form and meaning include a sentential level (phonological/graphological, lexical, morphosyntactic) and suprasentential level (cohesion, information management, and interactional). Grammatical knowledge is treated separately from pragmatic knowledge, which involves appropriateness, conventionality, naturalness, and acceptability. The pragmatic knowledge consists of factors of contextual (interpersonal), sociolinguistic (gender, age, status, dialect, politeness markers and registers), sociocultural (cultural references and norms), psychological (affective stance), and rhetorical (organizational modes and genres) meanings.
2.1.5. Unitary vs. Multi-componential Views of Communicative Language Ability

All the models of communicative language ability reviewed so far assume that such ability can be further separated into several components. However, there exists another approach in defining language proficiency as a single, unitary ability. This “unitary” approach was prevalent in the 1970s (Bachman and Clark 1987; Stansfield 2008), and has been adapted by the Interagency Language Roundtable (ILR) scale (See Herzog 2009 for overview) and the Proficiency Guidelines by the ACTFL (Breiner-Sanders et al. 2000). Bachman (1990: 329) terms the unitary view by ILR/ACTFL as “Real-Life” (RL) approach, and the multi-componenstal view by Bachman and Palmer (1996) as “Interactional/Ability” (IA) approach, and explains the difference between the two as follows:

These differences derive quite logically from the difference between the RL and IA definitions of language proficiency and test authenticity. If language proficiency is defined in terms of language performance, as is the case in the RL approach, there is no basis for distinguishing separate abilities. That is, abilities cannot be directly observed in performance, and if we define proficiency as performance, then it follows that separate abilities cannot be measured as part of proficiency. It should be noted that proponents of the ILR/ACTFL oral interview implicitly recognize the multicomponental nature of language proficiency when they speak of the ‘functional trisection’ of function, content, and accuracy that characterizes the scale definitions […], and when they provide raters with profile checklists as aids to rating […]. Nevertheless, to the best of my understanding, none of these components is actually reported in the scores. If, on the other hand, one chooses to define language proficiency in terms of component abilities, as is the case with the IA approach, then the implication is that we should attempt to measure these component abilities and report scores for these measures separately.

---

6 For an overview and critiques of the ACTFL OPI, see Yoffe (1997).
Makino (1991: 28) concurs with Bachman’s view and states that “currently, ACTFL’s single, global rating is determined according to an overall judgment based on the four pillars of proficiency assessment criteria (global tasks and functions, context/content, accuracy, and text type). It is desirable to provide assessment for each of the pillars along with the overall rating. This is because it would improve the validity of the test, and it could provide detailed linguistic diagnosis to the test takers” (my translation).

2.1.6. Summary

In sum, many influential models of communicative language ability are multicomponential in nature, and treat grammatical and pragmatic knowledge as distinct components. However, there remain many unanswered questions. Purpura (2008: 63) summarizes them as follows:

No consensus has been reached as to what exact components constitute a comprehensive model of CLA [Communicative Language Ability], how the components might interact, how the components of CLA are acquired and develop — alone or together, how knowledge of these components is organized in the test-taker’s mind or how these knowledge representations generally change as test-takers advance along the interlanguage continuum, or how knowledge of these discrete components might integrate in ways that they can be used as resources for accomplishing meaningful activities under different conditions across diverse contexts. In short, much research still remains.

There are a few studies, however, that explore the relationship among components of communicative language ability, including the relationship between grammatical and pragmatic knowledge. The following section reviews the representative studies that investigate the grammar-pragmatic relationship.
2.2. Components of Communicative Language Ability

2.2.1. Relationship between Grammatical and Pragmatic Competence

As many conceptualizations of communicative language ability are multicomponential in nature, possible relationships among the components — degree of independence and interdependence — have been on the research agendas. Kasper and Rose (2003) devote an entire chapter (Chapter 5) on the topic of the development of pragmatics and grammar, citing Bardovi-Harlig and Dörnyei (1998) as the work that renewed attention to the relationship between pragmatic and grammatical competence.

Bardovi-Harlig and Dörnyei investigated whether the learning environment influences learners’ awareness of pragmatic and grammatical errors. They used videotaped scenarios to test 543 ESL and EFL learners and their teachers in Hungary, Italy, and the U.S. The results showed that whereas the EFL learners and teachers consistently identified and ranked grammatical errors as more serious than pragmatic errors, the ESL learners and their teachers showed the opposite pattern, ranking pragmatic errors as more serious than grammatical errors. They define grammatical and pragmatic awareness by stating that “grammar [refers to] the accuracy of structure, including morphology and syntax, whereas pragmatics addresses language use and is concerned with the appropriateness of utterances given specific situations, speakers, and content” (p.233). The grammatical errors in the test reflect common errors reported in the ESL literature, including a zero object (*yes I would like), a double marking of the past (*I didn't brought it), non-use of do-insertion (*if you not need it), inversion in an embedded question (*can you tell me where is the library), etc. (p.240 footnote). The
pragmatic problems in the test included the lack of an explanation formula in a refusal addressed to a teacher, a bare imperative used for a request (without an alerter) addressed to a classmate, the lack of an explanation or a query preparatory formula with a request (I would like you to) addressed to a teacher, etc. (p.241 footnote) The speech acts included in the test include request, apology, suggestion, and refusal.

Their examples of the test items are as follows (p.241):

**Pragmatic Item**
It's Anna's day to give her talk in class, but she's not ready:
Teacher: Thank you, Peter, that was very interesting. Anna, it's your turn to give your talk.
!Anna: I can't do it today but I will do it next week.

**Grammar Item**
Peter has borrowed a book from his friend, George. George needs it back, but Peter has forgotten to return it.
George: Peter, do you have the book that I gave you last week?
!Peter: I'm really sorry but I was in a rush this morning and I didn't bring it today.

**Appropriate/Correct Item**
Anna invites her friend to her house, but she can't come.
Anna: Maria, would you like to come over this afternoon?
!Maria: I'm sorry, I'd really like to come, but I have a difficult history test tomorrow.

Schauer (2006) replicated the Bardovi-Harlig and Dörnyei study with 53 participants (16 German-speaking ESL group in a British University, 17 EFL group in Germany, and 20 British English native-speaking controls). The results were congruent with the original study, thus confirming that the learning environment plays an important role in priming the learners’ awareness, regardless of their native languages and the sites (different countries) of the learning environment.
Niezgoda and Röver (2001) also replicated the Bardovi-Harlig and Dörnyei study, and found that their EFL group rated pragmatic errors as being more severe than grammatical errors. The difference is attributed to the fact that their EFL group is highly selected and motivated group and more closely match the ESL group in the original study. They claim “this is evidence that pragmatic awareness can indeed be acquired in the FL environment, or more specifically, in the FL classroom. In addition, it strongly encourages further observational studies and instructional research to determine specifically how pragmatics can be taught in an EFL classroom” (p.78).

Xu, Case and Wang (2009) also used the Bardovi-Harlig and Dörnyei questionnaire to examine the influence of length of residence in the target language community and overall L2 proficiency (undergraduate vs. graduate) on L2 pragmatic competence with a reference to L2 grammatical competence. This study with 126 participants revealed that both length of residence and overall L2 proficiency influenced L2 pragmatics significantly with overall L2 proficiency demonstrating a stronger influence. Findings also showed that there was a strong and positive correlation between pragmatic and grammatical competence for advanced participants and all participants as a group.

Takamiya (2006) conducted a study to investigate if the trends seen in the ESL and EFL studies reported above are applicable to learners of Japanese. She recruited 49 JFL and 41 JSL learners to investigate if different learning environments and general proficiency levels influence pragmatic and grammatical awareness. The instrument used in this study was a written format, consisting of 23 scenes (10 with grammatical errors,
10 with pragmatic errors, and 3 correct and appropriate utterances). The task is the same as the Bardovi-Harlig and Dörnyei study: the subjects were asked to identify any errors, and rate the severity of the identified errors. The results showed that the JSL group noticed more errors, but rated grammatical and pragmatic errors almost identically. The JFL learners, on the other hand, rated grammatical errors as more severe than the pragmatic errors. There were differences in patterns of recognition between a low-proficiency group (at the intermediate level, in fifth semester) and a high-proficiency group (at the advanced level, in seventh semester). The results do not exactly corroborate the Bardovi-Halig and Dörnyei, nor the Niezgola and Röver study. She claims that learners’ learning environment does not always influence the learners’ awareness of errors, whereas proficiency level does.

The reported mismatch, however, is perhaps due to the difference in design of the instrument, and/or construct definition. In Takamiya’s study, a written survey, which involves reading, was used. Given the complexity of the Japanese writing system, the instrument might have inadvertently introduced a construct-irrelevant variable (reading ability). The reported difficulty conducting the study with low-proficiency students may be due to the difficulty in reading Japanese orthography. Another, more critical issue is the lack of precision in construct definition in Takamiya’s study. Consider the following two test items that, according to Takamiya, contain grammatical errors and not pragmatic ones:

友だちと話しています。^7

^7 English equivalents (provided by Takamiya):
While it is true that there are other linguistic forms preferred in the given situation above (“借りた方が”、“一時間しか勉強しなかったので”, respectively), the test utterances 図書館で借りる方がいいと思うよ and 昨日は一時間だけ勉強しました から, out of context, are both morpho-syntactically well-formed. These utterances are awkward only because they are deployed in situations they are not suited for, which is more of a pragmatic issue than a grammatical one. There are a few other items in Takayama’s study whose grammatical/pragmatic classification is questionable. Without clear definitions of what constitutes grammatical and pragmatic errors, Takamiya’s data need to be interpreted with care. Thus it remains to be seen, especially in the JSL/JFL context, if grammatical and pragmatic competence can be assessed separately, and if general proficiency and/or learning environment play a role in learners’ development in these two areas of competence.

You are talking to your friend.
Friend: I lost my textbook. Should I buy a new one?
I: You better borrow one from the library. We only have two more weeks before winter break.

You are talking at your teacher’s office.
Teacher: The result of the yesterday’s test was not good.
I: Sorry. I studied only for an hour. I will study more next time.
Bardovi-Halig (1999) points out that “research has not established that pragmatic competence is independent of grammatical competence. Although grammatical competence may not be a sufficient condition for pragmatic development, it may be a necessary condition.” Similar remarks are made by Hoffman-Hicks (1992): “linguistic competence is a necessary prerequisite to pragmatic competence but that it does not itself guarantee pragmatic competence.” However, even if grammatical competence precedes development of pragmatic competence, there are cases where “grammatical knowledge (a) is not put to pragmatic use; (b) enables non-target-like pragmatic use; and (c) [is] used in a way that is pragmalinguistically target-like but sociopragmatically non-target-like” (Kasper and Rose 2003). Bardovi-Harlig (2001) identifies four factors in determining L2 pragmatic competence: input, instruction, level of proficiency and length of stay, and first language and culture.

T. Takahashi and Beebe (1987) studied Japanese learners’ refusal strategies in English using a Discourse Completion Test (DCT), hypothesizing that pragmatic transfer is a function of learning context (ESL vs. EFL) and proficiency level. They conclude that pragmatic transfer is observed more with the EFL learners, and that higher proficiency ESL learners (graduate students in New York City) are more subject to pragmatic transfer. They refer to this finding as “their fluency gave them ‘the rope to hang themselves with’ — i.e., the control over English vocabulary to express Japanese sentiments” (p.153).

Kasper and Blum-Kulka (1993: 9) report that some learners exhibit a similar behavior, termed “waffle phenomenon,” characterized as “proficiency-dependent, being

---

8 See Section 2.2.2.1, following, for the notions of pragmalinguistics and sociopragmatics.
strongest at an intermediate stage when learners possess the linguistic means to say as much as they wish, yet at the same time feel more of a need to be explicit about their communicative goals and the reasoning behind them than more acculturated nonnative speakers do.” This waffling phenomenon is observed as lengthy responses in written responses to production questionnaires.

Kasper and Rose (2003: 190) speculate that “it appears that learners at different stages of pragmalinguistic development face different learning tasks: early learners have acquired the L2 grammatical means to express already existing pragmatic categories whereas later learners have to tease out the pragmatic meanings to which their now available L2 grammatical knowledge can be put.” As most studies are for learners of English, it remains to be seen if the same generalization applies to JSL/JFL learners.

2.2.2 Components of Pragmatic Knowledge

2.2.2.1. Interlanguage Pragmatics: Pragmalinguistics and Sociopragmatics

Interlanguage Pragmatics is defined as “the study of nonnative speakers’ use and acquisition of linguistic action patterns in a second language” (Kasper and Blum-Kulka 1993). A substantial body of research on interlanguage pragmatics has been conducted (Kasper and Rose 1999, 2003), with the majority of studies focuses on second language use (production of linguistic action, as well as pragmatic success and failure) rather than development (Kasper and Rose 1999).

Thomas (1983) proposes to divide “pragmatic failure” (inability to understand ‘what is meant by what is said’) into two subcategories: pragmalinguistics and
sociopragmatics. Pragmalinguistics has to do with “highly conventionalized usage” (p.91) of the language to convey communicative intentions. Pragmalinguistic failure occurs “when the pragmatic force mapped on to a linguistic token or structure is systematically different from that normally assigned to it by native speakers” (p.101). Sociopragmatics refers to “the social conditions placed on language in use” (p.99), including “size of imposition, cost/benefit, social distance, and relative rights and obligations” (p.104). Cohen (1996: 22-3) makes a similar distinction, using different terminology, sociolinguistic and sociocultural ability⁹. He defines these terms as follows:

Sociocultural ability refers to the respondents' skill at selecting speech act strategies which are appropriate given (1) the culture involved, (2) the age and sex of the speakers, (3) their social class and occupations, and (4) their roles and status in the interaction.

Sociolinguisitic ability refers to the respondents' skill at selecting appropriate linguistic forms to express the particular strategy used to realize the speech act (e.g. expression of regret in an apology, registration of a grievance in a complaint, specification of the objective of a request, or the refusal of an invitation). Sociolinguistic ability is the speakers' control over the actual language forms used to realize the speech act (e.g., "sorry" vs. "excuse me", "really sorry" vs. "very sorry"), as well as their control over register or formality of the utterance from most intimate to most formal language.

This distinction between pragmalinguistics and sociopragmatics has been proven useful (Cohen 2004; Rose and Kasper 2001), but the two are, of course, closely related and it is “not always obvious from performance data that indicate pragmatic failure which one is to blame: is the problem due to non-nativelike encoding or non-targetlike analysis of the social context of the utterance?” (Röver 2005: 4, added emphasis).

---

⁹ Cohen also uses the terms pragmalinguistics and sociopragmatics in other publications. See Cohen (2004, 2008).
The boundary between pragmalinguistics and sociopragmatics becomes even more blurry in the case of Japanese, which “has no neutral speech style; every speech act reflects choices involving the in-group/out-group membership, status, age, and gender of the participants” (Jorden and Walton 1987: 111). As such, utterances in Japanese include many linguistic forms that reflect such sociopragmatic notions. For example, Ide (1992) includes several personal pronouns\(^\text{10}\) used to express speakers’ masculinity/femininity, and perceived relationship with the interlocutor and/or a referent. Ide also reports 15 variants of a Japanese predicator meaning “go” in the context of “When do you go?”, ranging from most casual to more polite\(^\text{11}\). The pragmalinguistics (sociolinguistic ability) and sociopragmatics (sociocultural ability) distinction has been helpful in Second Language Acquisition (SLA) and testing research for ESL/EFL. However, in reality, they form a continuum and more research will be necessary to determine if such distinction may be useful in JSL/JFL testing and pedagogy contexts.

Jung (2002), in her review of research in second language pragmatic acquisition, identifies five aspects of competence that learners have to acquire in order to be pragmatically competent: (1) the ability to perform speech acts; (2) the ability to convey and interpret non-literal meanings (including inference and implicature); (3) the ability to

\(^{10}\) Such as 私（わたし、わたくし）、僕、俺、あたし、あたくし、君、お前、etc. Further, there is no requirement to express a grammatical subject in Japanese, and the first and second person pronouns are often left unmentioned when understood in context.

\(^{11}\) They are: 行く？ 行くの？ いらっしゃる？ 行きます？ 行くんですか？ 行きますか？ 行かれますか？ 行かれますか？ いらっしゃるの？ いらっしゃるんですか？ いらっしゃいますか？ おいでになりますか？ お出かけになりますか？ お出かけになられますか？ お出かけでいらっしゃいますか？ There are possibilities beyond what Ide listed (e.g., 行くのかい？ いらっしゃいます？, etc.)
perform politeness functions, (4) the ability to perform discourse functions (including sequencing of utterances, discourse markers, back-channeling\textsuperscript{12}, adjacent pairs, etc.); and (5) the ability to use cultural knowledge (cultural schemata or scripts, including a society’s “ways of speaking”).

The study this dissertation reports on focuses on the first three components, (1), (2) and (3) of Jung’s competencies, as much of the literatures on interlanguage pragmatics focus on these three, particularly in the area of speech acts (Roever 2006). This is also because the selection-type (online multiple-choice) test format employed for this study may not be the best tool to adequately address and measure discourse in terms of either functional knowledge or cultural knowledge, as these types of knowledge would require a more open-ended, highly interactive test format with human raters.

The following section reviews the three areas of investigation in interlanguage pragmatics pertinent to the current study: speech acts, non-literal meanings (as presented in routines), and politeness functions (as keigo, in particular).

2.2.2.2. Speech Acts

Speech acts can be defined as “functional units in communication” (Cohen 1996). Originally developed by Austin (1962) and further elaborated by Searle (1969), speech act theory has been influential in language pedagogy and interlanguage pragmatics. Kasper (2006) observes that “speech acts are the most widely examined object in interlanguage pragmatics, where they have been studied from a wide array of theoretical

\textsuperscript{12} For back-channeling (\textit{aizuchi}) and other discourse phenomena in Japanese, see Kita and Ide (2007).
perspectives and research methodologies.” See Geis (1995), Sadock (2004) and Schmidt and Richards (1980) for overviews of speech act theory. Only brief summaries of important works are provided below.

Austin (1962) notes that with regards to such utterances as “I do” in a wedding ceremony, or “I name this ship Queen Elizabeth” – what Austin called performative utterances – “to utter the sentence (in, of course, the appropriate circumstances) [...] is the performing of an action – it is not normally thought of as just saying something” (p.6-7). Not only is it true that “to say something is to do something,” often “by saying something or in saying something we are doing something” (p.12, original emphasis). Austin thus distinguishes three kinds of acts: an act of locution (uttering a certain sentence with certain sense and reference), illocutionary acts (with a certain conventional force such as informing, ordering, warning, undertaking, etc.), and perlocutionary acts (what we bring about or achieve by saying something) (p.109). Austin uses the following example to illustrate his point (p.102):

Act (A) or Locution
He said to me, ‘You can’t do that’.
Act (B) or Illocution
He protested against my doing it.
Act (C.a) or Perlocution
He pulled me up, checked me.
Act (C. b):
He stopped me, he brought me to my senses, &c.
He annoyed me.

Concerning Austin’s illocutionary force, Searle (1975: 60) discusses the case of indirect speech acts, “cases in which one illocutionary act is performed indirectly by way
of performing another.” The often-quoted example “Can you reach the salt?” illustrates the primary illocutionary act of requesting being accomplished by way of the secondary illocutionary act of questioning. Searle argues that such indirect illocutionary effect can be explained by “conditions of the felicitous performance of the speech acts […] — preparatory conditions, propositional content conditions, and sincerity conditions” (p.60), together with Gricean principles of cooperative conversation (p.61). For example, when performing a request, the felicity conditions are (1) a preparatory condition (H (hearer) is able to perform A (Action)), (2) a sincerity condition (S (speaker) wants H to do A), (3) propositional content condition (S predicts a future act A of H), and (4) an essential condition (Counts as an attempt by S to get H to do A)\(^\text{13}\). Seale also observes that what counts and what do not count as request, promise, etc. are conventionalized, and some linguistic forms are idiomatically used (i.e., an ability question for a request), while some lexical and syntactic choices work better than others (i.e., Can you do A? vs. Are you able to do A?).

Searle also distinguishes propositional indicators from illocutionary force indicator (or illocutionary-force-indicating devices). For example, in “I promise that I will come”, the indicator for propositional content is “I come”, and the indicator of illocutionary force is “I promise”. The illocutionary force identification device (IFID) “shows how the proposition is to be taken, or to put it another way, what illocutionary force the utterance is to have; that is, what illocutionary act the speaker is performing in the utterance of the sentence. IFIDs in English include at least: word order, stress,

\(^{13}\) For felicity conditions for other types of speech acts, see Searle (1969).
intonation contour, punctuation, the mood of the verb, and the so-called performative verbs” (ibid, p.30). Searle’s (Searle 1969) five classification categories for speech acts – directives, commissives, representatives, expressives, and declarations (Austin’s performative) – are probably the most well-known in the fields of linguistics and philosophy of language. The notional-functional framework draws on this classification, and establishes major functions (“Exchange factual and intellectual information,” “Exchange emotional and moral attitudes,” “suasion,” and “socializing”) with subfunctions and examples of linguistic patterns for each (van Ek and Trim 1991).

A certain level of the universality of speech act categories has also been proposed. Ochs (1996: 425-6) observes that “within the dimensions of social act meaning, acts such as greeting, asserting, […] and requesting goods and services appear across the world’s communities.” Coulmas (1981: 81) cautiously agrees with a caveat: “It seems to be a reasonable assumption that [there are] generic speech acts in every speech community. I would even go so far as to venture the hypothesis that every language provides a stock of conventionalized means for fulfilling these functions. However, the way in which a cultural context imposes restrictions on a kind of linguistic activity can often be seen to vary in a specific way from one socio-cultural system to another”. The differences in speech act realization patterns in different cultures have been investigated in both cross-cultural and interlanguage-pragmatics perspectives, focusing on the notion of pragmatic transfer, i.e., “falling back on their L1 […] norms and conventions in realizing speech acts in the target language” (Takahashi 1996: 189).
In interlanguage pragmatics, the Cross-Cultural Speech Act Realization Patterns (CCSARP) project (Blum-Kulka and Olshtain 1984) is perhaps the best-known study on speech acts, for its scale of investigation, research instrument, and the coding system developed for the study. The CCSARP project focuses on two speech acts, requests and apologies, and investigates inter-cultural and intra-cultural variability between native and non-native speakers’ speech act realization patterns, using a Discourse Completion Test (DCT) as the research instrument. Here is an example of an item used to elicit a request:

(p.198):

At a student’s apartment
Larry, John’s roommate, had a party the night before and left the kitchen in a mess.

John: Larry, Ellen and Tom are coming for dinner tonight and I’ll have to start cooking soon; __________________________________________________
Larry: OK, I’ll have a go at it right away.

There are eight items for each speech act, with various social parameters of ± social distance and ± dominance. The test has been administered in seven languages (Australian English, American English, British English, Canadian French, Danish, German, and Hebrew), to 200 natives and 200 non-natives (in their 2nd and 3rd years of study) for each language. The DCT has been used in numerous studies since, and there are some variants of DCTs in different formats and modalities. McNamara and Röver (2006) reviews the DCT’s strengths and weaknesses as a data collection tool.

Olshtain and Cohen (1983) describe the analysis procedure for the case of apology in the CCSARP project. They identify five strategies for apology, and semantic
formulae: a semantic formula is “a word, phrase, or sentence which meets a particular semantic criterion or strategy” (p.20), that is typically used to realize these strategies. This concept of semantic formula is akin to Searle’s illocutionary force identification device (IFID), and has been applied to many speech act studies. In order to analyze various strategies used for apologies, they proposed the notion of a speech act set, and identified semantic formula typically used to realize each strategy: (1) an expression of apology (“I’m sorry”, “I apologize”, “Forgive me”, etc.); (2) an explanation or account of the situation; (3) an acknowledgment of responsibility; (4) an offer of repair; and (5) a promise of forbearance (“It won’t happen again”, etc.). Other strategies (downgraders, intensifiers, etc.) have also been identified. The nonnatives’ realization patterns of these five strategies are compared with those of natives in terms of the frequency, order of appearance, and linguistic forms used for each strategy.

Limitations in the traditional Searlean speech act theory, which assumes that indirect speech acts are determinable by idiomatic conventions and by inference at the utterance-level\(^{14}\), have been pointed out in the literature. Some are mentioned in Hatch (1992), but the central limitation is that there is no absolute one-to-one correspondence between function and utterance, and speech acts are indeterminable in nature (Tanaka 1996; Geis 1995). Geis asserts that "any attempt to map Searlean primary acts into the utterances that comprise multiple-turn conversational sequence cannot succeed" (p.26) as "there are inevitable, unresolvable uncertainties in the assignment of primary illocutionary force to individual utterances" (p.21). Yotsukura (2005: 343) reports that

\(^{14}\) Searle (1969: 16) states, “the unit of linguistic communication is […] production or issuance of the symbol or word or sentence in the performance of speech acts”.

“recent research on Japanese conversation […] has moved beyond the traditional speech act theory […] to examine larger units of discourse. This has had important theoretical consequences, for it has been demonstrated that social actions such as requests and invitations are actually negotiated over a series of turns”. Kasper (2006) also calls for a discursive pragmatics. However, Cohen (2008: 214) maintains that “I can […] see for the sake of foreign language pedagogy the benefits of continuing to focus on the more traditional speech act research”. It is this call of Cohen’s for applying the traditional approach to speech act realization patterns that has been adopted in this study, and testing the ability to perform discourse functions is beyond the scope of this dissertation study. Testing of discursive pragmatics, employing the approach of conversation analysis, would require innovative test methods, with more interactive and open-ended performance elicitation and raters trained in conversation analysis.

In the traditional, utterance-level speech act research in interlanguage pragmatics, particularly important is the notion of negative transfer. Negative transfer is characterized as “the projection of first language-based sociopragmatic and pragmalinguistic knowledge onto second language contexts where such projections result in perceptions and behaviors different from those of second language users” (Maeshiba et al. 1995: 155). Learners’ negative transfer is observed in their choice of lexical and structural patterns to realize the various speech act strategies, and is of particular interest in language testing, as “it may allow generations of multiple-choice items with ‘attractive’ distractors, or produce items that invite transfer by less proficient learners” (Röver 2005: 7). Examples of negative transfer by JFL/JSL learners abound, as the products of making
overgeneralizations of certain moves in seemingly similar target situations. Tateyama (2007) reports cases among 46 second-year (4th semester) learners, in which they often misunderstood their teacher’s simple yes-no questions such as *yomemasu ka?* ‘are you able to read?’ and *simasen ka?* ‘don’t/won’t [you] do [it]?’ as requests. This is most likely a negative transfer from English, where ability questions (Can you …?) and intention questions (Will/won’t you …?) are conventionally associated with requests. Yotsukura (2003) contains a few anecdotes of negative transfer. One is served up by a learner of Japanese uttering “*zenbu ikaga desu ka?*” in his role play as a waiter in a restaurant context, based on his literal translation from “How’s everything?”. Unlike in the U.S., a waiter or waitress in Japan usually does not come back to the table to ask how things are, thus making the Japanese version syntactically accurate and pragmatically puzzling to the natives. Another anecdote concerns what a bilingual office assistant said to her supervisor — “*Katte kite agemasyoo ka?*” (a direct translation for “Shall I go and buy [them] for you”) — as an offer to go and buy stamps. She was reportedly immediately corrected by her mother, who was present in the office, as it is inappropriate in the Japanese context to use *-te agemasyoo ka?* form to a superior when merely performing one’s job duties. Using *-te agemasyoo ka?* makes it sound as if the act is a favor to her supervisor.

Unger (2001: 353) stresses that a foreign language is not a translation of one’s native language, and encourage teachers to help learners avert such negative transfers: “The greater the cultural difference, the more likely that literal translation will result in misunderstanding and a breakdown of communication. […] For example, the Japanese phrase *nomitai desu ka*, a direct translation of the English ‘Do you want [something] to
drink?’ is not an invitation in Japanese and is socially inappropriate even though it is syntactically possible. Teaching the structural patterns is not enough; as soon as sentences of this type are introduced, the teacher must take time to explain how they are used in Japanese discourse.” Oishi (1996) describes in detail how queries about desire manifest themselves differently in different cultures.

There are abundant studies on speech acts in interlanguage pragmatics involving various languages (see Barron (2003) for lists of cross-sectional and longitudinal studies). There are studies from the perspective of cross-cultural pragmatics, comparing Japanese and English, on apologies (Barnlund and Yoshioka 1990; Maeshiba et al. 1995; Ito 1998), compliments (Daikuhara 1986), and the use of excuses in refusals (Nishimura 2007). There are also many studies on speech act realization patterns from the perspective of interlanguage pragmatics that investigate the language use of Japanese learners of English. The speech acts investigated in this perspective include advice (Matsumura 2001), chastisement and disagreement (Beebe and Takahashi 1989), complaint (Boxter 1993; Rinnert, Nogami, and Iwai 2006), refusals (Beebe, Takahashi, and Uliss-Weltz 1990; Robinson 1996; Takahashi and Beebe 1987; Gass and Houck 1999), correction (Takahashi and Beebe 1993), and requests (Achiba 2003; Niki and Tajika 1994; Rinnert and Kobayashi 1999; Takahashi and DuFon 1989; Takahashi 1996). Yoshinaga et al. (1992) provides an extensive bibliography on Japanese pragmatics prior to the early 1990s, and Cornwell et al. (2007) provides a review on studies in interlanguage pragmatics published in Japan between 2000 and 2006.
There are smaller numbers of studies in interlanguage pragmatics that investigate language use by English-native adult JSL/JFL learners. Longitudinal studies are almost nonexistent, except for Cohen’s (1997) detailed account of his own in-class and outside-class experience as a learner of Japanese during an intensive language program. There are several cross-sectional studies involving JSL/JFL learners’ speech act production, which are briefly summarized in the following section.

For apologies, Tamanaha (2003) used three role-plays with different levels of gravity of offence, while keeping the power and social distance variables consistent (namely, between friends). Utilizing a speech act set classification similar to that of Olshtain and Cohen, she confirmed the findings in previous studies that Americans preferred an “explanation” or rational strategy, while Japanese more frequently chose a more emotional, “offer of repair” strategy. As for American learners of Japanese, she compared the performance by thirteen intermediate learners (three years of formal instruction who had not lived in Japan for more than a year) with eleven advanced learners (more than three years of formal instruction plus more than two years of living experience in Japan), and found that negative transfer was suspected to be present in both groups, that intermediate learners performed poorly, and that the advanced learners approximated those produced by the natives. More specifically, in terms of the linguistic forms used by the subjects, the intermediate learners produced more formal IFIDs for the core apology strategy such as sumimasen, gomennasai, while the Japanese subjects used gomen ne most frequently (only 2 intermediate learners used gomen ne), and did not use other strategies very often. The intermediate learners also produced fewer prefaces (i.e.,
tyotto hanasi ga aru n yakedo… ‘I have something to tell you but (is now a good time)?’; zitu wa ‘As a matter of fact,’ etc.), often displayed English-like account-then-apology sequence, and used the –te simau (completive aspect marker) form much less frequently than the other groups. The use of the –te simau structure appears to be a critical linguistic tool to appropriately express an unintended course of action, as this pattern is also reported as one of the specific linguistic gains made in Cohen and Ishihara’s study (Cohen and Ishihara 2005) of the development of web-based instructional units for five speech acts in Japanese, including apologies.

In investigating reactions to compliments, Saito and Beecken (1997) examined data obtained from 10 American learners of Japanese with at least 3 ½ years of study through 10 role-plays. The role-plays combine five different topics (compliments on their clothes, term papers, the house they live in, food they cooked, and winning a tennis match) with two interlocutors (a friend and a professor). The learners’ reaction data was classified into three types (positive, negative, and avoidance), and was compared with native Japanese and American English data. The “positive” compliment responses included Thank you (Arigatoo), I’m glad you like it (Soo itte itadaite kooei desu); the “negative” responses included No, I’m still not good, you know (iya, mada mada desu yo); and the “avoidance” responses included Really? (Hontoo?) Well, it’s OK (Soo? Maamaa ne). The results showed that the learners were able to use the negative reactions (nonexistent among English-speaking American subjects), and showed nonuse of avoidance reactions, both of which are possibly instruction effects. They also found that the learners’ negative reactions were qualitatively different from those of the native
Japanese, in that the learners used simple denials with no follow-up strategies (such as return, mitigation, affirmative explanations, etc.). Baba (1999) also studies compliment responses made by seventeen American learners of Japanese (self-reported as ACTFL Intermediate and Advanced), in data obtained from conversations with a female friend (the conversation leader). Four target topics (family-external, family-internal; self-external, and self-internal) were identified. “External” refers to compliments on cloths, possessions, looks, etc., and “internal” refers to personalities, abilities, etc. It was found that there was no statistically significant difference in reaction patterns for self-external, but the difference was most significant in the use of the negative strategy in family-external context. More specifically, she found that (1) in general, the learners used more negative strategies while the Japanese natives seemed to prefer self-denigration in a joking manner; (2) the learners showed resistance to using negative strategies in family-related compliments; and (3) the learners did not use “thank you” at all, and some used self-mockery in response to self-external compliments, while the Japanese natives tended to agree or appreciate compliments on their own clothes or fashion tastes in conversation with a friend.

In the case of complaints, Tamanaha (2003) again used three role-plays with different levels of gravity of the offence, while keeping the power and social distance variables consistent (again between friends), for the same subjects who participated in her apology study (intermediate and advanced groups). She reports that the learners had more difficulty in this speech act than with apologies, using more direct strategies than the
Japanese natives, who used “hints” and “annoyance” strategies more often. The intermediate learners also reportedly displayed many grammatical errors.

As for refusals, Ikoma and Shimura (1993) used 12 Discourse Completion Tests (DCTs) and analyzed data from 10 learners of Japanese at their 4th-year level of study. The 12 DCTs included 4 refusal patterns (refusal of request, invitation, offer, and suggestion) with 3 different roles (interlocutor with lower, equal, and higher social status). They report no pragmatic transfer in terms of the order of strategies (apology-then-excuse). They did observe, however, three cases of negative transfer among the learners which might cause miscommunication: (1) the infrequent proposing of alternatives in the refusals of requests by a superior, (2) the excessive use of formal-sounding kekkoo desu ‘no, thank you’ when refusing cake offered by a friend, and (3) the excessive use of direct refusals (whereas the Japanese used minor sentences as Nitiyoo wa tyotto… ‘Sunday is a bit… (inconvenient)’ more frequently).

In the case of requests, Tateyama (2007) examined the effects of instruction in teaching Japanese requests to JFL learners in their second year (4th semester) of study. The 3-week intervention in the middle of semester included consciousness raising activities (watching video clips, collecting more conversation data and analysis of the conversation) and additional conversation practice. The results showed that students in both the control group (N=22) and treatment group (N=24) improved their performances in the post-test, moving from direct, teacher-talk-like directives (V-te. / V-te kudasai) to more sophisticated ones (V-te itadakenai/kudasaranai desyoo ka). But there was no statistically significant difference between the two groups.
Cohen and Ishihara (2005) developed self-access, web-based instructional units for five speech acts in Japanese: apologies, compliments/responses to compliments, requests, refusals, and thanks. They list the following as a few of the research-based, empirical underpinnings Japanese speech acts in contrast with English: (1) directness/indirectness in requests as being largely dependent on the relative status of the interlocutor; (2) variation relative to age, status, and familiarity with the interlocutor rather than the intensity of the speech act (e.g., severity of imposition); (3) the selection of reasons to use in a refusal in light of who the interlocutor is; and (4) a tendency to deflect or reject a compliment (p. 4). In addition to these strategies, learners also need to know semantic formula — patterned, “routinized phrases used regularly to perform a variety of functions” (p.2).

The linguistic patterns typically used to realize the speech act intentions are important for learners to learn, as attested in the studies reviewed so far, and thus they are focus of this dissertation project.

2.2.2.3. Routines

The term “routine” or “formula” has been used in various contexts in the literature, referring to different degree of conventionality and indirectness of linguistic forms and strategies involved (Bardovi-Harlig 2006; Kasper 1995). Here, routines refer to conversational routines, defined as “highly conventionalized prepatterned expressions whose occurrence is tied to more or less standardized communication situations” (Coulmas 1981: 2-3). This is what Bardovi-Harlig (2006) refers to as the target formula,
and what Kecskés (2000) refers to as Situation-Bound Utterances\textsuperscript{15}. Examples for routines in English include *Hi, how are you?*, *What’s up?* *So long, Can I take a message?*, etc.\textsuperscript{16} The discussion in this section concerns conversational routines, and excludes other types of such routines as interactional routines, such as Initiation-Response-Follow-up routines (Ohta 1999) or classroom routines, such as taking attendance (Kanagy 1999).

Japanese is particularly rich in routine expressions, and there are many so-called “how-to” books in Japan, written for native consumption, on etiquette and protocols that involves routine expressions (for example, Sanseido 2002). The importance of routines for JSL and JFL learners has been stressed in the literature. Jorden with Noda (1987: 309-10, original emphasis) state in their Japanese language textbook:

\begin{quote}
… if we carefully memorize exchanges that are regularly used in specified situations, we are guaranteed to ‘say the right thing.’ But […] if we resist ‘clichés’ and depend on an original creation of our own making […] our chances of saying what is appropriate, or even what is comprehensible to the native speaker, are very slim. […] Obviously foreigners are excused for many of their errors, but the more one depends natural, comfortable communication [, or worse.] may be interpreted by the native speaker not as linguistic ignorance but rather as a totally different message — including, on occasion, social insult.
\end{quote}

These cautionary remarks are echoed by Coulmas (1981: 90), with the added note that there are cultural differences in the attitude toward the use of routines: “[in Japan,] formulaic utterances are not considered as hackneyed expressions lacking in any real content, as is often the case in Western cultures. More important than originality of

\textsuperscript{15} Following Bardovi-Harlig (2006), routines are differentiated from more flexible and productive types, including those known as semantic formulae (in the previous Speech Acts section).

\textsuperscript{16} Japanese examples of routines include いただきます, はじめまして, 乾杯! etc. See Appendix B for a list of routine expressions developed for this dissertation project.
expression is to say the right thing at the right time. [...] Indeed, linguistic etiquette requires the speaker of Japanese to make extensive usage of routines, often leaving little room for variation.” The tension between following social conventions and originality is also discussed in Ide and Yoshida (1999).

Among the numerous routine expressions, one that draws scholars’ attention is the ubiquitous sumimasen. Coulmas analyzes Japanese routines for gratitude and apology expressions, including sumimasen, and argues how these two seemingly very different speech acts are similar, and how Japanese sumimasen can be understood as both an expression of gratitude and apology (as well as a conversation opener and attention getter) by using the notion of indebtedness. R. Ide (1998) concurs and advances Coulmas’ ideas, and identifies seven functions of sumimasen from the ethnographic perspective, using recorded public discourse at a clinic as her primary data. The seven functions are: (1) as a sincere apology, (2) as quasi-thanks and apology, (3) as a request marker, (4) as an attention-getting device, (5) as a leave-taking device, (6) as an affirmative/confirmative response, and (7) in reciprocal exchange of agreement.

Kumatoridani (1999), utilizing a questionnaire and data from live conversations, focuses on cases when sumimasen and arigatoo are not interchangeable, and the sequential preference of the two. He concludes that the gratitude expression and the apology expression occur as a result of the empathy operation, that this alternation is subject to the affect and role-relationship constraints, and that sumimasen is used for local management and arigatoo for closing the exchange expressing gratitude.
Research on how learners gain competence in using and understanding routine expressions is scarce, both for ESL/EFL and JSL/JFL learners. Kecskés (2000) reports that, in a Dialogue Interpretation Task, when 88 ESL learners in the U.S. from ten different countries were asked to interpret such utterances as OK, shoot (Go ahead), Piece of cake (Easy), and Get out of here (Don’t fool me), they “relied on the literal meaning and compositional structure of the expressions which led to misunderstandings.” Another interesting finding from this study is from its Problem Solving Task, when the subjects were asked to answer the following question:

There is a TV show on. The anchor pauses the program for a commercial. S/he wants the audience to continue watching the program after the break. What does s/he say?

While “native speakers came up with the most frequently used expressions such as Stay tuned, We’ll be right back, We’ll have to take a break, Don’t go away” etc., the ESL subjects “having spent less than two years in the U.S. […] did very poorly on this task” and came up with responses such as Keep your channel (L1 Japanese)17, When we come back we will an action… (L1 Russian). This task “was not a problem for those students who have spent more than two years in the target language country” (pp. 153-154). This suggests that amount of input, exposure, and/or the learning environment may contribute to the learning of some routine expressions, although there are no detailed statistical data of the two groups provided in the article.

17 Most likely the direct translation from a common Japanese phrase チャンネルはそのまま.
Tateyama (2001) investigates the effect of explicit and implicit instruction in the use of *sumimasen* on twenty-seven beginning JFL learners. In her study, the treatment group received explicit instructions on three functions of *sumimasen* (attention-getting, expressing gratitude, and as an apology indicator), among other routine expressions (*gomen-nasai, arigatoo*, etc.). Results on a multiple-choice test and role-plays “indicated no significant differences between the two groups”, which is “inconsistent with the result of the pilot study […] which demonstrated an advantage for the explicit group after a single treatment of only 25 minutes” (p. 219).

More research on the process of learning routine expressions, and on factors that might influence their learning (exposure/input, types of instruction, etc.) are needed.

2.2.2.4. Speech Styles

Speech styles, which have traditionally been classified under sociopragmatics, are incorporated in the scope of this dissertation, as Japanese contains rich morphological and lexical forms to indicate various speech styles. As mentioned earlier, “Anglo-European learners of Japanese have to learn sociopragmatic categories such as *uchi* (in group) and *soto* (out group), sociolinguistic distinctions such as different speech styles, and pragmalinguistic knowledge such as donative verbs18” (Kasper and Rose 2003).

There are many categories of speech styles that index speakers’ gender, social roles and hierarchy, distance, etc. Jorden and Noda (1987) list various binarily opposed

---

18 Her use of the term “sociolinguistic” to refer to something in between sociopragmatics and pragmalinguistics may be an indication of the in-between nature of speech style markers in Japanese, and of the sociopragmatic-pragmalinguistic continuum mentioned in Section 2.3.
speech styles: careful vs. casual, distal vs. direct, polite vs. plain, gentle vs. blunt, masculine vs. feminine. Of these, polite styles, or *keigo*, have been identified as important for Japanese learners to learn (Cohen and Ishihara 2005, among many others). Polite styles include the choice and apt deployment of, e.g., direct vs. distal markers (e.g., direct *kaku* vs. distal *kakimasu* ‘write’) of the speaker-addressee relationship, plain vs. polite markers (e.g., plain *kaku* vs. honorific-polite *okaki ni naru* and humble-polite *okaki suru* ‘write’) of speaker-referent relationship, or combinations thereof (Wetzel 2001).

The importance of teaching and learning different speech styles is supported by many educators. Wetzel (2001) maintains that “skillful use of the Japanese language – *keigo* in particular – depends on a set of cultural categories and associated practices that should be integral to language training at all levels.” Konomi (1997) states, “Japanese is a language with no socially neutral expressions. Even to say ‘thank you’ properly in the range of situations one encounters in everyday life requires that the learner understand the different grammatical forms for indicating gratitude.” The treatment of different speech styles varies among major textbook series, however, in terms of the timing of introduction and sequencing of different speech style patterns (Itomitsu 2005).

There are difference in terms of linguistic politeness used by Americans and Japanese. Hill et al. (1986) conducted a large-scale quantitative study, involving 525 Japanese and 490 American subjects, to investigate on different polite levels in request. They used questionnaires and elicited 20 different ways of requesting a pen to 20 different categories of people, and asked the subjects to rank them using a 5-point scale, from most uninhibited to most careful. They found that (1) the Japanese responses are
more tightly clustered together, and that (2) the Japanese request forms show greater range span on the 5-point scale, while the distribution shows two larger groupings, reflecting the *uchi* and *soto* distinction.

Understanding how different speech levels are perceived among natives seems to be one of the most difficult aspects of learning Japanese for Americans. Cook (2001) investigated learners’ perception of nonnative speakers’ use of various speech styles in three scripted job interviews in Japanese, given as part of the listening comprehension exam at the University of Hawaii at Manoa. To the instructors’ surprise, more than 80% of the 120 JFL learners in their second year of study chose Applicant A, who says she has all the required qualifications, while the same applicant A was regarded by the instructors as by far the worst ("out of question") because of the impolite speech style used.

Speech Style is included in this study, as the grammatical forms to express various levels of speech styles are deemed critical in learning Japanese as a foreign language.

2.3. Tests of Interlanguage Pragmatics

In many studies in interlanguage pragmatics reviewed in the previous section, learners’ pragmatic ability has been observed in semi-spontaneous speech, role-play, DCT, questionnaire, etc. The observed data are typically categorized and rated by one or more human raters. Compared with the growing number studies in interlanguage pragmatics, however, there are only a small number of studies on test development in interlanguage pragmatics for learners of English, and much less for JSL/JFL learners.
McNamara and Röver (2006: 54) speculate that the possible reason for such imbalance as follows: “because of its highly contextualized nature, assessment of pragmatics leads to significant tension between the construction of authentic assessment tasks and practicality: Social context must be established and learner responses are often productive, so simulations of real-world situations and scoring by human raters would be ideal, but they are also very costly. It is indicative of these difficulties that only few tests are available in this area.” Probably the first comprehensive attempt to embark on testing pragmatics (sociopragmatics) was by Hudson and his colleagues at University of Hawai‘i at Manoa (Hudson, Detmer, and Brown 1992, 1995).

Hudson et al. (1995) used three types of test (indirect measures, semi-direct measures, and self-assessment measures) to measure interlanguage pragmatics, each with two test format, resulting in six measures (written DCT, an oral (language lab) DCT, a multiple-choice DCT, a role-play, self-assessment measures for the DCTs and the role-play). Each format has its own advantages and disadvantages (Brown 2001; McNamara and Röver 2006).

McNamara and Röver (2006: 55-6) observe that “because of the close connection between pragmalinguistics and sociopragmatics, it is difficult to design a test that tests pragmalinguistics to the exclusion of sociopragmatics or vice versa. Still, most tests of pragmatics have focused on one or the other aspect of pragmatic competence and can be classified as oriented more toward the sociopragmatic end (testing appropriateness in the context of social relationships) or the pragmalinguistic end (testing linguistic forms
necessary to achieve communicative ends).” Studies on sociopragmatic and pragmalinguistic tests are reviewed separately in the following section.

2.3.1. Tests of Sociopragmatics

As mentioned earlier, Hudson et al. (1992, 1995) designed a test battery for assessing sociolinguistic competence to assessing Japanese ESL learners’ ability to produce and recognize appropriate speech act patterns in requests, apologies, and refusals. Their battery consisted of six sections: written and oral DCTs, a multiple-choice DCT, a role-play, and two self-assessments. During the test item production stage, Hudson et al. systematically varied three variables – power, social distance, and degree of imposition – the ones that are deemed important (and universal) variables by Brown and Levinson’s politeness theory. For their multiple-choice DCT, they conducted a survey to collect frequently observed native speaker response patterns and used them as correct answers, and dissimilar non-native speaker responses as distractors. To assess test-taker performance on the oral DCTs and the role-plays, Hudson et al. trained English-native raters on the use of analytic 5-point scale. Hudson et al. (1995) contains all test items used in the study. Hudson (2001) reports studies on their test administration to 25 Japanese ESL learners in Honolulu on written DCT, oral DCT in a language lab, and role-play. They found a method effect (face-to-face role-play resulted in higher rating than DCT in the language lab) and that apology tasks seemed easier than the others. Brown (2001) reports findings from the test administered by Yoshitake-Strain to 25 Japanese EFL learners in Tokyo, reporting relatively low reliability estimates for the
multiple-choice and the written DCTs (K-R21 of 0.6050, 0.4960, respectively). This may have resulted in lack of general correlations among the six sections of the test (only three out of fifteen correlation coefficients were significant). Brown speculates that the results may be due to the relatively homogenous nature of the test participant group.

Yamashita (1996) developed a test battery for learners of Japanese, by translating Hudson et al.’s original English test into Japanese, with minor revisions. The test was administered to 47 English-speaking JFL learners. She employed various statistic analyses, and reports high reliability among the three test formats except for the multiple-choice DCT (overall Cronbach alpha of 0.4652), high inter- and intra-rater reliability, high correlation coefficients among the six sections of the test (13 out of 15 are statistically significant). She also found that there was a difference among Elementary, Intermediate, and Advanced groups in DCT, role-play, and the language lab DCT, and that the participants’ length of exposure to the target culture appears to have effects on role-play and the language lab DCT performance. Brown (2001) compares the findings from Yamashita and Yoshitake’s studies side-by-side.

Low reliability in multiple-choice DCTs, reported in both Brown’s and Yamashita’s studies, is particularly concerning. There are a few explanations that can be made to this point. One possible factor is that the multiple-choice DCT was in the 3-option format, making the chance level of selecting the correct answer as 33%. Another possibility for Yamashita’s data may lie in the translation process. Yamashita reports that in her pilot study there were some comments about the situations included in the test items as not being relevant or appropriate in the Japanese context (borrowing a pen from
your boss at a meeting, renting out a room in a big house, etc. may not occur frequently in Japanese context). Even when the tasks are plausible in the Japanese context, the difficulty of translating the options still remains. Here is an example of the pair of multiple-choice DCT items, first in the original English (Hudson, Detmer, and Brown 1995: 109), followed by the Japanese translation (Yamashita 1996: 176):¹⁹

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.

a: Excuse me, can I use an extra pen?
b: Oh, I’d like to take some notes, but it seems that I have no pen with me.
c: Excuse me, but do you have an extra pen I could borrow?

あなたは大きな会社の小さな部で働いています。今その部の会議に出席していますが、ノートを取るためにペンを借りる必要があります。となりに坐っている部長が一本余分に持っているかもしれません。

a: すみませんが、ペンをお借りできますか。
b: あ、メモを取りたいんですが、書くものがないんですよ。
c: すみませんが、お借りできるような、余分なペンをお持ちですか。

Specific lexical choices (借りる ‘borrow’ instead of ‘use’ in Option A, 書くもの ‘a thing to write with’ instead of ‘pen’ in Option B, etc.) are not direct, literal translations from the original English. There are many other linguistic issues (politeness level and other speech style issues, use of the extended predicate んです, etc.) that complicate the issue even further. Unger’s (2001) point that a foreign language is not a translation of one’s native language is particularly relevant here. Yamashita’s JSL/JFL items are based

¹⁹ The original Japanese test items in Yamashita contain furigana (reading over the Chinese characters).
on data obtained from English speakers, and not on a survey or observation of prototypical native communication patterns in Japanese culture, or on a consensus on acceptability judgment by the Japanese natives. The same speech act in the given situation may not even occur in the Japanese context. Furthermore, the very nature of assessing sociopragmatic knowledge involves acceptability judgments, which actually indicate a wide range of acceptability among the natives. In the above English multiple-choice example, the “correct” answer is most likely option C, but option A may also be acceptable (keyable) to some people. In the Japanese example, option A now seems to be more likely as the best option, more so than option C. As McNamara and Röver (2006: 57) observe, the central problem of sociopragmatically-oriented tests is that “judgments of what is and what is not appropriate differ widely among native speakers and are probably more a function of personality and social background variables than of language knowledge.” Creating reliable multiple-choice items to test sociopragmatic knowledge has been proven to be of major challenge. The multiple-choice format may be suitable for testing the more linguistic ‘end’ of the pragmatics continuum, where there is generally tighter consensus among target natives.

---

20 Yamashita (1996: 63) reports comments by native Japanese speakers participated in a pilot study: “it is not natural for the company worker (a) not to bring a pen to a meeting and (b) then to ask to borrow a pen from your superior or boss.”

21 No indications as to which option was intended as the key option are provided in both of the studies.
2.3.2. Tests of Pragmalinguistics

Previous studies of tests of pragmalinguistics are small in number as well. There are a few tests developed for ESL/EFL learners.

Röver (2005; 2006) developed a Web-based test of ESL pragmalinguistics, designed to assess learners’ knowledge of routine formulae, speech acts, and implicature. Both multiple-choice as well as short-answer formats were used. Here are his examples (2006: 238-9):

Implicature:
Jack is talking to his housemate Sarah about another housemate, Frank.
Jack: ‘Do you know where Frank is, Sarah?’
Sarah: ‘Well, I heard music from his room earlier.’
What does Sarah probably mean?
1. Frank forgot to turn the music off.
2. Frank’s loud music bothers Sarah.
3. Frank is probably in his room.
4. Sarah doesn’t know where Frank is.

Routine:
Jack was just introduced to Jamal by a friend. They’re shaking hands.
What would Jack probably say?
1. ‘Nice to meet you.’
2. ‘Good to run into you.’
3. ‘Happy to find you.’
4. ‘Glad to see you.’

Speech Act 22:

---

22 As for the Speech Act section, data from this short-answer format was chosen and analyzed in detail in Röver (2005) and Roever (2006). Röver (2005) also contains multiple-choice version of Speech Act section administered to smaller group of participants (p. 133-138 lists the 12 multiple-choice test items). No detailed statistics on the reliability of the multiple-choice items are provided. The multiple-choice version also seems to involve more reading and logical thinking ability.
Ella borrowed a recent copy of Time Magazine from her friend Sean but she accidentally spilled a cup of coffee all over it. She is returning the magazine to Sean.

Ella: .................................................................
Sean: ‘No, don’t worry about replacing it; I read it already.’

Röver conducted a number of quantitative and qualitative validation studies with 267 learners of English with various L1 backgrounds, and found high reliability in each section of the test (Cronbach alpha of 0.73 or better) as well as high inter- and intra-rater reliability for the speech act section (alpha = 0.96 and 0.87, respectively). Factor analysis and correlation studies revealed that the three subcomponents of pragmalinguistic knowledge can be considered interrelated, with each having some unique variance, too. He also found that learners’ knowledge of speech acts increased with proficiency, as did their knowledge of implicature, but the knowledge of routines was strongly dependent on L2 exposure.

Liu (2006) developed a test of requests for Chinese EFL learners in China. The test consisted of a multiple-choice DCT, a written DCT, and a self-assessment. Instead of basing the items on the test developed by the previous studies, Liu painstakingly developed his own items by going through a series of processes (exemplar generation, likelihood investigation, metapragmatic assessment by English natives, and a pilot test). The final test was administered to 89 Chinese EFL learners, consisting of high-level (in their 3rd year, N=31) and low-level (1st year, N=58) students. The resulting data were compared with the TOEFL scores of the same groups as well. The results showed that the internal consistency of each section was high (alpha = 0.8647 or better), including the multiple-choice DCT. The fact that the multiple-choice DCT achieved high reliability is
particularly commendable, and Liu attributes this finding to his unique test item generation process, as test situations are based on Chinese context and therefore closely related to the participants’ lives. The results also showed the three tests significantly correlated with one another, and factor analysis confirmed that these three tests might tap a similar construct, i.e., the tests seem to be testing the same underlying knowledge. The group difference between the high-level and the low-level was significant in the TOEFL score and the multiple-choice DCT (might be a method effect), but not in the written DCT and self-assessment. Liu concludes that the participants with higher grammatical proficiency (as indicated by the TOEFL score) did not necessarily possess higher concomitant pragmatic competence. Liu’s study is targeted at Chinese EFL learners, and it remains to be seen if similar results can be obtained by administering the same test to a group with a different L1. The appropriateness of generating the test contexts based on Chinese learners’ experience in China (their self-reported ten most recently occurring events involving request) is debatable if the similar situations are not equally frequently encountered by the learners in the target language use situations. Still, the high reliability of the tests is encouraging, and the generalizability of the findings and the test item generation methods to other types of speech acts will be of great future research interest.

2.3.3. Relationship among Components of Pragmatic Knowledge

In addition the relationship between the grammatical and pragmatic knowledge reviewed in Section 2.3.1, a few studies investigated the relationships among subcomponents of the pragmatic knowledge.
Bachman and Palmer (1982) investigated test results from 116 adult ESL students. The tests were designed to test three components of communicative competence — linguistic competence, pragmatic competence and sociolinguistic competence. A multitrait-multimethod design was used, in which each of the three hypothesized traits was tested using four methods: an oral interview, a writing sample, a multiple-choice test and a self-rating. They defined pragmatic competence as encompassing the areas of vocabulary, cohesion, and organization, and sociolinguistic competence as sensitivity to register, nativeness, and non-literal language (operationalized as “cultural reference”) (p.451). Examples from the 119 5-choice items (item stimuli included pictures, short sentences, dialogues and short paragraphs) includes the following (p.454):

#1: Sociolinguistic item (nativeness):
I went into town and _____
(1) got lost.
(2) got off the way.
(3) got off my way.
(4) lost the path.
(5) lost my path.

#2: Sociolinguistic item (cultural stereotypes):
"He's pretty absent-minded, don't you think?"
"He sure is. And he's not even a _____.”
(1) hockey player
(2) college professor
(3) sales person
(4) big executive
(5) secretary

The results from confirmatory factor analysis (a statistical analysis to examine whether the observed factors match the hypothesized test constructs) indicated that the model which best fits the data included a higher-order “general factor” and two specific trait
factors — grammatical/pragmatic competence and sociolinguistic competence. Bachman and Palmer argue “while we had originally posited separate components of grammatical and pragmatic competence, it is perhaps not surprising that these two appear to cluster together and are distinct from sociolinguistic competence. It may well be that grammar and vocabulary underlie, or are necessary for cohesion and organization, and that these are all functions of the organizational aspects of language, both structural and semantic. The sociolinguistic subtraits, on the other hand, may be related more to the affective aspect of language” (p.462).

Harley et al. (1990) report findings from a large-scale five-year research project entitled “Development of Bilingual Proficiency” at Ontario Institute for Studies in Education, to examine whether three hypothesized traits (grammar, discourse, and sociolinguistic competence) would emerge as distinct components of second language proficiency among 6th-grade French immersion students. Both multitrait-multimethod and confirmatory factor analysis were used, involving 175 students and 23 native speakers. The three hypothesized competence were measured by three methods: oral (interviews and storytelling), written multiple-choice, and written composition. The results showed that reliability of the multiple-choice test ranged from .58 to .75, and that the factor analysis identified two factors: a general factor and a method factor (written). The three-component distinction of grammar, discourse, and sociolinguistic competence was not strongly supported by this study.

Kim (2006) also investigated the relationships among the components of speaking ability. He operationally defined speaking ability as the product of grammatical
ability, grammatical meaning (the ability to produce and understand literal and intended meaning associated with an utterance), sociolinguistic competence, and discourse competence. Ninety-five ESL students participated in a semi-direct test involving ten tasks, which are given on computers and scored by human raters. The correlation analyses conducted in the study found “the hypothesized language domains were greatly correlated with one another […] thus, provided limited evidence for the construct validity of the four-dimension model of language ability measured by the speaking test. Although the four dimensions failed to emerge, it was found that the strengths of the relationships among the hypothesized components varied. Sociolinguistic competence, for example, appeared to have relatively weaker relationships with other components, indicating its distinctiveness to some degree” (p.29).

Röver (2005), on the other hand, found that his 3-section test (implicature, routine, and speech act) can be most easily interpretable as a four-factor solution in his exploratory factor analysis, closely resembling the test construct (each “factor” corresponding to one of the three test sections plus a “difficulty” factor) and concluded that his test structure “can be best viewed as tapping into a relatively homogenous pragmalinguistic knowledge base, with each section engaging slightly different aspects of that language.” Röver also conducted multimethod-multitrait analyses (multiple-choice, brief response) and found that the method effect is negligible.

Grabowski (2008) examined the construct validity of a test based on Purpura’s (2004) theoretical model, which specifies the grammatical and pragmatic (sociolinguistic, sociocultural, and psychological) components of language knowledge. The study
involved 90 ESL learners, utilizing six reciprocal written test tasks designed to elicit conventional conversation and turn-taking behavior in pairs, and required the test takers to both convey and interpret implied meanings in online performance. The study found high inter-rater reliability (.697 or better), high internal consistency (alpha = .850), moderately high correlation between grammatical knowledge and pragmatic knowledge (r = .569 to .828). Rasch analysis (a statistical analysis based on Item Response Theory) further revealed that achieving a high score for any of the pragmatic appropriateness dimensions was more difficult than for grammatical control in this test.

With only a few studies conducted so far, it remains to be seen if grammatical and pragmatic competence can be assessed separately, and if subcomponents of pragmatic knowledge can and should be identified and measured separately.

2.3.4. Test of JSL/JFL Pragmalinguistics

Few previous studies are found about testing pragmatic knowledge for JSL/JFL learners. There appears to be no large-scale standardized tests systematically designed to test pragmatic competence. The survey of currently available standardized tests of Japanese in the U.S. reveals no specific mention of pragmatic competence as test construct in most of the tests. They tend to measure grammar, vocabulary, and reading more extensively instead, and many test items are presented in isolated sentences with no contextual information available (Eda, Itomitsu, and Noda 2008).

There are, however, some sample test items for these large-scale tests that seem to be targeting pragmatic knowledge. For example, the Reading and Grammar section of the Japanese Language Proficiency Test (JLPT) contains a few test items that ask for an
appropriate expression to fill in a blank in a mini conversation (Japan Foundation 2008).

This test format does not always involve pragmatic knowledge, however. Okamoto et al. (2000) contains the following sample questions for the JLPT:

「使いかたを ごせつめい いたしましょうか。」
「ええ、________」。
1 せつめいします  2 わかりました
3 けっこうです  4 おねがいします

「田中さんも コーヒー ______。」
「はい、ありがとうございます。」
1 飲むの  2 飲むな
3 飲むわ  4 飲むかい

Both items can be thought of testing awareness of speech acts. The first item tests linguistic patterns associated with requests (and comprehension of offers), and the second one with offer. However, with no contextual information provided, it is questionable if the test item can invoke appropriate discourse domains (Selinker and Douglas 1985) or scripts (Schank 1999) among test takers — probable settings and contextual information they need to understand the given exchanges. For the second item, for instance, the intended context is most likely someone, probably a male, talking with someone named Tanaka. Option 4 then will be the most appropriate option. However, if, for instance, they are preparing cups of tea to a large number of people including Tanaka, a third person, options 3 (and marginally 1) will also be keyable. There is also no indication of intonation the options can take (options 1, 2, and 3 can take both rising and falling intonation, resulting in different illocutionary force, and the Japanese period “.” is often
used regardless of the intonation patterns). While JLPT provides a conversation as a stem (and not just isolated sentences), this lack of contextual information is detrimental in testing pragmatic knowledge.

There are additional examples in a large-scale standardized test in the Business Japanese Proficiency Test (BJT) created by the Japan External Trade Organization (JETRO), which is designed to assess communicative skills in typical daily business situations (Kato 2006). It is particularly noteworthy that the BJT includes control over ritual expressions and keigo in its test construct, as “facility to communicate at basic levels of politeness and formality” is listed among the criteria for rating (p.13). Thus the BJT sample questions contain a number of items that prompt different speech acts, as well as speech styles. Here are a few examples (p. 58, p.71):

A: あの〜、これたいしたものじゃないんですけど。
B: ___________________________.
   1 では、遠慮します
   2 では、ご辞退申し上げます
   3 あ、どうもご丁寧に
   4 あ、そうですか。

次の写真は、ある社員が海外駐在から帰国し、上司の部長にあいさつをしている場面です。この社員はどのようにあいさつすることが望ましいでしょうか。<photo of a man bowing to a senior man>

   1 今日付けで配属となった佐藤だよ。
   2 今日付けでご配属となる佐藤です。
   3 今日付けでご配属となりました佐藤です。
   4 今日付けで配属となりました佐藤です。

23 The stem and options are provided in both audio and in Japanese orthography.
24 The lead and options are provided only in audio. Only a photo is presented in the test booklet.
The first example tests knowledge of routines, presumably involving polite gift-giving and receiving, but the situation is not defined. The second example tests knowledge of speech style, and it is presented with a lead describing the situation and a photo. It is questionable, however, if all the test takers can understand the lead in Japanese, due to the sophisticated vocabulary used for the description of the context. Although at least they have visual clues as to what the intended speech events may be, this potential linguistic confounding is problematic, as situational information is critical in testing pragmatic competence. On the other hand, the relatively larger number of these types of items in the BJT is reassuring, and it may be an indication of increasing interest in measuring pragmatic knowledge among JFL learners, and the perceived importance of such knowledge on their part, particularly in business situations.

On a smaller scale, Taguchi (2008) developed an original instrument that measures pragmatic comprehension for JFL learners. Her study examined the ability to comprehend implied meaning and the effect of proficiency on comprehension. Sixty-three college students of Japanese at two proficiency levels completed a listening test measuring their ability to comprehend three types of implied meaning (12 items in each): indirect refusals (chotto, etc., with no explicit “No, I can’t”); conventional indirect opinions (chotto, amari, plus expressions like soo kana and to iu ki ga); and nonconventional indirect opinions (both negative and positive). Comprehension was analyzed for accuracy (scores on a multiple-choice measure) and comprehension speed. She found that there was a significant effect of item type on accuracy but not on
comprehension speed. Proficiency was observed to have an effect on accuracy, but not on comprehension speed. Here is an example of an Indirect Refusal item (p.575)\textsuperscript{25}:

Smith: *Satoo-san genki desu ka? Ima jikan arimasu ka?*  
Sato: *Ah Sumisu-san, doo shitan desu ka?*  
Smith: *Ee onegai ga arun desu ga. Kore nihongo no essei nandesu ga, boku no nihongo chekku shite moraemasen ka?*  
Sato: *Korekara compyutaan no jugyoo ga hachiji made arun desu.*

Question: Which statement is correct?  
1. The woman is going to check the man’s Japanese.  
2. The woman is taking a Japanese computer class.  
3. The woman can’t write Japanese essays.  
4. The woman can’t check the man’s Japanese now.

There are a few things that can be said about this test item. First, the stem (i.e., the question) is not very well directed (i.e., not specific), so that test takers have to read all the four options to come up with the right answer. Second, two of the four options (Options 1 and 2) check comprehension of details and are not directly related to the purpose of testing whether or not the test taker can understand the indirectness of the utterance as a refusal (the item could be a two-option item, asking whether Sato is complying with or rejecting the request). Third, the stimulus (the conversation) sounds incomplete or like part of a larger conversation, thus the distractors cannot be ruled out as clearly wrong, making all options keyable. (Sato says she is taking a computer class,

\textsuperscript{25} English equivalent (provided by Taguchi):  
Smith: Ms. Sato, how are you? Do you have time now?  
Sato: Oh, Mr. Smith, what’s up?  
Smith: Well, I have a favor to ask you. This is an essay I wrote for the Japanese class. Could you please check my Japanese?  
Sato: I have a computer class from now till eight.
which may or may not be in Japanese; her name suggests she is a Japanese native but
may not be; and she is saying she cannot help Smith now, but she may be going to help
him later.) However, these appear to be minor problems as the internal consistency
reliability estimates are relatively high (alpha = .74 or better in all three sections, .91 for
total).

Ishihara and her colleagues (Ishihara, Aoshima, and Akikawa 2008; Aoshima,
Ishihara, and Akikawa 2008) conducted a case study of classroom assessment of
pragmatic competence, with eight 3rd year JFL students as participants. They utilized
various assessment instruments (role-play, written DCT, reflective writing, self/peer-
assessment, student-teacher collaborated assessment, etc.) in order to investigate the
gains the learners made by using the web-based instructional units for the five speech acts
in Japanese identified by Cohen and Ishihara (2005). Analytical rubrics used for the
teachers’ evaluation contain such criteria as (1) overall directness, politeness, and
formality; (2) choice of requesting strategies (e.g., offering a reason for the request,
getting a pre-commitment, checking availability, promising to pay back, showing
consideration for the hearer, expressing apology/thanks); (3) grammar and word choice
(in terms of appropriateness, not accuracy); and (4) tone (e.g., use of intensifiers,
emphatic tone, bows). No statistical data is available for inter-rater reliabilities, etc. due
to the small number of the participants, but these studies show fine examples of
classroom teaching and testing in practice, utilizing various assessment tools to measure
both the process and the product of learning outcome.
Rating performance through role-play, DCT, etc. has an advantage in simulating authentic language use, but it also has its disadvantages in that it is time-consuming, labor-intensive, and costly to administer and score, as it involves human raters. It is also challenging to establish high reliability with such evaluation, and it may also involve potential logistics and test security issues (Brown and Hudson 2002). Norris et al. (1998) lists advantages and disadvantages of task-based performance assessments. An additional issue with role-play involves the interviewer-related variables that might affect the rating. Yamashita (1996) provides the following extract from her role-play test for the “at the car garage” scene, where the test taker is expected to ask a mechanic to expedite the repair process:

S3 = beginning level JFL male (ten days after arrival in Japan)
S3: Sumimasen.
   ‘Excuse me.’
IL [Interlocutor]: A, doomo irasshaimase.
   ‘Welcome.’
S3: Etto, senshuu boku no kuruma ga motte kimashita.
   ‘Umm.. Last week, I brought in my car.’
IL: A, hai, zya, suwatte kudasai.
   ‘Oh, yes, please sit down.’ [IL asked S3’s name]
IL: E.. XX-san no kuruma wa asatte dekirun desukedo.
   ‘Um.. Your (name) car will be ready the day after tomorrow.’
S3: Demo, etto, ashiita…
   ‘Well, but, tomorrow…’
IL: Ashita hoshii desu ka?
   ‘You want it tomorrow?’
S3: Ashita hoshii desu.
   ‘I want it tomorrow.’
IL: Ashita hoshii desu ka. Hai, wakarimashita.
   ‘You want it tomorrow, I see. Yes, I understand.’
As Yamashita observes, “the native Japanese interlocutor constantly supplied the necessary words and expressions in the sequence of conversation with a very poor beginner (JFL), so that the conversation could flow. As a result, the roleplays, even those of beginners, sounded as if the conversation went smoothly” (p. 67). Therefore the seeming success in accomplishing a task in role-play alone may not be a strong indicator of the test takers’ ability in Japanese. A few cases of simplifications (suwatte kudasai, asita hoshii desu ka?) might have been triggered as well, as the S3’s first utterance contains a grammatical error (misuse of a phrase particle ga instead of the correct o).

Although it is likely that such simplifications and word-supplying take place in actual, non-test conversational situations between a target-native and target-nonnative in the target culture, the amount, types of strategies, and skills to implement such “help” vary greatly from one interviewer to another. This brings in a host of variables that affect the rating of the performance that might affect the test reliability.

When it comes to testing pragmatic knowledge, it seems particularly important to utilize multiple test formats to compensate for the weakness of each test format.

Computer-assisted language testing (CALT) has been identified as a promising direction that might contribute to the development of measuring pragmatic ability. CALT has a number of advantages, as summarized by Chapelle and Douglas (2006): CALTs can be taken at many convenient locations, at convenient times, and largely without human intervention. CALTs’ multimedia capabilities allow for a variety of input modalities, and can minimize the interfering effect of reading comprehension ability. The multimedia

---

26 This may not be a negative effect if the test construct views such interviewer variables as (integral) part of co-constructional process that reflects testee’s pragmatic competence.
format can also be used to enhance contextualization and perceived authenticity of test items, which is of critical importance in measuring pragmatic knowledge. Web-based multiple-choice test has been created as tests of pragmalinguistics (Röver 2005; Liu 2007). While the multiple-choice format may not be suitable for online production and comprehension of discourse-level task completion (involving turn taking, backchanneling, hesitations, etc.), it can serve as a reliable measurement tool for an offline assessment of skills describable toward the linguistic “end” of a scale of pragmatic competence. As there are few previous studies on measuring pragmalinguistic knowledge of JSL/JFL learners, the potentials of CALT is worth exploring the possibilities.

2.4. Summary

Reviews of the literature show that pragmatic knowledge has been identified as one of the key components in communicative language ability, and a growing number of studies have been conducted on different communication realization patterns in English and Japanese. Subcomponents of pragmatic knowledge, especially the knowledge associated with particular speech acts, have received attention in the study of interlanguage pragmatics, and a few tests have emerged to test both sociopragmatic and pragmalinguistic knowledge for ESL/EFL learners. For JSL/JFL, few existing standardized tests measure pragmatic knowledge per se, and some items targeted at measuring pragmatic knowledge do not provide any contextual information. Among the existing test formats for measuring pragmatic knowledge, the multiple-choice format for testing sociopragmatic knowledge presents a challenge while there is a few study that
achieved high reliability in its test of pragmalinguistics. The advantages of CALT show potential in creating a test of pragmalinguistics for JSL/JFL learners, and Röver’s study has been adopted as a model for this study. The next chapter describes the project in detail, including the purpose of the study, definitions that comprise the test construct, and research questions.
Chapter 3: The Study

This chapter provides a detailed description of the study and the test construction process, including the targeted population, instrument design, content and format specifications, and the pilot testing. The test construct and design of the test follows Röver (2005), with a few modifications based on findings from previous studies and considerations specific to Japanese. Decisions on the test design have aimed to ensure the overall usefulness of the test, including its reliability, construct validity, content validity, and practicality.

3.1. Purpose

The purpose of this dissertation project is to produce a useful test of pragmatic knowledge for JSL/JFL learners. This test is expected to function as a complementary tool to measure learners’ communicative language ability in general. For our purposes, the usefulness of this test will be understood as a function of several different qualities, as defined in Bachman and Palmer (1996: 18):

\[
\text{Test usefulness} = \text{reliability} + \text{construct validity} + \text{authenticity} + \text{interactiveness} + \text{impact} + \text{practicality}
\]

Bachman and Palmer define \textit{reliability} as consistency of measurement; \textit{construct validity} as the meaningfulness and appropriateness of the interpretations that we make on the
basis of the test scores; *authenticity* as the degree of correspondence on the characteristics of a given test task to the features of target language use task; *interactiveness* as the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task; *impact* as effect of the test on society and educational systems and upon the individuals within those systems; and *practicality* as the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities.

In this study, reliability and construct validity will be investigated in detail and incorporated in our research questions, and the other aspects of the test usefulness will be included in the discussion section.

3.2. Construct

Pragmatic knowledge encompasses many aspects of language use. This study focuses on part of such knowledge. The targeted test construct for this study is off-line pragmatic knowledge, which can be theoretically defined as knowledge of linguistic forms that allows one to interpret and convey intentions appropriately in a variety of contexts. This test construct is further operationally defined as knowledge of linguistic forms typically associated with speech acts, routines, and speech styles in specific contexts. Test items will be created by sampling from these three domains of pragmatic knowledge that are deemed important for JFL learners.

Our construct for a *routine*, as tested in the Routines section, is defined as knowledge of highly conventionalized prepatterned expressions whose occurrence is tied
to more or less standardized aspects of communication (including ritual and other formulaic expressions), following Coulmas (1981).

The construct of speech acts, as tested in the Speech Acts section, is defined as knowledge of linguistic communication patterns conventionally associated with performing such functions as suggesting, inviting, requesting, offering, giving advice, etc. Prototypical semantic formulas (or IFIDs) are thus targeted as a test construct.

The construct of speech styles, as tested in the Speech Style section, is defined as knowledge of appropriate stylistic use of the language and selection of linguistic forms (accepted as prototypical and proper in the given context among experts in the target culture) associated with different levels of politeness with different interlocutors or referents, following Wetzel (2001).

The above construct definitions are similar to Röver’s, but different in a few significant ways. First, the theoretical construct definition for this study does not include a strategic aspect, i.e., learners’ ability in choosing a particular speech act or other courses of action appropriately from a number of options in the given context. Only the linguistic items that are typically associated with particular situations, interlocutors, or speech act will be measured. In other words, the linguistic choice the test takers make that reflects knowledge of conventions of form (Clark 1979) will be the focus of the test construct, rather than learners' speech act realization strategies, or conventions of means. Thus, an analysis of situations and speech act intentions will be provided to the test takers as a given, leaving the test takers to choose appropriate linguistic forms that would be easily understood as typical speech act realization patterns for those situations and
intentions. Second, Röver’s implicature section is replaced by the Speech Styles section in this study, for the following reasons: (1) a test for implicatures in Japanese already exists (see Taguchi 2008, reviewed in the previous chapter); (2) there are few substantial studies available on how implicatures are typically performed and understood in Japanese, and types and the amount of implicature strategies used — especially “Pope Questions” (answering a question whose answer is obvious by providing such a question as “Is the Pope Catholic?”) — are expected to differ greatly in Japanese; and (3) the relative importance of different speech styles in JSL/JFL perceived among teaching professionals and teaching materials (Itomitsu 2005). Third, in addition to the three subsections of pragmalinguistic knowledge, grammatical knowledge is also measured to investigate a possible relationship among grammatical and subcomponents of pragmalinguistic knowledge. Grammatical knowledge is defined as knowledge of the well-formedness of isolated sentences independent of contextual information.

3.3. Research Questions

Research questions for this study include the following:

(1) What is the reliability of the test, of its sections, and its test items? The reliability of the test is measured by calculating Cronbach’s alpha, a general formula for estimating internal consistency developed by Cronbach in 1951 (Bachman 1990). Reliability is an important component of test usefulness and a prerequisite for validation studies. Overall test reliability as well as reliability of each section will be investigated. This question is of particular concern, as some of the previous studies on testing
pragmatics with the multiple-choice format resulted in low reliability. Difficulty and item-total correlation indices of each item will also be investigated.

(2) What is the relationship between the test scores and length of study/exposure? What is the relationship between the test scores and levels of instruction? Are there significant differences among learners of different levels of instruction and/or length of study/exposure in the mean performances on the test? These questions concern group differences, and the results inform the construct validity of the test. It is hypothesized that the overall scores increase as the learners have longer exposures to language, and/or they advance to higher levels of study. However, each subcomponent may show different developmental trajectories, suggesting differences in typical learning rate across the sections.

(3) What is the relationship between the test scores and learners’ proficiency levels? Are there significant differences among learners of different proficiency ratings in the mean performances on the test? This question concerns concurrent validity, by comparing the test scores with the proficiency levels of the test takers, as defined in the ACTFL Guidelines and measured by procedures similar to the Oral Proficiency Interview (OPI), a well-established measure of spoken proficiency. As reviewed in Chapter 2, the ACTFL Guidelines views language proficiency as a unitary ability. Although the construct adopted for this dissertation study is multi-componential and different from how the ACTFL defines language proficiency, it is still revealing to see how the two compares, given the impact of the ACTFL Guidelines in the field. According to the ACTFL Guidelines, pragmatic competence is recognized as having subcomponents of
accuracy, along with grammar, vocabulary, pronunciation, sociolinguistic rules, and fluency (Makino 1991). As pragmatic competence is perceived as part of the global proficiency rating, it is hypothesized that the test scores positively correlate with ACTFL rating. It is hypothesized that the higher the learners’ proficiency rating, the higher overall scores as well as section scores will be.

(4) What are the relationships among the scores for grammar and the three sections (Speech Acts, Routines, Speech Styles) of pragmatics? Is there a relationship between grammatical and pragmatic knowledge? Correlation analyses as well as exploratory factor analysis will be conducted to investigate the interrelationship among the subcomponents of the test. Correlation of sections in different proficiency levels will also be investigated, as Kim (2006) found the degrees of correlations were the largest in the low proficiency group, indicating that different dimensions might not have emerged equally, due to significantly limited linguistic resources.

3.4. Target Population

The target population for this particular test, for the study reported in this dissertation, is learners of Japanese in U.S. universities and colleges. Emphasis will be placed on learners at non-beginning stages — (future) undergraduate Japanese majors. It is assumed that the test takers are competent in English enough to read directions and follow instructions, and that they have basic computer skills to access to a website, able to handle audio files on a browser, and navigate through a computer-delivered test. It is also assumed that the test takers have fundamental reading proficiency (able to read
hiragana, including furigana) and listening comprehension proficiency to understand the stems (i.e., test questions) and the options presented in Japanese.

3.4. Instrument Design

A web-based computer-assisted testing format has been adopted for its accessibility, ease of test administration and its multimedia capabilities to present contextual information and sounds. The multimedia presentation of context with color graphics, “potentially enhancing authenticity of both input and response […] particularly situational authenticity, which defines authenticity in terms of the features of context including setting, participants, content, tone, and genre” (Chapelle and Douglas 2006) is deemed indispensable for a test of pragmatics. Availability of sound would also “decrease the influence of reading proficiency on test scores, and […] could even lead to a test that requires minimal or no literacy skills. Such an approach would greatly broaden the possible test taker population, and improve the precision of measurement at lower levels of second language proficiency where the reading affect interferes most with measurement of the construct” (Röver 2005). This is particularly the case with JFL, given the complexity of the Japanese writing systems. Among many choices in creating such a Web-based, multimedia test, a commercial online survey creation site, SurveyGizmo, has been selected for its user-friendly interface design, multimedia capabilities (including multilanguage display), low maintenance cost, and rich data tracking, reporting and export capabilities.
The test has been designed to be self-paced, in which learners can take the time needed to complete the test. As a practical consideration, the number of items is limited so that the average time to complete the test will be approximately sixty minutes. The final operationalized test contains 12 items in each of the 4 sections (the Grammar, Routines, Speech Acts, and Speech Styles sections), making the total number of the test items 48. The multiple-choice format has been selected as the item type, for its advantage in providing consistent and standardized test environment, and its ease of test administration and scoring.

3.5. Content Specifications

Content specifications are of critical importance, as they directly affect the content validity of the test — if and to what degree the test items are representative and relevant samples of the knowledge the test has been designed to measure. This is far easier said than done, however. Röver (2005: 115-6) observes:

[…] Ensuring content validity] requires ethnographic studies of real world language use, or at least analyses of corpora of spoken language. To keep pace with diachronic changes in language use, however, ethnographic studies would have to be repeated at regular intervals and corpora would have to be regularly updated. Both are extremely expensive endeavors […] The second best option might be targeted elicitation of response plausibility in pilot studies with native speakers. They can identify seriously outdated response options and provide suggestions for revision but is still likely to be subject to native speakers’ beliefs of what they ‘should’ say, regardless of what is actually commonly said.

Brown and Hudson (2002) describe two approaches to ensure content validity in test design: a theoretical argument approach, and an expert judgment approach. The
theoretical argument approach “makes a case for the representativeness of the content of the a test by (a) carefully planning a test on the basis of theoretical, research, or practical perspectives and (b) arguing for the content validity of the test on the basis of those perspectives” (p.213); and the expert judgment approach asks “experts […] who are trained in or knowledgeable about the content/processes being tested […] to judge the degree to which the test (or each individual test item) is testing the relevant content” (p. 220). This study utilizes both approaches in test production design. The theoretical argument approach is taken for test item production, and the expert judgment approach for item revision and selection during the pilot test.

For item production that employs the theoretical argument approach, in addition to the established taxonomies on linguistic items and language use situations, data from a series of textbook analyses (Itomitsu 2005, 2008) have been incorporated in the creation of content specifications. This is motivated by the assumption that agreements observed in several textbook series popularly used in the U.S. reveal typical knowledge and language use situations that are deemed important among textbook authors as well as language teachers who decided to adopt one of the textbook series. This textbook-analyses approach has also been taken by Norris et al. (1998) as a theoretical argument approach for a creation of a task-based performance assessment when selecting task themes “that authors of a variety of different language teaching textbooks found important enough to include in their language teaching materials” (Brown and Hudson 2002: 214). With the absence of extensive ethnographic studies and corpora of contact situations the learners will most likely encounter, the data from these textbook analyses
provide valuable information with which the content specifications can be generated. Data from textbook analyses, however, should be taken as a second best option to ethnographic studies, as conversations and other examples in textbooks may not accurately reflect patterns of the actual language use in the target culture (Jones and Ono 2005; Sei 2006). To alleviate any negative effect of potentially inauthentic stimuli, and to complement the theoretical argument approach, each test item was examined by experts during Pilot Study as part of expert judgment approach (See Section 3.9, following).

For the Grammar section, a list of structural patterns was generated by adapting data from Itomitsu (2005), which compares five popular Japanese language textbook series in the United States in terms of their selection and sequencing of structural items. Itomitsu found that of the 303 structural items overall, 211 (approximately 70%) of the structural items appear in four or more textbooks, which shows a high level of consensus in the field. This study also yielded results similar to the JLPT content specification for grammar for Levels 3 and 4 (Japan Foundation and Association of International Education 1994), which is based on a similar textbook analysis of eight textbook series for the JSL/JFL learners. The 211 structural patterns are used as the basis for content specifications for the grammar section. Makino and Tsutsui (1986, 1995), well-known reference dictionaries on grammar, were also consulted. Muzutani (1994) was consulted in sampling typical, commonly observed errors among learners of Japanese. Appendix A contains the list of the structural patterns targeted in the Grammar section.

For the Routines section, a list of routine expressions was created by selecting approximately 100 expressions from "A handbook of Common Japanese Phrases"
(Sanseido 2002), which lists approximately 600 such phrases. The selected 100 include routine expressions (especially as found in telephone conversations, dialogues typical of visiting someone’s house, etc.) attested in Itomitsu’s (2008) textbook analysis of the speech events in conversation sections of five textbook series in the U.S. The list also includes all of the JLPT content specification for “greetings and other expressions” (あいさつ語等表現) for the JLPT Levels 4 and 3 (Japan Foundation and Association of International Education 1994). Appendix B contains the list of routine expressions (明けましておめでとうございます、いらっしゃいませ、はじめまして、etc.).

For the Speech Acts section, the four most frequently attested Speech Acts categories in popular textbooks used in the US were chosen, based on Itomitsu’s (2008) study (under the "suasion" section). The four categories are, in order of observed frequencies, as follows: requesting (includes instructing or directing, 201 tokens); suggesting a course of action (including encouraging and inviting in which both speaker and listener doing something together, 168 tokens); offering assistance and/or service (94 tokens), and advising (recommending to others what s/he should do, 65 tokens). A caveat is in order here, as Itomitsu’s analysis encountered difficulties interpreting conversations in textbooks in terms of traditional speech act perspective, and therefore should not taken as an absolutely accurate reflection of frequencies of particular speech acts in target language use situations. However, the relative frequency observed in that study, as well as the list of attested linguistic features associated with these speech acts, have been informative for narrowing the scope of the test, and those results have been used as a sounding board for the study. Appendix C contains the list of semantic formulae
for each of the four speech acts (〜てくださいませんか for request, 〜ませんか for suggestion, etc.).

For the Speech Styles section, Itomitsu’s (2005) data on speech styles included in the textbooks as well as the JLPT content specifications (Japan Foundation and Association of International Education 1994) for Levels 3 and 2 are considered. Both the lexical and productive types of two major categories of keigo (the honorific-polite and the humble-polite) are listed in Appendix D (e.g., lexical honorific/humble 召し上がる・いただく, productive honorific/humble お～になる・お～する, etc.)

3.6. Format Specifications

Each multiple-choice item contains directions in English, and an explicit, written lead (containing contextual information and description of speech act, etc., also in English), as well as illustration depicting the situation. Illustrations for the test items were created with materials originally produced for Japanese: The Spoken Language Parts 2 & 3, Interactive DVD Program (Noda and Itomitsu 2008).

English is used for the directions and descriptions of context, to ensure that the test takers understand the critical contextual information, as selecting typical linguistic forms associated with particular contextual information is the test construct of this project. This decision of course involves the risk of limiting the test taker population, as well as of having test takers process English and Japanese back and forth, but was deemed suitable for the development of the test for JFL learners in the U.S. Every attempt was made to make the directions and the descriptions concise and straight-forward.
The leads and illustrations minimally provide the following contextual information: (1) name(s) of participant(s), (2) gender of the participants, (3) social roles and the relationship between the participants, (4) scene of the interaction, and (5) age range of the participants. Both casual (minimally, the da-style) as well as careful (minimally, desu/masu-style) are included in all sections except for the Speech Styles section.

Both the stem (utterances in Japanese with a blank to fill in) and the four options are provided in Japanese orthography (with hurigana on every kanji) and in audio. This is to minimize the effect of reading proficiency and to provide appropriate intonation patterns for each option. Both male and female voices are included to balance the gender variable.

Here is an example of a test item (from the Routines section):

---

27 For conversations in casual-style interaction, first names (e.g., Takashi) were used. For more formal, careful-style interactions, last names with titles (e.g., Ms. Suzuki) were used.
28 Items in the Speech Style section, for its very nature of testing humble- and honorific-polite forms, are all presented in the desu/masu-style.
Figure 3.1: Item Format Example

English equivalents of the stem and options:
“Well, (then), ________.”
“Bye. (lit. come, having gone!)”
A. (I’m back) just now.
B. (Oh, you are) back home!
C. Bye (lit. will be back)
D. Bye (lit. come, having gone!)
Options are created so that each test item can tap into the different types of knowledge (i.e., grammatical and pragmatic knowledge) that it is designed to measure.

For example, observe the following pair of items:

**Item A:** Takashi is at an electric appliance store, talking with a clerk at a counter. What would Takashi probably say?
すみませんが、ちょっとそれ見せて____________。
- くれませんか。
- きませんか。
- しませんか。
- お願いします。

**Item B:** Takashi is at an electric appliance store, talking with a clerk at a counter. Takashi is requesting assistance. What would Takashi probably say?
すみませんが、ちょっとそれ____________。
- 見てくれませんか。
- 見せたいんですか。
- 見せてあげましょうか。
- 見せてもらいたいでしょうか。

Although the leads and completed stems are identical, Item A is a grammar item, as くれませんか is the only option that can directly follow a verbal gerund. Test takers only need grammatical knowledge, and do not require contextual information (who is speaking to whom and for what purpose) to eliminate the other options. Item B, on the other hand, is a speech act item, as the key 見てくれませんか is the only option that can be interpreted as a typical request in the given context. Other options are created as such that they would create grammatically correct utterances, but would not be understood as requests in the given context. Therefore, the contextual information is vital for the test takers to choose the correct answer.
For the Grammar section, the distractors are the ones that form ungrammatical sentences when combined with the stem, regardless of contextual information. For the Routines sections, the distractors are routine expressions that would not fit to the given situation. For the Speech Acts section, the distractors are in the same level of speech style as the key, and would form a grammatical sentence, but fail to deliver the specified speech act in the given conversational context. For the Speech Styles section, the distractors contain different types of politeness that resulting in inappropriate utterance due to its mismatch with referent(s) and/or with the interlocutor(s).

3.7. Pilot Testing

A pilot test was conducted in order to gather information about each test item, and to revise and select items for the actual, operationalized test.

First, 16 items were created for each of the four sections (64 items total). The 64 items were then delivered on the Internet, just as the operationalized test would be. Experts, teachers of Japanese in U.S. colleges and universities, were asked to participate in the pilot test to evaluate each of the 64 items. Based on the results, revised 48 items were chosen for the operationalized test.

Twelve teachers of Japanese at U.S. colleges and universities participated in the pilot, which ran from October 10th through November 5th, 2008. The twelve experts (3 males and 9 female) from 5 different institutions consisted of 4 in professorial rank, 4 in instructor/lecturer position, and 4 teaching assistants. 8 of the 12 are native speakers of
Japanese. Their average years of teaching Japanese are 7.3 years, with 4 teachers teaching 10 years or more.

For each item, the experts were asked to indicate the correct answer, as well as rate its perceived importance (“This is what a Japanese major should know”), authenticity (“The situation and language use are natural and authentic”), and ease (“This item is easy for a Japanese major”) on 5-point Likert scales. The experts were also provided opportunities to freely comment on the test items. It took approximately 90 minutes for the experts to answer all the questions.

The following is a screen capture of a pilot test page:
Pilot Test: Japanese

Part 3: Pilot Test -- Routines --

1. Mr. Abe, an employee, is talking with his wife, as he leaves his house for work.

What would he probably say to his wife?

Read and/or listen to the Japanese:

A. ただいま
B. お帰りなさい
C. 行ってきます
D. 行ってらっしゃい

66. Choose the best option from above:

○ A
○ B
○ C
○ D

67. What do you think about this test item?

<table>
<thead>
<tr>
<th>strongly agree</th>
<th>agree</th>
<th>somewhat</th>
<th>disagree</th>
<th>strongly disagree</th>
</tr>
</thead>
</table>

This is what a Japanese major should know.

The situation and language are natural and authentic.

This item is easy for a Japanese major.

68. Comments:

Figure 3.2: Pilot Test Screen Capture
Based on the findings, a number of revisions were made, and 48 items, all with 100% agreement on the key option among experts\textsuperscript{30}, were selected for the operationalized test. The items are then ordered in terms of the perceived difficulty level, from easy to difficult. See Appendix F for the samples of test items and deleted items for the operationalized test. The following tables (Tables 3.1, 3.2, 3.3, and 3.4) summarize the results of the pilot test and selection procedures for each section:

\textsuperscript{30} One item in Speech Acts section with a 92% agreement among the experts was included after revision, because all “advising” items were eliminated. See the explanation following Table 4.
<table>
<thead>
<tr>
<th>Item #</th>
<th>decision</th>
<th>Agreement</th>
<th>Importance</th>
<th>Authenticity</th>
<th>Ease</th>
<th>Notes$^{31}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR01</td>
<td>selected</td>
<td>100%</td>
<td>5.00</td>
<td>3.62</td>
<td>4.46</td>
<td>Lead: role changed to “senior student” (speech style)</td>
</tr>
<tr>
<td>GR02</td>
<td>selected</td>
<td>100%</td>
<td>4.92</td>
<td>4.08</td>
<td>4.38</td>
<td>Lead: role changed to “middle-aged man”</td>
</tr>
<tr>
<td>GR03</td>
<td>selected</td>
<td>100%</td>
<td>4.92</td>
<td>4.77</td>
<td>4.08</td>
<td></td>
</tr>
<tr>
<td>GR04</td>
<td>selected</td>
<td>100%</td>
<td>4.85</td>
<td>4.38</td>
<td>4.00</td>
<td>Stem: changed to “X da yo” (speech style)</td>
</tr>
<tr>
<td>GR05</td>
<td>selected</td>
<td>100%</td>
<td>4.77</td>
<td>4.62</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>GR06</td>
<td>selected</td>
<td>100%</td>
<td>4.92</td>
<td>4.77</td>
<td>3.62</td>
<td></td>
</tr>
<tr>
<td>GR07</td>
<td>selected</td>
<td>100%</td>
<td>4.77</td>
<td>4.62</td>
<td>3.46</td>
<td></td>
</tr>
<tr>
<td>GR08</td>
<td>selected</td>
<td>100%</td>
<td>4.62</td>
<td>4.54</td>
<td>3.23</td>
<td></td>
</tr>
<tr>
<td>GR09</td>
<td>selected</td>
<td>100%</td>
<td>4.77</td>
<td>4.62</td>
<td>3.08</td>
<td>Stem: changed to “nakunattyatta n da”</td>
</tr>
<tr>
<td>GR10</td>
<td>selected</td>
<td>100%</td>
<td>4.23</td>
<td>4.46</td>
<td>2.62</td>
<td>Stem: changed to “DVD pureeyaa ga”; Option C and D changed</td>
</tr>
<tr>
<td>GR11</td>
<td>selected</td>
<td>100%</td>
<td>4.69</td>
<td>4.54</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>GR12</td>
<td>selected</td>
<td>100%</td>
<td>4.00</td>
<td>4.54</td>
<td>2.46</td>
<td></td>
</tr>
<tr>
<td>GR-A</td>
<td>deleted</td>
<td>100%</td>
<td>4.62</td>
<td>3.85</td>
<td>3.69</td>
<td>A few experts commented the stem was awkward and unnatural.</td>
</tr>
<tr>
<td>GR-B</td>
<td>deleted</td>
<td>92%</td>
<td>4.69</td>
<td>4.77</td>
<td>3.00</td>
<td>Not in 100% agreement on key</td>
</tr>
<tr>
<td>GR-C</td>
<td>deleted</td>
<td>92%</td>
<td>4.77</td>
<td>4.77</td>
<td>3.38</td>
<td>Not in 100% agreement on key</td>
</tr>
<tr>
<td>GR-D</td>
<td>deleted</td>
<td>92%</td>
<td>4.77</td>
<td>4.62</td>
<td>3.00</td>
<td>Not in 100% agreement on key; one expert commented this is a vocabulary test</td>
</tr>
</tbody>
</table>

Table 3.1: Pilot Test Results Summary (Grammar, GR)

---

$^{31}$ The Notes section includes revisions prompted by the pilot test, or reasons why the items were deleted.
<table>
<thead>
<tr>
<th>Item #</th>
<th>decision</th>
<th>Agreement</th>
<th>Importance</th>
<th>Authenticity</th>
<th>Ease</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RT01</td>
<td>selected</td>
<td>100%</td>
<td>4.85</td>
<td>4.69</td>
<td>4.69</td>
<td></td>
</tr>
<tr>
<td>RT02</td>
<td>selected</td>
<td>100%</td>
<td>4.77</td>
<td>4.62</td>
<td>4.46</td>
<td>Lead: role changed to “sister”.</td>
</tr>
<tr>
<td>RT03</td>
<td>selected</td>
<td>100%</td>
<td>4.85</td>
<td>4.69</td>
<td>4.31</td>
<td></td>
</tr>
<tr>
<td>RT04</td>
<td>selected</td>
<td>100%</td>
<td>4.62</td>
<td>4.77</td>
<td>3.92</td>
<td></td>
</tr>
<tr>
<td>RT05</td>
<td>selected</td>
<td>100%</td>
<td>4.69</td>
<td>4.62</td>
<td>3.77</td>
<td></td>
</tr>
<tr>
<td>RT06</td>
<td>selected</td>
<td>100%</td>
<td>4.69</td>
<td>4.69</td>
<td>3.62</td>
<td></td>
</tr>
<tr>
<td>RT07</td>
<td>selected</td>
<td>100%</td>
<td>4.31</td>
<td>4.62</td>
<td>3.46</td>
<td>Lead: role changed to “middle-aged man”</td>
</tr>
<tr>
<td>RT08</td>
<td>selected</td>
<td>100%</td>
<td>4.69</td>
<td>4.69</td>
<td>3.38</td>
<td></td>
</tr>
<tr>
<td>RT09</td>
<td>selected</td>
<td>100%</td>
<td>4.62</td>
<td>4.54</td>
<td>3.38</td>
<td></td>
</tr>
<tr>
<td>RT10</td>
<td>selected</td>
<td>100%</td>
<td>4.62</td>
<td>4.69</td>
<td>3.08</td>
<td></td>
</tr>
<tr>
<td>RT11</td>
<td>selected</td>
<td>100%</td>
<td>4.31</td>
<td>4.62</td>
<td>2.85</td>
<td>Option D changed</td>
</tr>
<tr>
<td>RT12</td>
<td>selected</td>
<td>100%</td>
<td>4.23</td>
<td>4.69</td>
<td>2.69</td>
<td></td>
</tr>
<tr>
<td>RT-A</td>
<td>deleted</td>
<td>92%</td>
<td>4.31</td>
<td>4.62</td>
<td>3.54</td>
<td>Not in 100% agreement on key</td>
</tr>
<tr>
<td>RT-B</td>
<td>deleted</td>
<td>100%</td>
<td>3.69</td>
<td>4.62</td>
<td>2.15</td>
<td>Potentially sensitive topic and may raise affective filters (funeral)</td>
</tr>
<tr>
<td>RT-C</td>
<td>deleted</td>
<td>92%</td>
<td>3.92</td>
<td>4.54</td>
<td>2.92</td>
<td>Not in 100% agreement on key</td>
</tr>
<tr>
<td>RT-D</td>
<td>deleted</td>
<td>100%</td>
<td>4.15</td>
<td>4.69</td>
<td>3.31</td>
<td>“importance” rating relatively low</td>
</tr>
</tbody>
</table>

Table 3.2: Pilot Test Results Summary (Routines, RT)
<table>
<thead>
<tr>
<th>Item #</th>
<th>decision</th>
<th>Acts</th>
<th>Agree-ment</th>
<th>Importance</th>
<th>Authenticity</th>
<th>Ease</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA01</td>
<td>selected</td>
<td>Request</td>
<td>100%</td>
<td>4.92</td>
<td>4.85</td>
<td>3.46</td>
<td></td>
</tr>
<tr>
<td>SA02</td>
<td>selected</td>
<td>Request</td>
<td>100%</td>
<td>5.00</td>
<td>4.77</td>
<td>3.46</td>
<td></td>
</tr>
<tr>
<td>SA03</td>
<td>selected</td>
<td>Request</td>
<td>100%</td>
<td>4.92</td>
<td>4.85</td>
<td>3.62</td>
<td></td>
</tr>
<tr>
<td>SA04</td>
<td>selected</td>
<td>Request</td>
<td>100%</td>
<td>4.69</td>
<td>4.62</td>
<td>2.31</td>
<td></td>
</tr>
<tr>
<td>SA05</td>
<td>selected</td>
<td>Offer</td>
<td>100%</td>
<td>4.92</td>
<td>4.62</td>
<td>4.31</td>
<td>Option C changed</td>
</tr>
<tr>
<td>SA06</td>
<td>selected</td>
<td>Offer</td>
<td>92%</td>
<td>4.92</td>
<td>4.85</td>
<td>3.15</td>
<td>Options re-recorded with consistent intonation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Stem changed</td>
</tr>
<tr>
<td>SA07</td>
<td>selected</td>
<td>Offer</td>
<td>100%</td>
<td>4.69</td>
<td>4.62</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>SA08</td>
<td>selected</td>
<td>Offer</td>
<td>100%</td>
<td>4.69</td>
<td>4.85</td>
<td>2.62</td>
<td></td>
</tr>
<tr>
<td>SA09</td>
<td>selected</td>
<td>Suggest</td>
<td>100%</td>
<td>4.92</td>
<td>4.85</td>
<td>4.15</td>
<td></td>
</tr>
<tr>
<td>SA10</td>
<td>selected</td>
<td>Suggest</td>
<td>100%</td>
<td>4.62</td>
<td>4.69</td>
<td>4.08</td>
<td></td>
</tr>
<tr>
<td>SA11</td>
<td>selected</td>
<td>Suggest</td>
<td>100%</td>
<td>4.92</td>
<td>4.38</td>
<td>3.62</td>
<td>Lead: role changed to “coworker” (authenticity)</td>
</tr>
<tr>
<td>SA12</td>
<td>selected</td>
<td>Suggest</td>
<td>100%</td>
<td>4.62</td>
<td>4.62</td>
<td>3.31</td>
<td></td>
</tr>
<tr>
<td>SA-A</td>
<td>deleted</td>
<td>Advice</td>
<td>92%</td>
<td>4.69</td>
<td>4.77</td>
<td>3.00</td>
<td>“Advice” category eliminated</td>
</tr>
<tr>
<td>SA-B</td>
<td>deleted</td>
<td>Advice</td>
<td>100%</td>
<td>4.54</td>
<td>4.69</td>
<td>3.54</td>
<td>“Advice” category eliminated</td>
</tr>
<tr>
<td>SA-C</td>
<td>deleted</td>
<td>Advice</td>
<td>100%</td>
<td>4.23</td>
<td>4.62</td>
<td>2.69</td>
<td>“Advice” category eliminated</td>
</tr>
<tr>
<td>SA-D</td>
<td>deleted</td>
<td>Advice</td>
<td>100%</td>
<td>4.62</td>
<td>4.31</td>
<td>3.38</td>
<td>“Advice” category eliminated</td>
</tr>
</tbody>
</table>

Table 3.3: Pilot Test Results Summary (Speech Acts, SA)
<table>
<thead>
<tr>
<th>Item #</th>
<th>decision</th>
<th>Keigo group</th>
<th>Agree -ment</th>
<th>Import -ance</th>
<th>Authen -ticity</th>
<th>Ease</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS01</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>5.00</td>
<td>4.83</td>
<td>4.00</td>
<td></td>
</tr>
<tr>
<td>SS02</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.92</td>
<td>4.77</td>
<td>3.69</td>
<td></td>
</tr>
<tr>
<td>SS03</td>
<td>selected</td>
<td>honorific</td>
<td>100%</td>
<td>4.85</td>
<td>4.69</td>
<td>4.31</td>
<td>Lead, stem: “company” changed to “client (Kato-san)” (authenticity)</td>
</tr>
<tr>
<td>SS04</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.67</td>
<td>4.50</td>
<td>3.08</td>
<td></td>
</tr>
<tr>
<td>SS05</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.58</td>
<td>4.75</td>
<td>3.00</td>
<td></td>
</tr>
<tr>
<td>SS06</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.83</td>
<td>4.83</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>SS07</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.50</td>
<td>4.83</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>SS08</td>
<td>selected</td>
<td>honorific</td>
<td>100%</td>
<td>4.77</td>
<td>4.77</td>
<td>2.92</td>
<td></td>
</tr>
<tr>
<td>SS09</td>
<td>selected</td>
<td>honorific</td>
<td>100%</td>
<td>4.62</td>
<td>4.85</td>
<td>2.77</td>
<td></td>
</tr>
<tr>
<td>SS10</td>
<td>selected</td>
<td>honorific</td>
<td>100%</td>
<td>4.88</td>
<td>4.83</td>
<td>2.67</td>
<td></td>
</tr>
<tr>
<td>SS11</td>
<td>selected</td>
<td>humble</td>
<td>100%</td>
<td>4.58</td>
<td>4.75</td>
<td>2.58</td>
<td></td>
</tr>
<tr>
<td>SS12</td>
<td>selected</td>
<td>honorific</td>
<td>100%</td>
<td>4.25</td>
<td>4.75</td>
<td>2.58</td>
<td></td>
</tr>
<tr>
<td>SS-A</td>
<td>deleted</td>
<td>humble</td>
<td>100%</td>
<td>5.00</td>
<td>4.69</td>
<td>4.23</td>
<td>potentially testing set-phrase (self-introduction)</td>
</tr>
<tr>
<td>SS-B</td>
<td>deleted</td>
<td>neutral</td>
<td>100%</td>
<td>3.92</td>
<td>4.17</td>
<td>3.67</td>
<td>One expert commented that the neutral-polite is too polite in the given situation</td>
</tr>
<tr>
<td>SS-C</td>
<td>deleted</td>
<td>humble</td>
<td>67%</td>
<td>4.5</td>
<td>4.5</td>
<td>2.83</td>
<td>Not in 100% agreement on key</td>
</tr>
<tr>
<td>SS-D</td>
<td>deleted</td>
<td>humble</td>
<td>100%</td>
<td>4.75</td>
<td>4.5</td>
<td>2.75</td>
<td>One expert commented the stem as not authentic</td>
</tr>
</tbody>
</table>

Table 3.4: Pilot Test Results Summary (Speech Styles, SS)

For the Speech Acts, a few experts commented that use of the word “advising” (to name a type of speech act) was confusing. The distinction between “advising” (that the speaker advises the interlocutor to do something) and “suggesting” (that the speaker
suggests that both the speaker and the interlocutor do something together) stemmed from the distinction made in van Ek and Trim (1991), but was not transparent enough for the experts. For this reason, all the “advising” items, which belong to a less frequently observed speech act category in Itomitsu’s (2008) study, were eliminated. This decision was also motivated by the fact that the previous studies on testing speech acts tend to focus on three speech acts (apology, request, and refusal) at a time (Röver 2005; Hudson, Detmer, and Brown 1995), and that it may also be more meaningful to focus on fewer speech acts types to see more of the range of linguistic repertoire the test takers know for a given type.

In addition to the minor changes noted above, the following changes were made: (1) the questions have been formatted as “What would NAME probably say to the/his/her INTERLOCUTOR ROLE?” instead of “What would she (or he) probably say?”; (2) for items involving telephone conversation, the main interlocutor (whose line is to be filled) is placed on the left, and the interlocutor’s image is scaled smaller and placed on the right, to depict the physical distance between the two.

Table 3.5 below summarizes the average scores in importance, authenticity, and difficulty ratings of the selected items in each section:
Table 3.5: Average Scores of the Expert Ratings in Sections

<table>
<thead>
<tr>
<th>Section</th>
<th>Importance</th>
<th>Authenticity</th>
<th>Ease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>4.705</td>
<td>4.463</td>
<td>3.481</td>
</tr>
<tr>
<td>Routines</td>
<td>4.604</td>
<td>4.661</td>
<td>3.634</td>
</tr>
<tr>
<td>Speech Acts</td>
<td>4.819</td>
<td>4.714</td>
<td>3.424</td>
</tr>
<tr>
<td>Speech Styles</td>
<td>4.704</td>
<td>4.763</td>
<td>3.120</td>
</tr>
</tbody>
</table>

Table 3.5 shows a consensus among the experts on all items as highly important for a Japanese major to know, and the situations and language use in the items are natural and authentic. The Routines section seems to be regarded as slightly easier, while the Speech Styles section as slightly more difficult than the other sections.

The pilot test process incorporated expert judgment to ensure the content validity and authenticity of each test item. The pilot test also enabled the test developer to select items with virtually 100% agreement among the experts on the key options. The selected 48 test items, with minor revisions, are thus of higher quality. The pilot stage is indispensable if we are to obtain meaningful results in the operationalized test.
Chapter 4: Methods

This chapter describes the participants in this study, materials used for the study, and data collection procedures.

4.1. Participants

Through the researchers’ personal connections, solicitation e-mail messages were sent to learners of Japanese in four different universities in the midwest U.S. (states of Kansas, Missouri, and Ohio). Students in their second year of Japanese or above were recruited. Three of the four universities offer both a major and a minor in Japanese, and the other, a major in Asian Studies. See Appendix E for the solicitation e-mail message.

Data collection was conducted in two phases. The first phase was carried out in Fall Quarter of 2008, where the majority of data was collected from three different universities. During the second phase, conducted in the Spring Quarter of 2009, thirteen additional cases with proficiency rating data were collected.

Participants were provided with a link to a web site that was made available through a web survey program Survey Gizmo (www.surveygizmo.com). On the very first page, the solicitation e-mail message was shown again, and the participants indicated their consent to participate in the study by clicking a radio button at the bottom of the page. The site was designed so that the participants could not proceed to the next page.
unless they clicked the “consent to participate” button. $10 was given to each participant as a token of appreciation for their participation, either by post or handing over cash/check in person.

Overall 110 JFL learners participated in the study. Tables 4.1 and 4.2 below summarize the numbers of participants from the four universities in the two phases of data collection:

<table>
<thead>
<tr>
<th></th>
<th>University A</th>
<th>University B</th>
<th>University C</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd Level</td>
<td>13</td>
<td>19</td>
<td>9</td>
<td>41</td>
</tr>
<tr>
<td>2nd/3rd Intensive</td>
<td>9</td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>3rd Level</td>
<td>9</td>
<td>16</td>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>4th Level</td>
<td>7</td>
<td>7</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>5th Level</td>
<td>8</td>
<td></td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Not currently enrolled</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL</td>
<td>46</td>
<td>42</td>
<td>17</td>
<td>105</td>
</tr>
</tbody>
</table>

Table 4.1: Participants during Phase One (Fall 2008)

<table>
<thead>
<tr>
<th></th>
<th>University A</th>
<th>University D</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Level</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2nd Level</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>2nd/3rd Intensive</td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>3rd Level</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>4th Level</td>
<td>2</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>5th Level</td>
<td>4</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Not currently enrolled</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>9</td>
<td>5</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 4.2: Participants during Phase Two (Spring 2009)

32 2nd/3rd year is an intensive track, meeting 10 hours per week to cover the 2nd and 3rd level materials in one academic year.
Note that all nine participants from University A in Phase Two are repeaters: they took
the online test during Phase One, and took the test again after proficiency rating
interviews in Phase Two. Thus, the total number of students is 110, and total number of
cases is 119.

Japanese language programs at these four universities differ in terms of the local
conditions and instructional practices. The first of the four universities (“University A”)
is a large state university with a language requirement, which offers smaller class sizes
taught mainly by lecturers and teaching assistants. The second university (“University
B”) is another large state university, with larger class sizes taught by professors and
teaching assistants. The third university (“University C”) is a private university with no
language requirements, where small-size classes are taught by instructors. The fourth
university (“University D”) offers courses in Japanese as part of Asian Studies
Major/Minor, taught by professors. Universities A and C use the same textbook series,
although the pace, class hours, and amount of coverage in each level differ significantly.

4.2. Materials

4.2.1. The Questionnaire (“About You”)

The first part of the online material involves questions on learners’ background
information. See Appendix F for the sample questions included in this section.
The average age of the 110 participants was 21.65 (minimum 18 years old, and maximum 40 years old). The majority of the participants (89.1%) were English-native, and 38.2% of the participants are Japanese majors. Most of them (88.2%) had never lived in Japan for longer than 6 months.

Some of the important background information is summarized in Table 8, following:
<table>
<thead>
<tr>
<th>categories</th>
<th>subcategories</th>
<th>N</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>58</td>
<td>52.7%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>52</td>
<td>47.3%</td>
</tr>
<tr>
<td>Native Language</td>
<td>English</td>
<td>98</td>
<td>89.1%</td>
</tr>
<tr>
<td></td>
<td>Mandarin</td>
<td>5</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>Cantonese</td>
<td>2</td>
<td>1.8%</td>
</tr>
<tr>
<td></td>
<td>others (Arabic, Russian, Spanish)</td>
<td>5</td>
<td>4.5%</td>
</tr>
<tr>
<td>Age</td>
<td>18 years old</td>
<td>9</td>
<td>8.2%</td>
</tr>
<tr>
<td></td>
<td>19 years old</td>
<td>21</td>
<td>19.1%</td>
</tr>
<tr>
<td></td>
<td>20 years old</td>
<td>20</td>
<td>18.2%</td>
</tr>
<tr>
<td></td>
<td>21 years old</td>
<td>18</td>
<td>16.4%</td>
</tr>
<tr>
<td></td>
<td>22 years old</td>
<td>13</td>
<td>11.8%</td>
</tr>
<tr>
<td></td>
<td>23-25 years old</td>
<td>15</td>
<td>13.6%</td>
</tr>
<tr>
<td></td>
<td>26 years old or older</td>
<td>14</td>
<td>12.7%</td>
</tr>
<tr>
<td>Academic Standing</td>
<td>Freshman</td>
<td>11</td>
<td>10.0%</td>
</tr>
<tr>
<td></td>
<td>Sophomore</td>
<td>21</td>
<td>19.1%</td>
</tr>
<tr>
<td></td>
<td>Junior</td>
<td>28</td>
<td>25.5%</td>
</tr>
<tr>
<td></td>
<td>Senior</td>
<td>32</td>
<td>29.1%</td>
</tr>
<tr>
<td></td>
<td>Graduate</td>
<td>18</td>
<td>16.4%</td>
</tr>
<tr>
<td>Major</td>
<td>Japanese</td>
<td>42</td>
<td>38.2%</td>
</tr>
<tr>
<td></td>
<td>Business and Economics</td>
<td>12</td>
<td>10.9%</td>
</tr>
<tr>
<td></td>
<td>Language, Linguistics, Literature</td>
<td>8</td>
<td>7.3%</td>
</tr>
<tr>
<td></td>
<td>Undeclared/Undecided</td>
<td>5</td>
<td>4.5%</td>
</tr>
<tr>
<td></td>
<td>others</td>
<td>38</td>
<td>34.5%</td>
</tr>
</tbody>
</table>

Table 4.3: Summary of Learners’ Background Information
Table 4.3 continued

<table>
<thead>
<tr>
<th>Length of Study</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year or less</td>
<td>16</td>
<td>14.5%</td>
</tr>
<tr>
<td>2 years</td>
<td>36</td>
<td>32.7%</td>
</tr>
<tr>
<td>3 years</td>
<td>23</td>
<td>20.9%</td>
</tr>
<tr>
<td>4 years</td>
<td>15</td>
<td>13.6%</td>
</tr>
<tr>
<td>5 years</td>
<td>14</td>
<td>12.7%</td>
</tr>
<tr>
<td>6 years or more</td>
<td>6</td>
<td>5.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lived in Japan for 6 months+?</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>97</td>
<td>88.2%</td>
</tr>
<tr>
<td>Yes – 1 year or less</td>
<td>8</td>
<td>7.3%</td>
</tr>
<tr>
<td>Yes – more than 1 year</td>
<td>5</td>
<td>4.5%</td>
</tr>
</tbody>
</table>

4.2.2. The Operationalized Test (‘‘Test of Pragmatics’’)

The main material used for this dissertation study was the operationalized version of the test, a 48-item multiple-choice test, a revised and improved version of the pilot test described in the previous chapter. The test consists of four sections (Grammar, Routines, Speech Acts, and Speech Styles), each containing twelve items. Each item contained a picture, sound, and texts written in English and in Japanese (with furigana on every kanji). See Appendix F for sample test materials used for the study.

The test was self-paced, that is, learners can take the time necessary to complete the test. The website provided an option of saving the test in progress by bookmarking the website, and returning later to complete the test. Most of the test takers seemed to complete the test in one continuous session, although a few did not seem to return to the site for a couple of days. Of all the 119 cases, 7 cases took more than 5 hours to complete the test (the maximum taking over 193 hours, most likely returning to the test after a
96

Excluding these 7 cases, the average time to completion (including the “About You” section) was 1 hour 7 minutes 25 seconds, the minimum 20 minutes 45 seconds, and the maximum 4 hours 28 minutes 2 seconds. 10 out of the 112 cases took more than two hours.

The test takers were asked to answer all questions even when they were not sure, and were informed that there was no penalty for guessing. They were instructed to answer the questions on their own, and not to consult textbooks or dictionarys. They were also asked not to ask their teacher or a native speaker of Japanese for help, and to think of the conversations in the test items as normal, everyday talk between normal everyday people (i.e. not between people being extremely rude, sarcastic, funny, etc.).

4.2.3 Proficiency Rating Data

In order to investigate the relationship between the test scores and proficiency rating based on the ACTFL Guidelines (Research Question 4), data from interview tests were also collected in both Phase One and Phase Two of the study. All participants received the proficiency ratings within one month prior to or after they participated in the online multiple-choice test. A total of 50 ratings were collected. The participant data are summarized below.

---

33 There was no way of telling whether the test takers took breaks in between, or spent the entire amount of time on the test without breaks.
<table>
<thead>
<tr>
<th></th>
<th>University A (ACTFL OPI)</th>
<th>University B (Interview Test)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced High</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Advanced Mid</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Advanced Low</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Intermediate Mid</td>
<td>9</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>Intermediate Low</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Novice High</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Novice Mid</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>TOTAL</td>
<td>5</td>
<td>33</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 4.4: Participants with Proficiency Rating (Phase One)

<table>
<thead>
<tr>
<th></th>
<th>University A (Interview-ILR)</th>
<th>University D (Institutional OPI)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced High</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Advanced Mid</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Advanced Low</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Intermediate High</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Intermediate Mid</td>
<td>3</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Intermediate Low</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Novice High</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Novice Mid</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>TOTAL</td>
<td>9</td>
<td>3</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 4.5: Participants with Proficiency Rating (Phase Two)

There were four types of interviews through which the proficiency ratings were obtained.

“ACTFL OPI” refers to the official ACTFL telephonic OPI conducted by certified ACTFL OPI testers. Five participants from University A took the official ACTFL OPI as part of a scholarship application. “Institutional OPI” interviews refer to those conducted by a certified ACTFL OPI tester, who interviewed students of the same academic
institution. The institutional OPIs were conducted as part of course/program evaluation. “Interview Test” refers to those conducted by an individual who completed the ACTFL OPI training and who interviewed the students of the same institution. The interview tests were conducted as part of program evaluation. “Interviews-ILR” refers to those conducted by the researcher who underwent training and was certified as OPI Interviewer at the Defense Language Institute, where the proficiency scales developed by the Interagency Language Roundtable (ILR) are used. The ACTFL OPI was in fact developed as an academic version of the ILR-based OPI (Yoffe 1997), and the interviewer studied and followed the ACTFL OPI format and procedures. The Institutional OPIs, Interview Tests, and Interviews-ILR were very similar to the official ACTFL OPI except that the interviewers know the interviewees as they were teaching and studying at the same institution, and there was no second independent rater. Interrater reliability is not addressed for the interview tests, and the test is referred to here as "interview test", rather than as "OPI," to distinguish them from those recognized by the ACTFL as official OPI. All ratings were based on the same criteria, the ACTFL Guidelines.
Chapter 5: Results

This chapter summarizes the findings and statistical analyses of the obtained data. SPSS 17 for Mac was used for the statistical analyses.

5.1. General Test Characteristics

This section provides an overview of the data from the 110 participants, as 119 cases (there were 9 repeaters). The multiple-choice answers were scored by giving the correct answers a score of 1, and incorrect answers 0. All participants answered all questions, except for three items, where there were no answers given. Such cases receive a score of 0. The maximum possible point total was 48.

Table 5.1 below is a descriptive statistics of the test and its sections from the entire 119 cases:
Table 5.1: Descriptive Statistics of the Test and Sections

<table>
<thead>
<tr>
<th>section</th>
<th>N</th>
<th>mean</th>
<th>SD</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>119</td>
<td>7.39</td>
<td>2.688</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Routines</td>
<td>119</td>
<td>8.08</td>
<td>2.293</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Speech Acts</td>
<td>119</td>
<td>7.35</td>
<td>2.698</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Speech Styles</td>
<td>119</td>
<td>6.66</td>
<td>2.865</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>TOTAL</td>
<td>119</td>
<td>29.44</td>
<td>9.030</td>
<td>10</td>
<td>48</td>
</tr>
</tbody>
</table>

The means in each section are similar, although the Routines section was slightly easier, and the Speech Styles section slightly more difficult than the other sections. Standard deviations show that the spread of the test scores are slightly wider in the Speech Styles section, and slightly narrower in the Routines section. Two participants scored 100% (48 points).

5.2. Reliability of the Test and Sections

Research Question 1 concerns the reliability of the test, a prerequisite for validation studies of the test. Table 5.2 below is a summary of Cronbach’s alpha of internal consistency coefficients for the test and its sections from the entire 119 cases:
<table>
<thead>
<tr>
<th>section</th>
<th>Cronbach’s alpha</th>
<th>N of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>.686</td>
<td>12</td>
</tr>
<tr>
<td>Routines</td>
<td>.640</td>
<td>12</td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.709</td>
<td>12</td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.727</td>
<td>12</td>
</tr>
<tr>
<td>TOTAL</td>
<td>.893</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 5.2: Cronbach’s Alpha for the Test and Sections

Cronbach’s alpha ranges from 0 to 1.00, and values close to 1.00 indicate high consistency. In general, longer tests produce higher reliabilities. The test’s overall alpha was .893, which is interpreted as satisfactory. Wells and Wollack (2003) explains that “professionally developed high-stakes standardized tests should have internal consistency coefficients of at least .90. Lower-stakes standardized tests should have internal consistencies of at least .80 or .85. For a classroom exam, it is desirable to have a reliability coefficient of .70 or higher.”

Although the overall reliability of the test was close to the standard desired for professionally developed high-stakes tests, the Routines section showed lower reliability. To further investigate the reliability and quality of each item, the following four statistics were gathered for each item, following item analysis procedures after Brown (2003): (1) Item Facility, (2) Item Discrimination, (3) Corrected Item-Total Correlation, and (4) Cronbach’s alpha when the item is deleted. The Item Facility (IF) refers to the proportion of subjects who answered a particular item correctly; thus IF of 1.00 means all subjects answer the item correctly. Item Discrimination (ID) is the difference between the IF for
the upper group (top 1/3, 40 cases) and the IF for the lower group (bottom 1/3, 40 cases).

An item is considered to be discriminating if the “better” students tend to answer the item correctly while the “poorer” students tend to respond incorrectly (Wells and Wollack 2003). Brown (1996) suggests that items with an ID of .40 and up can be considered as very good items; .30 to 39 as reasonably good, .20 to .29 as marginal, and .19 and below as poor items, to be rejected or improved by revision. Corrected Item-Total Correlation refers to the point-biserial correlation between the right/wrong scores that students receive on a given item and the total scores that they receive when summing up their scores across the remaining items, without that item considered part of the scale. A value of 0.20 or better is preferred for classroom tests (Wells and Wollack 2003). Finally, a reliable item positively contributes to the overall reliability, thus its deletion results in lowering Cronbach’s alpha.

The four statistics, IF, ID, Corrected Item-Total Correlation, and Cronbach’s Alpha if Item Deleted are summarized separately for each section of the test in Tables 5.3, 5.4, 5.5, and 5.6 below:
In the Grammar section, the IFs show that the actual difficulties were showing a trend similar to the difficulty predicted by the experts consulted, except that GR01 and GR07 were more difficult than they predicted. (Recall that the test items were ordered according to the predicted difficulty level by the experts during the pilot study.) All items display a satisfactory ID, as well as good reliability; removing any of the items results in equal or lower reliability.

### Table 5.3: Item Statistics for the Grammar Section

<table>
<thead>
<tr>
<th>Grammar item #</th>
<th>IF</th>
<th>ID</th>
<th>Corrected Item-Total Correlation</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR01</td>
<td>.55</td>
<td>.38</td>
<td>.267</td>
<td>.892</td>
</tr>
<tr>
<td>GR02</td>
<td>.71</td>
<td>.40</td>
<td>.326</td>
<td>.891</td>
</tr>
<tr>
<td>GR03</td>
<td>.77</td>
<td>.45</td>
<td>.464</td>
<td>.890</td>
</tr>
<tr>
<td>GR04</td>
<td>.79</td>
<td>.50</td>
<td>.511</td>
<td>.889</td>
</tr>
<tr>
<td>GR05</td>
<td>.63</td>
<td>.58</td>
<td>.466</td>
<td>.889</td>
</tr>
<tr>
<td>GR06</td>
<td>.66</td>
<td>.33</td>
<td>.216</td>
<td>.893</td>
</tr>
<tr>
<td>GR07</td>
<td>.40</td>
<td>.23</td>
<td>.224</td>
<td>.893</td>
</tr>
<tr>
<td>GR08</td>
<td>.73</td>
<td>.25</td>
<td>.193</td>
<td>.893</td>
</tr>
<tr>
<td>GR09</td>
<td>.65</td>
<td>.53</td>
<td>.452</td>
<td>.890</td>
</tr>
<tr>
<td>GR10</td>
<td>.56</td>
<td>.65</td>
<td>.463</td>
<td>.889</td>
</tr>
<tr>
<td>GR11</td>
<td>.44</td>
<td>.60</td>
<td>.479</td>
<td>.889</td>
</tr>
<tr>
<td>GR12</td>
<td>.50</td>
<td>.28</td>
<td>.282</td>
<td>.892</td>
</tr>
</tbody>
</table>
In the Routines section, items RT01 and RT03 were the easiest items of the entire test; RT12 turned out to be the most difficult item. The above table indicates that RT06 and RT10 are problematic items, indicated by a lower or negative ID. Also, removing these two items would actually improve the overall reliability slightly. Overall reliability of the section would have been better if the quality of these two items had been better.
<table>
<thead>
<tr>
<th>Speech Acts item #</th>
<th>IF</th>
<th>ID</th>
<th>Corrected Item-Total Correlation</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA01 (request)</td>
<td>.58</td>
<td>.35</td>
<td>.234</td>
<td>.893</td>
</tr>
<tr>
<td>SA02 (request)</td>
<td>.84</td>
<td>.23</td>
<td>.289</td>
<td>.892</td>
</tr>
<tr>
<td>SA03 (request)</td>
<td>.71</td>
<td>.48</td>
<td>.431</td>
<td>.890</td>
</tr>
<tr>
<td>SA04 (request)</td>
<td>.26</td>
<td>.55</td>
<td>.536</td>
<td>.889</td>
</tr>
<tr>
<td>SA05 (offer)</td>
<td>.79</td>
<td>.38</td>
<td>.407</td>
<td>.890</td>
</tr>
<tr>
<td>SA06 (offer)</td>
<td>.68</td>
<td>.40</td>
<td>.341</td>
<td>.891</td>
</tr>
<tr>
<td>SA07 (offer)</td>
<td>.48</td>
<td>.58</td>
<td>.460</td>
<td>.890</td>
</tr>
<tr>
<td>SA08 (offer)</td>
<td>.50</td>
<td>.28</td>
<td>.241</td>
<td>.893</td>
</tr>
<tr>
<td>SA09 (suggestions)</td>
<td>.67</td>
<td>.53</td>
<td>.491</td>
<td>.889</td>
</tr>
<tr>
<td>SA10 (suggestions)</td>
<td>.83</td>
<td>.23</td>
<td>.249</td>
<td>.892</td>
</tr>
<tr>
<td>SA11 (suggestions)</td>
<td>.45</td>
<td>.73</td>
<td>.598</td>
<td>.887</td>
</tr>
<tr>
<td>SA12 (suggestions)</td>
<td>.57</td>
<td>.65</td>
<td>.533</td>
<td>.888</td>
</tr>
</tbody>
</table>

Table 5.5: Item Statistics for the Speech Acts Section

In the Speech Acts section, all items show high reliability, ID, and Item-Total Correlation. There seems to be a slight gap between the difficulty predicted by the experts and the actual difficulty among the learners. The average item facility indices of the three types of speech acts were very similar (.60 for requests, .61 for offers, and .63 for suggestions).
<table>
<thead>
<tr>
<th>Speech Styles item #</th>
<th>IF</th>
<th>ID</th>
<th>Corrected Item-Total Correlation</th>
<th>Alpha if Item Deleted</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS01 (humble)</td>
<td>.78</td>
<td>.38</td>
<td>.368</td>
<td>.891</td>
</tr>
<tr>
<td>SS02 (humble)</td>
<td>.66</td>
<td>.43</td>
<td>.357</td>
<td>.891</td>
</tr>
<tr>
<td>SS03 (honorific)</td>
<td>.74</td>
<td>.60</td>
<td>.496</td>
<td>.889</td>
</tr>
<tr>
<td>SS04 (humble)</td>
<td>.39</td>
<td>.73</td>
<td>.541</td>
<td>.888</td>
</tr>
<tr>
<td>SS05 (humble)</td>
<td>.49</td>
<td>.43</td>
<td>.332</td>
<td>.891</td>
</tr>
<tr>
<td>SS06 (humble)</td>
<td>.44</td>
<td>.60</td>
<td>.513</td>
<td>.889</td>
</tr>
<tr>
<td>SS07 (humble)</td>
<td>.48</td>
<td>.48</td>
<td>.368</td>
<td>.891</td>
</tr>
<tr>
<td>SS08 (honorific)</td>
<td>.45</td>
<td>.58</td>
<td>.445</td>
<td>.890</td>
</tr>
<tr>
<td>SS09 (honorific)</td>
<td>.72</td>
<td>.43</td>
<td>.384</td>
<td>.891</td>
</tr>
<tr>
<td>SS10 (honorific)</td>
<td>.68</td>
<td>.28</td>
<td>.196</td>
<td>.893</td>
</tr>
<tr>
<td>SS11 (humble)</td>
<td>.46</td>
<td>.40</td>
<td>.296</td>
<td>.892</td>
</tr>
<tr>
<td>SS12 (honorific)</td>
<td>.37</td>
<td>.35</td>
<td>.316</td>
<td>.892</td>
</tr>
</tbody>
</table>

Table 5.6: Item Statistics for the Speech Styles Section

The items in the Speech Styles section show high reliability, ID, and Item-Total Correlation as well. Items SS09 and SS10 seem to be easier than the experts’ prediction. The average item facility indices of the two types of speech styles were very similar (.53 for humble-polite, .59 for honorific-polite).

5.3. Relationship between the Test Scores and Length of Study/Exposure

Research Question 2 concerns group differences among learners in different lengths of study/exposure. It is hypothesized that the longer the learners study, the more instances of language use they are exposed to, and therefore the higher the test scores will be. One-way ANOVA was conducted to see if there is a difference in total scores among students who are studying the language for different period of time (regardless of the
level of course in which they are enrolled). A correlation analysis between the test scores and length of study/exposure was also conducted.

Of all the 119 cases, 9 cases of repeated measures in Phase Two were excluded from the analysis\textsuperscript{34}. Remaining 110 subjects are divided into four groups: those who has been studying Japanese for 1 year or less, 2 years, 3 years, 4 years, and 5 years or more.

Table 5.7 below is a summary of descriptive statistics, followed by a box plot, Figure 5.1:

\textsuperscript{34} There are nine subjects who too the test twice, both in Phase One and Phase Two. The results from Phase Two were excluded in this analysis.
Table 5.7: Descriptive Statistics of Total Scores by Length of Study/Exposure

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year or less</td>
<td>16</td>
<td>24.81</td>
<td>6.645</td>
<td>14</td>
<td>38</td>
</tr>
<tr>
<td>2 years</td>
<td>36</td>
<td>26.28</td>
<td>7.770</td>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>3 years</td>
<td>23</td>
<td>29.91</td>
<td>8.913</td>
<td>14</td>
<td>46</td>
</tr>
<tr>
<td>4 years</td>
<td>15</td>
<td>30.20</td>
<td>11.047</td>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>5 years or more</td>
<td>20</td>
<td>35.25</td>
<td>7.887</td>
<td>21</td>
<td>48</td>
</tr>
<tr>
<td>TOTAL</td>
<td>110</td>
<td>28.99</td>
<td>8.977</td>
<td>10</td>
<td>48</td>
</tr>
</tbody>
</table>

Figure 5.1: Box-Plot of Total Scores by Length of Study/Exposure

The boxes in the figure indicate the middle 50% of the data: the upper edge of the box indicates the 75th percentile of the data set, and the lower edge, the 25th percentile.
The thick line in the box indicates the median value of the data. The ends of the vertical lines or “whiskers” indicate the minimum and maximum data values.

The normality assumption was satisfied for all levels (Shapiro-Wilk p>.469), and homogeneity of variance is satisfied (Levene’s statistic F(4, 105)=1.755, p=.143). One-way ANOVA showed that the main effect of groups on test scores was significant, F(4, 105)=4.849, P=.001. Post-hoc Tukey HSD tests reveal that the differences were significant between “1 year or less” and “5 years and more”, and between “2 years” and “5 years and more”. It was also found that the total scores were significantly correlated with the length of study (Pearson r=.380, p=.000). The positive correlation between the test scores and length of study can be taken as evidence in support of the construct validity of the test, although the observed correlation is not very strong. Also, significant differences in the total scores are found only between those who were studying Japanese more than five years and those who were studying two years or less.

Relationships between the section scores and length of study were also investigated. The following Table 5.8 is a summary of descriptive statistics, followed by box-plots in Figure 5.2:
<table>
<thead>
<tr>
<th>Sections</th>
<th>Length of study</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1 year or less</td>
<td>16</td>
<td>5.88</td>
<td>2.156</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2 years</td>
<td>36</td>
<td>7.00</td>
<td>2.484</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3 years</td>
<td>23</td>
<td>7.17</td>
<td>2.887</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>15</td>
<td>7.67</td>
<td>3.063</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5 years or more</td>
<td>20</td>
<td>8.90</td>
<td>2.315</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>110</td>
<td>7.31</td>
<td>2.695</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Routines</td>
<td>1 year or less</td>
<td>16</td>
<td>7.31</td>
<td>2.243</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2 years</td>
<td>36</td>
<td>7.53</td>
<td>2.443</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3 years</td>
<td>23</td>
<td>7.61</td>
<td>1.948</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>15</td>
<td>8.07</td>
<td>2.658</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>5 years or more</td>
<td>20</td>
<td>9.30</td>
<td>1.625</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>110</td>
<td>7.91</td>
<td>2.285</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Speech Acts</td>
<td>1 year or less</td>
<td>16</td>
<td>5.75</td>
<td>2.408</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2 years</td>
<td>36</td>
<td>6.14</td>
<td>1.759</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3 years</td>
<td>23</td>
<td>8.00</td>
<td>2.860</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>15</td>
<td>7.60</td>
<td>2.874</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5 years or more</td>
<td>20</td>
<td>9.10</td>
<td>2.426</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>110</td>
<td>7.21</td>
<td>2.651</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Speech Styles</td>
<td>1 year or less</td>
<td>16</td>
<td>5.88</td>
<td>1.746</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2 years</td>
<td>36</td>
<td>5.61</td>
<td>2.901</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3 years</td>
<td>23</td>
<td>7.13</td>
<td>2.668</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4 years</td>
<td>15</td>
<td>6.87</td>
<td>3.441</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5 years or more</td>
<td>20</td>
<td>7.95</td>
<td>2.625</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>110</td>
<td>6.56</td>
<td>2.843</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5.8: Descriptive Statistics of Section Scores by Length of Study/Exposure
ANOVA was conducted for each section. Again, significant differences in the section scores were found only between those who were studying Japanese “5 years or more” and those who were studying two years and/or one year or less.

In the Grammar section, the normality assumption was satisfied (Shapiro-Wilk p>.162), and the homogeneity of variance assumption was satisfied, with Levene’s statistic $F(4,105)=.720$, $p=.580$. One-way ANOVA indicates that the main effect of
groups on test scores is significant, \( F(4, 105)=3.338, P=.013 \). Post-hoc Tukey HSD tests reveals that the difference is significant between the “1 year or less” and “5 years and more” \( (p=.006) \).

In the Routines section, the normality assumption was satisfied (Shapiro-Wilk \( p>.311 \)) except for the “4 years” group (Shapiro-Wilk \( p=.017 \)). The homogeneity of variance assumption was satisfied, Levene’s statistic \( F(4, 105)=1.784; p=.137 \). One-way ANOVA indicates the main effect of groups on test scores was significant, \( F(4, 105)=2.644, p=.038 \). Post-hoc Tukey HSD tests revealed that the difference was significant between the “2 years” and “5 years and more” \( (p=.040) \).

In the Speech Acts section, the normality assumption was satisfied (Shapiro-Wilk \( p>.063 \)), but the assumption of homogeneity of variance was violated, Levene’s statistic \( F(4, 105)=3.591, p=.009 \). One-way ANOVA indicates the main effect of groups on test scores was significant, Brown-Forsythe \( F(4, 76.074)=6.423, p=.000 \). Post-hoc Games-Howell tests revealed that the differences were significant between the “1 year or less” and “5 years and more” \( (p=.002) \), and between “2 years” and “5 years and more” \( (p=.000) \).

In the Speech Styles section, the normality assumption was satisfied (Shapiro-Wilk \( p>.082 \)), except for “1 year or less” (Shapiro-Wilk \( p=.015 \)), and the homogeneity of variance assumption was satisfied, Levene’s statistic \( F(4, 105)=2.344, p=.059 \). One-way ANOVA indicated the main effect of groups on test scores was significant, \( F(4, 105)=2.893, p=.026 \). Post-hoc Tukey HSD tests revealed that the difference was significant between the “2 years” and “5 years and more” \( (p=.024) \).
The following Figure 5.3 is a graph of mean plots for the four sections:

![Figure 5.3: Mean Plots of the Section Scores by Length of Study/Exposure](image)

The mean plots indicates that different sections show different tendencies: while for the Grammar and the Routines sections the increase in score is constantly positive as learners study longer, this is not the case for the Speech Acts and Speech Styles Sections. The decreases in the Speech Acts and Speech Styles scores at “4 years” are
counterintuitive. This implies that length of study is not as good of an indicator to learners’ grammatical and pragmatic knowledge. Perhaps, this is owing to differences in quality of experience depending on various factors not identifiable at this point.

5.4. Relationship between the Test Scores and Levels of Instruction

Research Question 2 also concerns group differences among learners in different levels of instruction. It is hypothesized that the overall scores increase as the learners advance to higher levels of study, because they have more exposures to language use and gain more pragmatic knowledge. One-way ANOVA was conducted to see if there is a difference in total scores among students who are in different levels of instruction. A correlation analysis between the test scores and levels of instruction was also conducted.

Of all the 119 cases, 1 case of currently not enrolled, and 9 cases of repeated measures in Phase Two were excluded from the analysis. Also, 9 cases from the 2\textsuperscript{nd}/3\textsuperscript{rd} intensive track were removed, so that the analyses were only of data from learners in the regular track. The remaining 100 subjects were divided into four groups: those in 2\textsuperscript{nd}, 3\textsuperscript{rd}, 4\textsuperscript{th}, and 5\textsuperscript{th} levels of instruction.

Table 5.9 below is a summary of the actual number of subjects from four different universities:
Table 5.9: Learners in Different Levels of Instruction from the Four Universities

<table>
<thead>
<tr>
<th>Level</th>
<th>University A</th>
<th>University B</th>
<th>University C</th>
<th>University D</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd Level</td>
<td>13</td>
<td>19</td>
<td>9</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td>3rd Level</td>
<td>9</td>
<td>16</td>
<td>5</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>4th Level</td>
<td>7</td>
<td>7</td>
<td></td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>5th Level</td>
<td>8</td>
<td></td>
<td>2</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>TOTAL</td>
<td>37</td>
<td>42</td>
<td>17</td>
<td>5</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5.10 below is a summary of descriptive statistics, followed by a box plot Figure 5.4. Again the maximum possible total score is 48.

<table>
<thead>
<tr>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2nd Level</td>
<td>41</td>
<td>23.44</td>
<td>6.819</td>
<td>10</td>
<td>38</td>
</tr>
<tr>
<td>3rd Level</td>
<td>35</td>
<td>28.14</td>
<td>7.964</td>
<td>10</td>
<td>47</td>
</tr>
<tr>
<td>4th Level</td>
<td>14</td>
<td>36.71</td>
<td>4.177</td>
<td>30</td>
<td>43</td>
</tr>
<tr>
<td>5th Level</td>
<td>10</td>
<td>42.80</td>
<td>4.022</td>
<td>38</td>
<td>48</td>
</tr>
<tr>
<td>TOTAL</td>
<td>100</td>
<td>28.88</td>
<td>9.221</td>
<td>10</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 5.10: Descriptive Statistics of Total Scores by Levels of Instruction
The normality assumption was satisfied for all levels (Shapiro-Wilk p>.073), and homogeneity of variance was satisfied (Levene’s statistic F(3, 96)=2.455, p=.068). One-way ANOVA showed that the main effect of groups on test scores was significant, F(3, 96)=29.382, P=.000. Post-hoc Tukey HSD tests revealed that the differences were significant between all levels (p≤.017) except between the 4th and 5th levels (p=.138). The results show that the total scores are different among learners in different levels of instruction. It was also found that the total score correlated significantly with the levels of instruction (Pearson r=.685, p=.000). Together, the results can be taken as another piece of evidence in support of the construct validity of the test, given that learners are
provided additional pragmatic knowledge as they advance to progressively higher levels of instruction.

To further investigate the group differences in different components of communicative language ability, relationships between section scores and levels of instruction were also investigated. Table 5.11 below is a summary of descriptive statistics, followed by box-plots in Figure 5.5:
<table>
<thead>
<tr>
<th>Section</th>
<th>Levels</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammar</strong></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Level</td>
<td>41</td>
<td>6.00</td>
<td>2.42</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Level</td>
<td>35</td>
<td>7.23</td>
<td>2.59</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>14</td>
<td>9.07</td>
<td>1.64</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>10</td>
<td>10.90</td>
<td>0.99</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>100</td>
<td>7.35</td>
<td>2.75</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td><strong>Routines</strong></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Level</td>
<td>41</td>
<td>6.66</td>
<td>2.09</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Level</td>
<td>35</td>
<td>7.94</td>
<td>2.39</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>14</td>
<td>9.21</td>
<td>1.12</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>5&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>10</td>
<td>10.20</td>
<td>1.13</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>100</td>
<td>7.82</td>
<td>2.33</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td><strong>Speech Acts</strong></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Level</td>
<td>41</td>
<td>5.54</td>
<td>1.76</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Level</td>
<td>35</td>
<td>6.94</td>
<td>2.41</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>14</td>
<td>10.00</td>
<td>1.41</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>10</td>
<td>11.00</td>
<td>1.15</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>100</td>
<td>7.20</td>
<td>2.71</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td><strong>Speech Styles</strong></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Level</td>
<td>41</td>
<td>5.24</td>
<td>2.48</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Level</td>
<td>35</td>
<td>6.03</td>
<td>2.48</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>14</td>
<td>8.43</td>
<td>2.21</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>5&lt;sup&gt;th&lt;/sup&gt; Level</td>
<td>10</td>
<td>10.70</td>
<td>1.42</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>100</td>
<td>6.51</td>
<td>2.91</td>
<td>1</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5.11: Descriptive Statistics of Section Scores by Levels of Instruction
The variances in the box-plots were seemingly narrower in the upper levels, but this was probably due to a ceiling effect (there were only 12 items in each section, and the true variance among upper levels might have been obscured by this).

In the Grammar section, the normality assumption was satisfied except for the 5th Level (Shapiro-Wilk=.713, p=.001), although ANOVA is robust (i.e., insensitive) to the violation of the normality assumption. The homogeneity of variance assumption was

---

35 Dots indicates extreme scores, and numbers indicate participants’ ID numbers.
satisfied, Levene’s statistic F(3, 96)=2.581, p=.058. One-way ANOVA indicates that the main effect of groups on test scores was significant, F(3, 96)=15.291, P=.000. Post-hoc Tukey HSD tests revealed that the differences were significant between all non-adjacent levels (e.g., 2nd Level was different from the 4th and 5th Levels; 3rd Level was different from the 1st and 5th Levels, etc., p=.000) but not between any of the adjacent levels (p=.100 for 2nd/3rd Levels, p=.061 for 3rd/4th Levels, and p=.227 for 4th/5th Levels).

In the Routines section, the normality assumption was satisfied for all levels (Shapiro-Wilk p>.169). The assumption of homogeneity of variance is violated, with Levene’s statistic F(3, 96)=4.027; p=.010; therefore, the Brown-Forsythe F-ratio is reported. One-way ANOVA indicated that the main effect of groups on test scores was significant, Brown-Forsythe F(3, 89.595) =21.981, P=.000. Post-hoc Games-Howell tests revealed that the differences were significant among all non-adjacent levels (p<.001) but not between any of the adjacent levels (p=.073 for 2nd/3rd Levels, .069 for 3rd/4th Levels, and .186 for 4th/5th Levels).

In the Speech Acts section, the normality assumption was satisfied except for the 5th Level (Shapiro-Wilk=.797, p=.014), and the assumption of homogeneity of variance is violated, with Levene’s statistic at F(3, 96)=4.261; p=.007. One-way ANOVA indicates the main effect of groups on test scores is significant, Brown-Forsythe F(3, 81.629) =43.342, P=.000. Post-hoc Games-Howell tests reveal that the differences are significant between all levels (p<.029) except for the 4th/5th Level difference (p=.256).

In the Speech Styles section, the normality assumption was satisfied (Shapiro-Wilk p>.087), and the homogeneity of variance assumption was satisfied, Levene’s
statistic $F(3, 96)=1.110, p=.349$. One-way ANOVA indicates the main effect of groups on test scores was significant, $F(3, 96)=17.964, P=.000$. Post-hoc Tukey HSD tests reveals that the differences were significant between all levels ($p<.010$) except for the 2nd/3rd Level difference ($p=.476$) and for 4th/5th Level difference ($p=.100$).

Figure 5.6 below is a graph of mean plots for the four sections:

![Figure 5.6: Mean Plots of the Section Scores by Levels of Instruction](image)

The above mean plots again suggest that different sections reflect different trajectories: while the Grammar and Routines sections show relatively steady increments
of test scores as the learners advance to the next level of instruction, that is not the case with the Speech Act and Speech Style scores. It seems that for the Speech Acts, the increase of scores is most sizable between the 3rd and 4th Levels. For the Speech Styles, the gain is smallest between the 2nd and 3rd Levels, and the scores increase significantly afterwards.

5.5. Relationship between the Test Scores and Proficiency Ratings

Research Question 3 concerns the relationship between the test scores and learners’ proficiency levels as characterized in terms of the ACTFL Guidelines, measured by the ACTFL OPI or similar interview tests. One-way ANOVA was conducted to see if there is a difference in total scores among students who are at different proficiency levels. Correlation analyses were also conducted.

Fifty cases with proficiency ratings were analyzed. The subjects were divided into five groups: Novice (including Novice-Mid and Novice-High), Intermediate-Low, Intermediate-Mid, Intermediate-High, and Advanced (including Advanced-Low, Advanced-Mid, and Advanced-High).

Table 5.12 below is a summary of descriptive statistics, followed by a box plot, Figure 5.7:
<table>
<thead>
<tr>
<th>Proficiency Rating</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>10</td>
<td>21.30</td>
<td>5.716</td>
<td>16</td>
<td>36</td>
</tr>
<tr>
<td>Intermediate-Low</td>
<td>10</td>
<td>25.80</td>
<td>5.712</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Intermediate-Mid</td>
<td>12</td>
<td>28.50</td>
<td>5.334</td>
<td>23</td>
<td>39</td>
</tr>
<tr>
<td>Intermediate-High</td>
<td>10</td>
<td>36.00</td>
<td>8.069</td>
<td>21</td>
<td>47</td>
</tr>
<tr>
<td>Advanced</td>
<td>8</td>
<td>37.75</td>
<td>6.649</td>
<td>27</td>
<td>48</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50</td>
<td>29.50</td>
<td>8.553</td>
<td>14</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 5.12: Descriptive Statistics of Total Score by Proficiency Rating

Figure 5.7: Box-Plots of Total Scores by Proficiency Rating

---

36 The 10 Novices include 2 Novice-Mid and 8 Novice-High cases. The 8 Advanced include 3 Advanced-Low, 3 Advanced-Mid, and 2 Advanced-High cases.
The normality assumption was satisfied for all levels (Shapiro-Wilk p>.175) except for the Novice group (Shapiro-Wilk p=.005), and homogeneity of variance was satisfied (Levene’s statistic F(4, 45)=.632, p=.642). One-way ANOVA showed that the main effect of groups on test scores was significant, F(4, 45)=11.201, P=.000. Post-hoc Tukey HSD tests revealed that the differences were significant between the Novice and Intermediate-High (p=.000), Novice and Advanced (p=.000); Intermediate-Low and Intermediate-High (p=.007), Intermediate-Low and Advanced (p=.002); and Intermediate-Mid and Advanced (p=.020).

Relationships between the section scores and proficiency rating were also investigated. The following table, Table 5.13, is a summary of descriptive statistics, followed by box-plots (Figure 5.8):
<table>
<thead>
<tr>
<th>Sections</th>
<th>Proficiency Rating</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>Novice</td>
<td>10</td>
<td>5.90</td>
<td>1.912</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Int-Low</td>
<td>10</td>
<td>6.70</td>
<td>2.214</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Int-Mid</td>
<td>12</td>
<td>7.42</td>
<td>1.782</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Int-High</td>
<td>10</td>
<td>9.20</td>
<td>2.150</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>8</td>
<td>9.50</td>
<td>1.773</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>50</td>
<td>7.66</td>
<td>2.335</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Routines</td>
<td>Novice</td>
<td>10</td>
<td>5.70</td>
<td>2.263</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Int-Low</td>
<td>10</td>
<td>7.80</td>
<td>2.440</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Int-Mid</td>
<td>12</td>
<td>7.75</td>
<td>2.137</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Int-High</td>
<td>10</td>
<td>9.20</td>
<td>1.619</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>8</td>
<td>9.13</td>
<td>1.642</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>50</td>
<td>7.86</td>
<td>2.347</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Speech Acts</td>
<td>Novice</td>
<td>10</td>
<td>4.80</td>
<td>1.687</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Int-Low</td>
<td>10</td>
<td>6.40</td>
<td>1.897</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Int-Mid</td>
<td>12</td>
<td>6.67</td>
<td>2.103</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Int-High</td>
<td>10</td>
<td>9.50</td>
<td>2.321</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>8</td>
<td>10.50</td>
<td>2.000</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>50</td>
<td>7.42</td>
<td>2.815</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Speech Styles</td>
<td>Novice</td>
<td>10</td>
<td>4.90</td>
<td>2.283</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Int-Low</td>
<td>10</td>
<td>4.90</td>
<td>1.287</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Int-Mid</td>
<td>12</td>
<td>6.67</td>
<td>1.826</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Int-High</td>
<td>10</td>
<td>8.10</td>
<td>3.348</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Advanced</td>
<td>8</td>
<td>8.63</td>
<td>2.387</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>TOTAL</td>
<td>50</td>
<td>6.56</td>
<td>2.681</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5.13: Descriptive Statistics of Section Scores by Proficiency Rating
In the Grammar section, the normality assumption was satisfied (Shapiro-Wilk $p > .070$) except for the Intermediate-Mid group (Shapiro-Wilk $p = .007$), and the homogeneity of variance assumption was satisfied, Levene’s statistic $F(4, 45) = .306$, $p = .900$. One-way ANOVA indicates the main effect of groups on test scores was significant, $F(4, 45) = 5.878$, $P = .001$. Post-hoc Tukey HSD tests revealed that the difference was significant between Novice and Intermediate-High ($p = .005$), Novice and Advanced ($p = .003$), Intermediate-Low and Advanced ($p = .035$).
In the Routines section, the normality assumption was satisfied (Shapiro-Wilk p>.136) except for the Novice group (Shapiro-Wilk p=.027). The homogeneity of variance assumption was also satisfied, Levene’s statistic F(4, 45)=.834; p=.511. One-way ANOVA indicated that the main effect of groups on test scores was significant, F(4, 45) =4.537, p=.004. Post-hoc Tukey HSD tests reveal that the difference was significant between Novice and Intermediate-High (p=.004) and between Novice and Advanced (p=.009).

In the Speech Acts section, the normality assumption was satisfied (Shapiro-Wilk p>.118) except for the Advanced group (Shapiro-Wilk p=.011). The homogeneity of variance assumption was satisfied, Levene’s statistic F(4, 45)=.369, p=.829. One-way ANOVA indicated the main effect of groups on test scores was significant, F(4, 76.074) =6.423, p=.000. Post-hoc Tukey HSD tests revealed that the differences were significant between the Novice and Intermediate-High (p=.000), Novice and Advanced (p=.000); Intermediate-Low and Intermediate-High (p=.011), Intermediate-Low and Advanced (p=.001); and Intermediate-Mid and Intermediate-High (p=.016).

In the Speech Styles section, the normality assumption was satisfied (Shapiro-Wilk p>.173) except for the Novice group (Shapiro-Wilk p=.005), but the homogeneity of variance assumption was violated, with Levene’s statistic F(4, 45)=4.664, p=.003. One-way ANOVA indicated the main effect of groups on test scores was significant, Brown-Forsythe F(4, 31.800)=5.192, p=.002. Post-hoc Games-Howell tests revealed that the difference was significant between the Novice and Advanced (p=.031), and between Intermediate-Low and Advanced (p=.017).
The following Figure 5.9 depicts mean plots of the section scores by proficiency rating.

![Figure 5.9: Mean Plots of Section Scores by Proficiency Rating](image)

The above mean plots show similar tendencies observed in the mean plots for different levels of instruction: the Grammar section shows steady increment of increase as proficiency rating goes up; the Speech Acts scores increase most noticeably between
Intermediate-Mid and Intermediate-High; and the Speech Styles section does not show much increase in the early stages of proficiency.

In order to further investigate the relationships between the section scores and the proficiency rating, correlation analyses were conducted. Following Kenyon and Tschirner (2000), the ratings from the oral interview test were converted into a numerical scale, as shown in Table 5.14 below:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Point</td>
<td>0.3</td>
<td>0.8</td>
<td>1.1</td>
<td>1.3</td>
<td>1.8</td>
<td>2.1</td>
<td>2.3</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Table 5.14: Conversion Scheme for Proficiency Rating

The above conversion procedure is not free of controversy, however, as “[n]o other rationale for converting the ordinal scale to an interval scale, nor the fractional distances between the intervals, was given” (Yoffe 1997). The ACTFL proficiency scales are exponential, rather than linear; that is, the higher the level, the greater the distance between one rating and the next. Kenyon and Tschirner’s linear conversion fails to capture this feature of OPI ratings. Their scale does reflect, however, the fact that ACTFL’s Novice level corresponds to the Foreign Service Institute (FSI) rating of 0, ACTFL’s Intermediate, to FSI 1, and ACTFL’s Advanced to FSI 2. The scale is adopted in the present study for this reason. No regression analyses were conducted, given this problematic issue of giving numerical figures to the proficiency levels by converting the
ordinal scale to an interval scale. Only the correlation analysis is presented below in Table 5.15:

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Proficiency Rating</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficiency</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar</td>
<td>.582</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.499</td>
<td>.542</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.727</td>
<td>.665</td>
<td>.652</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.542</td>
<td>.647</td>
<td>.379</td>
<td>.723</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5.15: Correlation between Proficiency Rating and Section Scores (N=50)

All correlations were significant, p=.007 for Routines/Speech Styles, and p=.000 for the rest (two-tailed). It was also found that the total score correlated significantly with the proficiency rating (Pearson r=.705, p=.000). The positive correlation between the test scores and the proficiency rating can be taken as more evidence to argue for the construct validity of the test. The proficiency rating was most highly correlated with the Speech Act scores, followed by the Grammar scores.

5.6. Relationship among the Section Scores

Research Question 4 concerns the relationship between the grammatical and pragmatic knowledge, as well as relationships among the sections of the pragmatic knowledge. Correlation analyses were conducted to examine the relationships among the section scores.
As seen in Table 5.15 above, all section scores were correlated significantly among students who received proficiency ratings. To further investigate the relationships, correlation analyses at different levels of proficiency (low and high) were conducted. The following tables, Tables 5.16 and Table 5.17, summarize the correlations at the two different levels of proficiency (** = significant at the 0.01 level, * = significant at 0.05 level, both two-tailed):

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.382*</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.341</td>
<td>.581**</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.317</td>
<td>.221</td>
<td>.448*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5.16: Correlations of Section Scores among Low-Proficiency Students (Intermediate-Mid and below), N=32

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.518*</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.727**</td>
<td>.514**</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.797**</td>
<td>.245</td>
<td>.796**</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5.17: Correlations of Section Scores among High-Proficiency Students (Intermediate-High and above), N=18

The strengths of the correlation relationships among the sections varied depending on the levels of proficiency. Among the low-proficiency group, the Grammar section score was not correlated significantly with the Speech Acts or with the Speech Styles.
section scores. However, among the high proficiency group, all correlations were significant except for the Routines and Speech Styles scores.

To further investigate the relationships, correlation analyses for different levels of instruction were also conducted. The following Table 5.16 is a summary of correlation coefficients among the 110 subjects (excluding the nine repeats). All correlations were significant at the 0.01 level (two-tailed):

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.626</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.666</td>
<td>.662</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.660</td>
<td>.536</td>
<td>.707</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5.18: Correlation among Section Scores (all subjects, N=110)

The following tables (Tables 5.19-22) summarize the correlations at different levels of instruction separately (** = significant at the 0.01 level, *= significant at 0.05 level, both two-tailed):

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.536**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.362*</td>
<td>.586**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.420**</td>
<td>499**</td>
<td>.445**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.19: Correlation among Section Scores (2nd Level students, N=41)
<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.501**</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.623**</td>
<td>.510**</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.608**</td>
<td>.318</td>
<td>.635**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.20: Correlation among Section Scores (3rd Level students, N=35)

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.284</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.332</td>
<td>.000</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.373</td>
<td>-.257</td>
<td>.419</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.21: Correlation among Section Scores (4th Level students, N=14)

<table>
<thead>
<tr>
<th>Pearson Correlation</th>
<th>Grammar</th>
<th>Routines</th>
<th>Speech Acts</th>
<th>Speech Styles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routines</td>
<td>.512</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech Acts</td>
<td>.871**</td>
<td>.597</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Speech Styles</td>
<td>.764*</td>
<td>.317</td>
<td>.814**</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.22: Correlation among Section Scores (5th Level students, N=10)

The strengths of the correlation relationships among the sections varied depending on the levels of instruction. All section score correlations were significant at the 2nd Level; that is, at the early stage in the test takers’ development, the scores on the four sections were highly related and perhaps indistinguishable. However, among learners at the upper levels of instruction, correlations among the variables became lower, suggesting the related and yet separate nature of the dimensions. The Routines scores became uncorrelated with other section scores at the upper levels. None of the section
scores were correlated for the 4th Level students. Grammar, Speech Acts, and Speech Styles scores again showed significant correlations among the 5th Level students.

To further investigate the underlying structure and relationships among the test items, an exploratory factor analysis was also conducted. The 48 test items from all 119 participants were analyzed to see if there might be common underlying factors for each section of the test.

Exploratory factor analysis identified 16 factors with the following eigenvalues (which measures the variance in all the variables which is accounted for by each factor) greater than 1.00 (the Kaiser criterion):
<table>
<thead>
<tr>
<th>Component</th>
<th>Eigenvalue</th>
<th>% of variance</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.806</td>
<td>18.346</td>
<td>18.346</td>
</tr>
<tr>
<td>2</td>
<td>3.204</td>
<td>6.675</td>
<td>25.021</td>
</tr>
<tr>
<td>3</td>
<td>2.236</td>
<td>4.659</td>
<td>29.680</td>
</tr>
<tr>
<td>4</td>
<td>2.151</td>
<td>4.481</td>
<td>34.161</td>
</tr>
<tr>
<td>5</td>
<td>1.938</td>
<td>4.038</td>
<td>38.199</td>
</tr>
<tr>
<td>6</td>
<td>1.880</td>
<td>3.916</td>
<td>42.116</td>
</tr>
<tr>
<td>7</td>
<td>1.693</td>
<td>3.527</td>
<td>45.642</td>
</tr>
<tr>
<td>8</td>
<td>1.591</td>
<td>3.316</td>
<td>38.958</td>
</tr>
<tr>
<td>9</td>
<td>1.460</td>
<td>3.042</td>
<td>51.999</td>
</tr>
<tr>
<td>10</td>
<td>1.341</td>
<td>2.981</td>
<td>54.981</td>
</tr>
<tr>
<td>11</td>
<td>1.340</td>
<td>2.792</td>
<td>57.773</td>
</tr>
<tr>
<td>12</td>
<td>1.270</td>
<td>2.646</td>
<td>60.419</td>
</tr>
<tr>
<td>13</td>
<td>1.193</td>
<td>2.486</td>
<td>62.905</td>
</tr>
<tr>
<td>14</td>
<td>1.152</td>
<td>2.401</td>
<td>65.305</td>
</tr>
<tr>
<td>15</td>
<td>1.087</td>
<td>2.266</td>
<td>67.571</td>
</tr>
<tr>
<td>16</td>
<td>1.019</td>
<td>2.122</td>
<td>69.693</td>
</tr>
</tbody>
</table>

Table 5.23: Factors with eigenvalues > 1

The scree plot below (Figure 5.10) illustrates that first high eigenvalue and the precipitous decline following it.
Various solutions were fitted to the data. A three-factor solution, which would suggest itself based on the above scree plot, yielded low communalities after extraction (9 items with communalities over 0.4, average communalities 0.296). Garson (Garson 2009) explains that communalities ($h^2$) are the squared multiple correlation for the variable as dependent using the factors as predictors, and communalities over 0.7 are considered high; .04 is usually the cut-off point. The factor loadings, which are summarized in the
next table, are difficult to interpret, as the patterns do not reflect the test’s internal structure and seem unrelated to difficulty levels, involved structural patterns, speech styles, etc. The following is a summary of factor loadings:
<table>
<thead>
<tr>
<th>items</th>
<th>Communalities ($h^2$)</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>targeted knowledge/patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR01</td>
<td>0.151</td>
<td>0.368</td>
<td>-0.128</td>
<td>0.156</td>
<td>～くなかった</td>
</tr>
<tr>
<td>GR02</td>
<td>0.247</td>
<td>-0.121</td>
<td>0.515</td>
<td>0.129</td>
<td>～だと思う</td>
</tr>
<tr>
<td>GR03</td>
<td>0.345</td>
<td>0.566</td>
<td>0.049</td>
<td>-0.022</td>
<td>～れでしょう</td>
</tr>
<tr>
<td>GR04</td>
<td>0.327</td>
<td>0.293</td>
<td>0.242</td>
<td>0.221</td>
<td>～の本</td>
</tr>
<tr>
<td>GR05</td>
<td>0.299</td>
<td>0.149</td>
<td>0.426</td>
<td>0.081</td>
<td>～でいい</td>
</tr>
<tr>
<td>GR06</td>
<td>0.261</td>
<td>-0.07</td>
<td>0.021</td>
<td>0.524</td>
<td>～る前に</td>
</tr>
<tr>
<td>GR07</td>
<td>0.297</td>
<td>-0.052</td>
<td>-0.036</td>
<td>0.562</td>
<td>～を卒業する</td>
</tr>
<tr>
<td>GR08</td>
<td>0.16</td>
<td>-0.179</td>
<td>0.463</td>
<td>0.02</td>
<td>～についての～</td>
</tr>
<tr>
<td>GR09</td>
<td>0.397</td>
<td>0.69</td>
<td>-0.154</td>
<td>0.031</td>
<td>買ったばかりの～</td>
</tr>
<tr>
<td>GR10</td>
<td>0.373</td>
<td>0.215</td>
<td>0.054</td>
<td>0.491</td>
<td>自動詞／他動詞</td>
</tr>
<tr>
<td>GR11</td>
<td>0.308</td>
<td>0.412</td>
<td>0.12</td>
<td>0.16</td>
<td>～みたいだ</td>
</tr>
<tr>
<td>GR12</td>
<td>0.111</td>
<td>0.191</td>
<td>0.05</td>
<td>0.192</td>
<td>～なんて、</td>
</tr>
<tr>
<td>RT01</td>
<td>0.167</td>
<td>0.335</td>
<td>0.12</td>
<td>-0.013</td>
<td>行ってきます</td>
</tr>
<tr>
<td>RT02</td>
<td>0.32</td>
<td>0.642</td>
<td>-0.208</td>
<td>0.024</td>
<td>ごちそうさま</td>
</tr>
<tr>
<td>RT03</td>
<td>0.229</td>
<td>0.302</td>
<td>0.264</td>
<td>-0.187</td>
<td>お久しぶりです</td>
</tr>
<tr>
<td>RT04</td>
<td>0.364</td>
<td>0.69</td>
<td>-0.14</td>
<td>-0.183</td>
<td>少々お待ちください</td>
</tr>
<tr>
<td>RT05</td>
<td>0.317</td>
<td>0.243</td>
<td>0.024</td>
<td>0.432</td>
<td>お大事に</td>
</tr>
<tr>
<td>RT06</td>
<td>0.252</td>
<td>-0.106</td>
<td>0.472</td>
<td>-0.337</td>
<td>ご苦労様</td>
</tr>
</tbody>
</table>

Table 5.24: Factor Loadings, Promax rotation with Kaiser normalization
Table 5.24 continued

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RT07</td>
<td>0.247</td>
<td><strong>0.568</strong></td>
<td>-0.158</td>
<td>-0.019</td>
<td>ごめんください</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT08</td>
<td>0.5</td>
<td><strong>0.615</strong></td>
<td>0.13</td>
<td>0.042</td>
<td>せっかくですが</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT09</td>
<td>0.355</td>
<td><strong>0.419</strong></td>
<td>0.266</td>
<td>-0.033</td>
<td>具合はどうですか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT10</td>
<td>0.059</td>
<td>0.027</td>
<td>0.046</td>
<td>-0.252</td>
<td>お世話になっております</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT11</td>
<td>0.167</td>
<td><strong>0.408</strong></td>
<td>0.034</td>
<td>-0.128</td>
<td>何もございませんが</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT12</td>
<td>0.18</td>
<td>-0.255</td>
<td><strong>0.358</strong></td>
<td>0.293</td>
<td>よいお年をお迎えください</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA01</td>
<td>0.506</td>
<td>-0.116</td>
<td>-0.027</td>
<td><strong>0.738</strong></td>
<td>〜ていただけませんか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA02</td>
<td>0.214</td>
<td>0.241</td>
<td>-0.109</td>
<td><strong>0.37</strong></td>
<td>〜てくれない？</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA03</td>
<td>0.539</td>
<td>0.291</td>
<td><strong>0.536</strong></td>
<td>-0.377</td>
<td>お願いできますか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA04</td>
<td>0.345</td>
<td><strong>0.257</strong></td>
<td>0.231</td>
<td><strong>0.3</strong></td>
<td>〜させてもらいたいんだけど</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA05</td>
<td>0.284</td>
<td>0.007</td>
<td><strong>0.518</strong></td>
<td>0.045</td>
<td>しましょうか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA06</td>
<td>0.37</td>
<td><strong>0.435</strong></td>
<td>-0.29</td>
<td><strong>0.419</strong></td>
<td>〜てあげようか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA07</td>
<td>0.276</td>
<td><strong>0.253</strong></td>
<td>0.164</td>
<td><strong>0.284</strong></td>
<td>させてください</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA08</td>
<td>0.185</td>
<td>-0.119</td>
<td><strong>0.484</strong></td>
<td>-0.018</td>
<td>〜させていただきます</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA09</td>
<td>0.412</td>
<td><strong>0.542</strong></td>
<td>-0.132</td>
<td>0.311</td>
<td>〜ようよ。</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA10</td>
<td>0.119</td>
<td>-0.066</td>
<td><strong>0.349</strong></td>
<td>0.09</td>
<td>〜ませんか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA11</td>
<td>0.44</td>
<td>0.335</td>
<td><strong>0.386</strong></td>
<td>0.092</td>
<td>〜でもいかがですか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA12</td>
<td>0.414</td>
<td><strong>0.448</strong></td>
<td>0.305</td>
<td>-0.089</td>
<td>〜がいいんじゃない？</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS01</td>
<td>0.204</td>
<td><strong>0.335</strong></td>
<td>0.029</td>
<td>0.199</td>
<td>いただきます</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS02</td>
<td>0.293</td>
<td>0.096</td>
<td><strong>0.498</strong></td>
<td>-0.125</td>
<td>おりませんが</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS03</td>
<td>0.578</td>
<td><strong>0.69</strong></td>
<td>0.164</td>
<td>-0.314</td>
<td>ご存知ですか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS04</td>
<td>0.492</td>
<td>-0.052</td>
<td>0.436</td>
<td><strong>0.501</strong></td>
<td>拝見しました</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS05</td>
<td>0.405</td>
<td>-0.088</td>
<td><strong>0.684</strong></td>
<td>-0.186</td>
<td>〜できません</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS06</td>
<td>0.398</td>
<td>0.076</td>
<td><strong>0.543</strong></td>
<td>0.137</td>
<td>〜します</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS07</td>
<td>0.226</td>
<td>-0.028</td>
<td><strong>0.356</strong></td>
<td>0.274</td>
<td>お知らせいたします</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS08</td>
<td>0.309</td>
<td>-0.024</td>
<td><strong>0.476</strong></td>
<td>0.233</td>
<td>行かれましたか</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS09</td>
<td>0.224</td>
<td><strong>0.403</strong></td>
<td>-0.05</td>
<td>0.198</td>
<td>〜ください</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS10</td>
<td>0.083</td>
<td><strong>0.298</strong></td>
<td>-0.102</td>
<td>0.076</td>
<td>お越しください</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS11</td>
<td>0.145</td>
<td>-0.021</td>
<td><strong>0.265</strong></td>
<td>0.24</td>
<td>ご指導いただいた</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS12</td>
<td>0.31</td>
<td>-0.18</td>
<td>0.309</td>
<td><strong>0.482</strong></td>
<td>お尋ねください</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

139
The highest loadings are in bold and underlined, and loadings within .05 of the highest loading are in bold. An oblique (unorthogonal) Promax rotation was applied, following Bachman (1990): “Although orthogonal solutions (uncorrelated factors) have historically been preferred because they tend to yield more interpretable solutions, researchers who deal with abilities that they expect to be related are increasingly turning to oblique solutions (correlated factors).” The factor loadings did not match the test section structure, suggesting that the hypothesized constructs for the test may not be distinguishable. The first factor with high eigenvalue can also be interpreted as a one-factor structure, which is supported by the high correlations among the sections. However, with low communalities and this small data set, the factor loadings table should be interpreted with caution. Garson (2009) further points out issues with factor analysis with dichotomous data (0 or 1 scoring, as is the case with the multiple-choice items), noting that “dichotomous data tend to yield many factors (by the usual Kaiser criterion), and many variables loaded on these factors (by the usual .40 cutoff), even for randomly generated data.” In the present case, then, more data (300 subjects or more) will be ideal to conduct a more reliable factor analysis.
Chapter 6: Discussions and Conclusion

This chapter provides a summary of the findings and general discussions, followed by comments on limitations of the study and an outlook for the project.

6.1. Summary of the Findings and General Discussions

This dissertation study concerns the development and validation study of a web-based, multiple-choice test to measure off-line pragmatic knowledge of JFL learners. Data from 110 participants revealed that the test was overall reliable (alpha = .893). The total scores of the test positively correlated with length of study/exposure (r = .380), levels of instruction (r = .685) and proficiency ratings based on the ACTFL Guidelines (r = .705), suggesting the validity of the test. As for the relationship among grammatical knowledge and the domains of pragmatic knowledge, the study found complex interrelationships: correlation analyses showed that the four sections (Grammar, Routines, Speech Acts, and Speech Styles) are significantly correlated overall (r > .536), but the strengths of correlation varied depending on proficiency levels and the levels of instruction. The data also suggest that the knowledge of speech styles, grammar, and speech acts are in closer relationship among JFL learners than among ESL/EFL learners. Exploratory factor
analysis found unexpected patterns of factor loadings, not matching the internal structure of the test, suggesting that the hypothesized components of language ability may not be empirically separable.

Finding for each research questions are discussed separately in more details in the following sections.

6.1.1 Test Reliability

In response to Research Question 1, the data showed that the test was overall highly reliable (Table 12). The reliability of each section was slightly lower than Röver’s (2005) ESL/EFL web-based test, and notably higher than the multiple-choice tests in Yoshitake’s EFL and Yamashita’s JSL studies reported in Brown (2001). It is particularly encouraging that the multiple-choice format of the test in this dissertation study achieved acceptable reliability in the Speech Acts section, as that has been a challenge in some of the previous studies (Hudson, Detmer, and Brown 1995; Yamashita 1996). One of the possible reasons for this higher reliability is that the test items in the present study focus more on the form-function association, instead of strategies, order of strategy use, etc. In other words, the test in the present study deals with the more discretely coded, linguistic “end” of the pragmatics continuum, where we find particular linguistic patterns in specified contexts acknowledged by a high level of consensuses among target natives.

The relatively lower reliability in the Routines section in this study shows patterns similar to Röver’s (2005) ESL/EFL studies, in which the alpha for the Routines in his study was reported as .73, while the Implicature section was .82, and the Speech Acts .89.
The relatively low reliability of the Routines sections in the present test is largely due to the two relatively unreliable items, RT06 and RT10, which are presented below, with the number of test takers that selected each of the four options:

6: Mr. Noguchi, an employee, is saying good-bye to his supervisor, as she leaves the office. What would Mr. Noguchi probably say to his supervisor?

「じゃ、あとはよろしくね。また明日。」
「あ、はい、どうも______。」
A. ご愁傷さまでした (N=11)
B. ご苦労さまでした (N=19)
C. お疲れさまでした (N=83)
D. お気の毒さまでした (N=6)

10: Mr. Kubota, an employee, is answering a phone call from another company. What would Mr. Kubota probably say to his business associate?

「もしもし。あ、田中商事の山本ですねけれども。」
「あ、どうも______。」
A. ごきげんよう (N=16)
B. お元気そうで (N=3)
C. お世話になっております (N=78)
D. 何のお構いもしませんで (N=22)

The first item, RT06, essentially tests if the test takers can distinguish お疲れさまでした from ご苦労様でした to acknowledge the others’ hard work, depending on to whom the comment is being addressed. This distinction, however, is apparently not a black-and-
white one even among the natives. The Agency for Cultural Affairs (2006) reports that only 69.2% of the 2,107 who responded said that they would use お疲れさまでした to a superior. Although there was a 100% agreement among the experts on the use of お疲れさまでした during the Pilot Test, this test item may be targeting at a blurry distinction, which may have contributed to the low reliability.

The second item, RT10, is lowest in reliability estimate, and the negative ID indicates that more participants in the bottom one-third in overall scores answered this item correctly than those in the top one-third. This is probably an instruction effect: it is very likely that this expression was taught in the 2nd Level and 2nd/3rd Intensive levels at University A, a few weeks before the test administration.

6.1.2 Relationships between the Test Scores and Length of Study/Exposure, Levels of Instruction

In response to Research Question 2, the significant positive correlations between the test scores and levels of instruction, as well as between the test scores and length of study, suggest that the test scores are indicative of the learners’ increase in pragmatic knowledge as they advance in their language learning and accumulate more pragmatic knowledge.

The correlations, however, show strikingly different strengths (with levels of instruction $r = .685$; with length of study/exposure $r = .380$). This can be again an indication of overall instructional effect: the contents of instruction at different levels among the four universities might be showing high levels of agreement. This speculation
is particularly plausible for the JFL setting in the U.S. that this study is concerned with, where the instruction accounts for majority of input for the learners, and where a diversity of educational practices can be found between high schools and colleges (e.g., “4 years” of study length can mean “3 years in high school + 1 year in college”, “1 years in high school + 3 years in college”, among other numerous possibilities). In addition, correlation coefficient between the proficiency ratings and length of study/exposure is .414, \( p=.003 \), while between the proficiency ratings and levels of instruction is .631, \( p=.000 \). The stronger correlation between the proficiency rating and levels of instruction also supports the speculation that level of instruction is a better indicator of proficiency, compared to sheer length of study. Or, perhaps it is due to the imprecise (and potentially confusing) categorical measure for length of study (continuous value) such as “1 year or less,” “2 years,” etc. More precise measure of length of study (in months, or contact hours of instruction, for example) could have yielded a different result.

The one-way ANOVA showed significant difference in overall scores among learners at different levels of instruction except for the 4th and 5th Levels. This is both good and bad news: good in the sense that the test can differentiate learners at different levels for the most part, and bad in a sense that it shows a ceiling effect of the test. The ceiling effect, observed at the higher levels of instruction, suggests that the current 48-point scale might need to be expanded (to 0 to 60, for example). While the test is targeted at Japanese majors at college level (with typically four years of instruction), the test might be more useful if it were to include more challenging items that can distinguish learners at the higher levels. In developing the test, the decision to limit the number of
items to 48 was based on practical considerations, such as many institutions’ inability to spend more than an hour on a test (Itomitsu 1996).

A somewhat puzzling finding is the slight decrease of scores in the Speech Acts and Speech Styles sections among students who have been studying for 4 years (Figure 5.3). Comparing students in this group (N=19) with those who have been studying for 3 years (N=26), we see that both groups are similar in terms of the composite percentage of students from three different universities. The mean proficiency rating for both groups are similar when the rating are converted to the numerical scale (“4 years” = 1.409, N=11; “3 years” = 1.421, N=14), and the mean overall test scores are actually higher among the “4 years” group (“4 years” = 32.37; “3 years” = 29.88). This might be due to the small size of the data sample. This may also be because one student in the “3 years” group, whose proficiency level was rated “Advanced-High,” received 42 points out of the possible 48, which may have inflated the data of that group. It should be noted, however, that the drop in the mean scores for both sections is approximately 0.5 point, and that the difference in the mean scores between the “3 years” and “4 years” groups is not statistically significant.

6.1.3 Relationship between the Test Scores and Proficiency ratings

In response to Research Question 3, significant positive correlation between the test scores and the proficiency rating was found (r = .705), indicating the convergent validity of the test. All section scores were significantly correlated with the proficiency ratings, with the Speech Act sections showing the strongest correlations, and the Routines
showing the weakest (Table 5.15). The mean plot of section scores by proficiency rating (Figure 5.9) shows similar tendencies as the mean plot of section scores by levels of instruction (Figure 5.6).

A few reasons can be hypothesized for the different strengths of correlation between the proficiency rating and the section scores. Recall that the ACTFL Guidelines define proficiency in terms of the four assessment criteria: global tasks and functions, context/content, accuracy, and text type. Pragmatic competence belongs to the Accuracy criterion, along with grammatical and sociolinguistic competences. It is plausible that the four sections of the test in the present study all tap into the “accuracy” criteria of the ACTFL Guidelines, while the Speech Acts section’s items include more variety of “tasks and functions” and “content/context” than those in the other sections. Knowledge of speech acts increasing with proficiency is in accord with Röver’s (2005) study.

The weakest correlation, which is between the ACTFL-based proficiency rating and the Routines scores, can be explained by the format of the proficiency interview test, mainly consisting of question-and-answer on everyday topics with one or two role-plays (especially for the Novice and Intermediate levels). The interview setting does not call for a heavy use of routine expressions: initial greetings, leave-taking, and other routine expressions most likely appear in the warm-up and wind-down stages, but probably not during the level check and probing stages.

Another interesting observation with regards to the test scores by proficiency rating is their relationship between proficiency rating and the scores in the Speech Styles section. The box-plots (Figure 5.8) show a generally positive correlation between the
Speech Styles scores and proficiency rating, despite the fact that the ACTFL Guidelines include the notion of register only in the descriptions for Advanced and above levels. The widest variance, which we see at the Intermediate-High level, however, shows a wide range of difference in Speech Style knowledge among students below the Advanced level. The box-plots also seem to indicate that reaching a competent level in Speech Style takes more time, compared to grammatical and other domains of pragmatic knowledge.

6.1.4 Relationship among Components of Communicative Language Ability

In response to Research Question 4, the answer is not straight-forward: the study found that all four sections of the test show significant levels of correlation overall (Tables 5.15 and 5.18), but strengths of correlation differed depending on the proficiency levels (Tables 5.16 and 5.17), as well as levels of instruction (Tables 5.19 -- 5.22). The Routines section shows weakest correlation with other sections. The correlations among the Grammar, Speech Acts, and Speech Style sections are significant at high-proficiency group as well as those who are in the 5th level of instruction. The factor analysis results are difficult to interpret and it did not provide strong support for the distinct components of communicative language ability that the test construct hypothesized (Table 5.24).

The fact that the Routines scores were not significantly correlated, especially at the upper levels, concurs with Rover’s (2005) factor analysis in ESL/EFL, which detected a separate factor for each of the sections of Implicature, Routines, and Speech Acts. This is probably because the Routines section is somewhat different in terms of the tasks required of the test takers: unlike the other sections where knowledge of grammar and
vocabulary may help the test takers “puzzle out” an answer, perhaps in a process of eliminations with the given options, the Routines section is “entirely memory-based, and does not require much problem-solving” (Röver 2005). Put it in another way, it could be argued that knowledge of structural patterns are relevant in the Grammar, Speech Acts, and Speech Styles sections, but not so much in the Routines section. This relevance of the three types of knowledge, however, may simply be a method effect caused by the multiple-choice test format.

The different strengths of correlation depending on levels found in this study are congruent with Kim’s (2006) study on ESL role-play data with 90 subjects, although in her study the correlations among the four hypothesized components of speaking ability were still significant at all levels. In this study, the Routines section is significantly correlated with all the three other sections at the 2nd level, only with the Speech Acts section at the 3rd level, and with none of the section at the 4th and the 5th levels. It is intriguing that none of the sections correlated at the 4th level: The Routines section has no correlation (.000) with Speech Acts and a negative correlation with Speech Styles at the 4th Level. This might be due to the small size of the data set (N=14).

This study finds high overall correlations among grammatical knowledge, speech acts and speech styles. This is in accord with Grabowski’s (2008) correlation analysis. The close relationship between the grammatical and pragmatic knowledge found in this study is also consistent with previous studies in ESL/EFL (Bachman and Palmer 1982; Xu, Case, and Wang 2009). The finding that no separate factors for grammatical and pragmatic knowledge were found in exploratory factor analysis is in agreement with
Harley et al.’s (1990: 24) study reporting that “confirmatory factor analysis did not strongly support the hypothesized distinctions”, and also with findings by Roever and Niezgoda (cited in Röver 2005: 33) from their factor analysis on a replication study of the Bardovi-Harlig and Dörnyei study: “no relationship between grammatical and pragmatic awareness was found: items related to these traits loaded very clearly on independent and uncorrelated factors”.

Röver’s findings in his factor analysis, on the other hand, identified separate factors for his Implicature, Routines, and Speech Acts section. This might be attributed to his large sample size (N=267), with a more diverse test taker population (ESL/EFL, 3rd-9th grades). Or, the relationships among the components of communicative language ability found in ESL/EFL studies may be different from the relationships among those that make up JFL communicative language ability.

Previous studies (Bachman and Palmer 1982; Kim 2006) showed that sociocultural competence (register, nativeness, etc.) can be separable from the other components of communicative language ability. The present study, however, found that the Speech Styles scores were highly correlated with the Grammar and Speech Act sections, regardless of the levels of instruction, except for the 4th level. The Speech Styles scores were also significantly correlated with the Grammar and Speech Acts sections in high-proficiency group. This suggests that, as far as JFL is concerned, the knowledge of speech styles is in closer relationship with the knowledge of grammar and speech act, compared with ESL/EFL. This is understandable, given the various lexical forms and
syntactic structures the learners need to know to identify the unexpressed agent of actions, and/or to realize different speech styles (Wetzel 1984).

The results from the correlation studies and factor analysis do not fully support the distinctness of the hypothesized components of language ability (grammatical knowledge and domains of pragmatic knowledge). Thus, construct validity of the test — the extent of appropriateness of the inference and prediction one can make based on the test scores about test takers’ ability in the target language use situation — is defensible only for the overall score being a valid indication of communicative language ability in general. The extent in which each section scores are valid indications of corresponding hypothesized separate components of the language ability remains unknown, and needs to be further investigated by such future research as multimethod-multitrait study, experimental treatment, etc. (Bachman 1990).

Schachter (1990: 43-4) criticizes Harley et al.’s (1990) study, which investigated the distinctness of the components of the hypothesized language ability yet failed to provide an empirical evidence in factor analysis:

[Harley et al.’s] study points out the difficulties involved in distinguishing coherent and theoretically separable competences. […] I’m not even sure if there is a conceptual distinction to be made between pragmatic and sociolinguistic competence. It seems we must always resort to talking in terms of ‘more or less’ […]. It seems at least possible that communicative competence is best viewed as consisting of two kinds of competence — grammatical and pragmatic — and that sociological phenomena interact with these two components at all levels. Is it not the case that cultural or sociological criteria influence all levels of a grammar, from phonological to syntactic to pragmatic? Why should we want to claim that they constitute yet a third component? (original emphasis)
The findings of this study, which imply close relationships among grammar, speech acts, and speech style knowledge in JFL, underscore Schachter’s view that sociolinguistic competence influences both grammatical and pragmatic knowledge. This study also confirmed that the hypothesized components of language ability are difficult to distinguish empirically, probably due to the complexity of the relationship among them: “Theoretically, the variables should be highly correlated with one another since they are hypothesized components of the same language competence, and yet the magnitude of correlation should not be too high since they are also supposed to be distinct” (Kim 2006: 19).

Nonetheless, the web-based test developed in this study, in conjunction with other types of tests, should contribute to the continued investigation on the possibility and usefulness of theorizing different competences in foreign languages, and to measuring them to inform the learners, teachers, and curriculums.

6.2. Limitations of the Study

Several limitations of the study should be acknowledged in this section. First, the data set comes from four universities in the American Midwest. The subjects were not randomly sampled from the entire population of all learners of Japanese in the U.S. The generalizability of the findings in this study needs to be confirmed by a larger-scale study in the future. A data set of 119 cases is regarded to be too small for factor analysis by some scholars (Garson 2009). More data is necessary to investigate the togetherness or distinctness of the components of communicative language ability in more confidence.
Second, the proficiency rating data were obtained through four different sources, and the inter-rater reliability was not addressed in this study. Although all the interviewers have received formal, official training by ACTFL or at the Defense Language Institute, it is unknown how reliable the ratings are.

Third, this test measures only a limited aspect of pragmatic knowledge. The grammar, routines, speech acts, and speech styles operationalized in this study were not meant to represent everything that is involved in communicative competence. The usefulness of the test depends on how well the content specifications match the views of different domains of competences among test users (teachers, program administrators, researchers, etc.). Most significantly, this test does not tap into the discourse competence — the organization of ideas, order of strategies (apologies before request, mitigating, etc.) and other such Japanese-specific elements as use of final particles, aizuchi or back-channeling, etc., that are reviewed in Kita and Ide (2007). The discourse competence is tied in with “text type” in the ACTFL Guidelines, which was introduced to the ACTFL guidelines in 1989 as an independent criterion, and has been proven to be relatively transparent (Makino 1991). The test in the present study restricts the tested speech acts to speech acts that are more discretely packaged (e.g. in a single utterance, with a single construction and few grammatical parts), and not spread out over several turns of a conversation. The more discretely packaged speech act is a speech act just as much as a more diffuse one is, and is more easily tested. However, future development of a test of discourse competence would shed light as to yet additional aspect of communicative language ability.
Fourth, the content specification of the test was derived from textbook analysis, and it is arguable that the test construct may be underrepresented in such data. The construct validity of the test scores as an appropriate indicator of test takers’ language ability in general is based on the assumption that the data obtained through the textbook analyses contain relevant and representative samples of the target language use situations. This assumption needs to be empirically demonstrated by future studies.

Fifth, the test is a discrete-point, multiple-choice test of offline pragmatic knowledge, with no time limitation. It does not include production of utterances nor interaction/negotiation with an interlocutor. This test measures test-takers’ offline knowledge by giving them ample time to study the surrounding information and to contemplate on what would work best. In real-life situations, however, interactions require immediate decision and performance. More studies are necessary to reveal the relationship between offline pragmatic knowledge and the production skills that drive more open-ended, discourse-level interactions. While discrete-point tests may be suitable for assessing various components of pragmatic knowledge, “communicative testing must be devoted not only to what the learner knows about the second language and about how to use it (competence) but also to what extent the learner is able to actually demonstrate this knowledge is a meaningful communicative situation (performance)” (Hudson, Detmer, and Brown 1995).

Sixth, the test was given online as a “no-stake” test, with no identification checks. Test security and other measures need to be implemented if this test is to be used as part of official formative or summative test by an institution. Also, the scores did not affect
the grades of the participants, which was one of the research requirements. Some students indeed contacted the researcher and wanted to know how well they did on the test, but others did not want to be informed of their scores.

Seventh, the test is designed primarily for English-native JFL learners, and its usability for other first language (L1) groups are unclear. The distractors were created in consultation with documentations (Mizutani 1994) and researcher’s experience of typical errors the English-native learners in the U.S., and therefore the “effectiveness” of the distractors is expected to be different depending on the test takers’ first language. It will be an interesting future study on negative transfer, however, to compare the results of subjects of different native languages.

6.3. Pedagogical Implications

Since this dissertation project is motivated by the question “What do JFL learners need to know in order to be successful in communicating in Japanese?”, some pedagogical considerations and viewpoints are offered in this section.

First of all, the test is based on the premise of the teachability of pragmatic knowledge, which is gaining support in the literature (Rose and Kasper 2001). The entire volume of System, volume 33 (2005) is dedicated to the topic of the effect of instruction on the development of pragmatic knowledge in the classroom, reflecting an immersing consensus that pragmatic knowledge can be taught.

Second, what can be regarded as “successful” communication involving second/foreign language learners differs depending on one’s theoretical and pedagogical
viewpoints. Often, convergence to native norms is valued by many teachers and researchers as one of the learning goals. One of the indicators of successful communication is described as verbal and non-verbal behavior that makes native interlocutors comfortable (Christensen and Warnick 2006). Since intercultural communication in Japanese usually involves a Japanese-native interlocutor, any communicative language use of Japanese to express intentions effectively must therefore consider the native audience. The reason for the test question formatted as “What would a native Japanese say?”, instead of “What would you (a learner) say?”, is based on the assumption that “knowledge of what a native speaker is likely to say in a given context is […] a crucial component of second language learners’ competence to understand second language communication and to express themselves in a native-like way” (Canale and Swain 1980).

However, such a view is not free from controversy. Liu (2006: 1) reports that during the plenary talk of the 2005 Second Language Research Forum (SLRF), Vivian Cook “reiterated the concept of successful L2 users that she proposed earlier in 1999, and contended that the norms of successful L2 users, rather than native speaker norms, should be used as the standards for L2 teaching and learning.” Also, some researchers and teachers of Japanese in the U.S. express concern for presenting a native model and asking learners to imitate the model. “We have to examine our ways of teaching in order to stay away from the colonial legacy of assimilationism or the sanctification of Japanese language and culture” (Tai 2003: 1). The same article cite Seiko Kataoka (2000) and notes that “In [Kataoka’s] search for a non-assimilationist Japanese culture course, she
finds the idea of ‘smooth communication’ problematic. This idea, often listed as a teaching goal, makes teachers inclined to push students to assimilate to a ‘Japanese-style communication’” (Tai 2003: 21).

It is important to explain to the learners why they might benefit from learning prototypical conversation patterns among natives. At the same time, intracultural diversity should also be addressed in a language curriculum, and teachers should refrain from presenting model behaviors in the target culture as fixed, invariant entities. It follows that a test of pragmatic knowledge, such as the one presented in this study, should undergo the same scrutiny: “A fundamental concern in constructing items for tests of pragmalinguistic knowledge is that they be indeed representative of real-world language use, [keeping] pace with diachronic changes in language use” (Röver 2005: 115) What is “correct” in answering the question “What would a native say in this situation?” in the test developed in this study should be based on contemporary agreements among the people in the target culture, and should be viewed as fundamental knowledge on which learners can build, as cultural capital, their repertoires in and for language use. Models of native norms should be presented as models, so that the learners can build their communication repertoire by modifying the model and applying and adapting it to slightly different situations as they gain guided experience in actual language use. Learners can make invaluable use of knowledge of native norms to gauge real-life interactions, and continue engaging in the long-term modification of the models. Properly understood, prototypical models are only the starting point of this endeavor, and not a fixed entity or endpoint.
There are at least two issues surrounding the notion of convergence to the native norm in a JFL context. One is learners’ resistance to using perceived L2 pragmatic norms, resulting in unwillingness to learn certain language forms that conflict with their own subjective position, or avoidance of particular patterns that they have mastered linguistically and are capable of producing (Cohen 2009). Siegal (1996: 356-7) points out “while constructing a ‘face’ within an interaction, learners might experience conflict concerning sociolinguistic appropriateness in their L2. […] For most native English speakers learning Japanese, this point takes on a particular saliency because of the way register variation is marked in Japanese with grammatical and lexical variation. Changing registers involves the ability to envision oneself in multiple ways.”

The other issue is a variety of expectations and attitudes among natives toward intercultural communication. Yokoyama (National Institute for Japanese Language 2006) found that native Japanese speakers engage in “foreigner talk”: her study investigated 80 Japanese natives’ refusal patterns to both Japanese and non-Japanese speakers through role-play. She found that Japanese natives tend to modify their speech by avoiding the use of linguistic mitigating devices when they speak to non-natives, and monolingual Japanese simplified their speech to non-natives more than bilingual Japanese did.

Also “being a gaijin (literally ‘foreigner’) in Japanese society creates different expectations […] among native speakers and thus further distinguishes and complicates language use for Japanese language learners” (Siegal and Okamoto 2003: 59). The prevailing expectations among native Japanese of gaijin speakers of their language are sometimes not very productive: “Jorden (1980) discusses the contrasts in language
instruction and learning between Japanese native teachers and American students whereby the widespread Japanese assumption that foreigners will not be able to speak is shared by the teachers, resulting in self-fulfilling low classroom expectations. Moreover, L2 speaker shortcomings are not only expected but excessively tolerated and even regarded as ‘reassuring’ (Loveday 1982: 48).

McNamara and Röver (2006: 73) summarize the issues as follows:

Research on pragmatics assessment tends to be referenced to target language norms (based on NS intuitions), and greater convergence of learners toward these norms is taken as an indication of greater competence. This view is not unproblematic, however. Pragmatic performance serves an indexical function as well, and nativelike performance might be interpreted by NSs as a claim to group membership. Where such a claim is evaluated as inappropriate or preposterous, group members might exclude learners rather than applaud their nativelike competence.

Given the wide range of perceptions found in both the learners and the native-speakers of Japanese, what is the role of language instructors, and language testers of Japanese?

Loveday (1982: 101) argues:

The reason that communicative competence should be a goal of L2 teaching is that without such knowledge and ability, varying degree of dissatisfaction and frustration for both native and non-native are bound to arise. The purpose of teaching communicative competence, then, is to avoid and overcome this potential dissatisfaction and frustration. This does not mean that perfect nativeness should be principally aimed for, unless so desired by the learner, but rather, more subtly, that those aspects of non-nativeness which disturb and rupture communication and which can lead to negative stereotyping concerning personality and intelligence should be avoided.

Appropriate language tests can be used to investigate learners’ knowledge of communicative competence, and further aid research and instruction. The Web-based test
in this study, which asks if learners know what a native speaker would say in the given situation, by design, neither requires nor assumes that learners adopt such norms as their own when they communicate in Japanese. Therefore, it is possible that the test can be combined with a performance test (role-play, etc.) and other measurements to investigate whether the learners know of the native norms yet decided to diverge from such norms, or the divergence observed in performance test may be simply learner’s lack of knowledge of the native norms. The decision of convergence or divergence is ultimately up to the learners. The pragmatic knowledge tested in this study is considered beneficial to learners in that it empowers them to make an informed decision, by providing learners with metacognitive awareness to gauge the effect of convergence or divergence in a particular interaction situation. “In all societies, members may vary which stances they display and in so doing build different sorts of social identities. In Japanese society, females do not necessarily display hesitancy and delicacy in every situation but rather select when to display these stances. Women the world over may play up or down their female gender identity” (Ochs 1996: 424). Knowing how to play up or down certain aspects of their personae is ultimately knowing how to exist among the interpretation among audiences by strategic use of indexes. “The adoption of sociocultural rules [of L2 natives] as one’s own in an L2 may have to be an individual decision. Providing the information so that a learner can make that choice is a pedagogical decision” (Bardovi-Harlig 2001: 32).

Language pedagogues can utilize this test of pragmatics to communicate to learners the importance of instructional goals, to monitor learners’ progress toward such
goals, to increase the accountability of the curriculum, and to facilitate learner’s understanding. Siegel (1996: 377) recommends “formal language classrooms can contribute to language learning (in particular, sociolinguistic competency) through focused discussion of how the language is used in the target language society. Such a discussion needs to include the intersection of language, culture, society (and all that reflects and creates power relations such as nationality, race, gender, social class, age, and so forth), history, and the learner’s position in that society.” Presenting self effectively is only possible with understanding of appropriate ways to do so in the given culture, whether in L1 or L2, and knowledge of how such presentation might be interpreted and accepted by the audience. Self cannot exist without being understood in the given society: “Human beings do not terminate at their own skins; they are expressions of a culture” (Bruner 1990). In cross-cultural or any types of communication, open-mindedness, or “a willingness to construe knowledge and values from multiple perspectives without loss of commitment to one’s own values [is] the keystone of we call a democratic culture” (Bruner 1990). This is not to deny learners’ rights to be who they are, nor is it changing what they feel to be their true identity, or forcing them to abandon their value systems. What learners are learning in a second language curriculum is additional ways of presenting their selves in a different culture, with different audiences.

In Performance-Based Pedagogy, a notion of “game” is found useful in fostering this notion of developing additional persona in learning an East Asian language. Walker and Noda (2000: 191) describe the notion as follows:
The main purpose of a pedagogical emphasis on gamesmanship is to foster a long-term metaphorical association of culture to game. If students of Chinese can think of playing Chinese culture rather than American culture in the same way they might play tennis rather than baseball, they may gain a more durable attitude toward the undertaking […] Games do not have to “make sense” in any absolute way. No one will seriously contend that baseball is a more rational game than tennis or vice versa. Different games share features, but shared features may have contrasting, even conflicting meanings and outcomes. Competence in one game often does not translate into competence in another. […] If we consider a game a performance with a shared system for keeping score, this is where playing games is beneficial to the general enterprise because games are conducive to creating the kind of openness and spontaneity that many recognize as necessary for successful language learning.

The game metaphor may be applied to pragmatic knowledge as well. Conforming to native paradigm of pragmatic practice does not negate a person’s identity, but adds a new dimension to the persona of those individuals who manage to utilize the pragmatic knowledge in appropriate contexts.

6.4. Suggestions for Future Research

This project can be further developed and improved, and can be utilized as a research tool in the future. This section lists some possible ways to improve the quality of the test, as well as a few research topics in language testing and language pedagogy to which this project can be applied.

6.4.1. Test Improvement

As mentioned, the current test can be improved further in terms of reliability. In addition to revising particularly problematic items, adding more challenging items can
increase the usefulness of the test. This presents a dilemma, however, as more items means more reliability but longer administration time and therefore lower practicality. The information gained through the item analysis reported earlier can help in selecting a few items that are low in ID and high in IF, and in adding a few more challenging items in the future.

In addition to item revisions, the test format itself could use a few revisions. Use of video to present context and stimuli will further improve authenticity of the test items, as such presentation can include body language (head and hand movements, facial expressions, etc.) and appear more realistic than illustrations. Production of video clips can be costly, however, and the test administration with video adds an extra burden on the network and software, at least at present. Furthermore, a study shows that the test takers engage differently to different media: in computer-based listening tests, while test takers have engaged minimally and similarly with still images, wide variation has been observed in the ways and degree to which they engaged with video stimuli (Ockey 2007). “A great deal of validation work would have to be undertaken to ensure that audio and video items engage pragmalinguistic knowledge the same way or more directly than text-based items” (Röver 2005: 117).

If video clips or photos are to replace the illustrations of the test, keeping the characters consistent throughout the test would make the production less costly and perhaps help the test takers to understand the given situation quickly. This may reduce the item independence, however, as understanding the situation for a given test item might be affected by whether or not the same character appeared in the previous items.
Illustrations may also have an advantage of having a simpler and more abstract quality, and less “distracting” than the photos or videos.

6.4.2. Future Research Projects

In order to further investigate the relationships among hypothesized components of communicative language ability, a multimethod-multitrait study involving this web-based discrete-point test and other production-type tests (oral interviews that are specifically designed to test pragmatic and sociolinguistic knowledge, etc.) is highly desirable. Role-plays with well-defined rating criteria for grammatical and pragmatic ability of the test takers are needed to investigate the relationships of the hypothesized grammatical and pragmatic knowledge in JSL/JFL.

There are a few rating criteria for grammatical and pragmatic, sociolinguistic, and/or socio-cultural abilities developed for ESL/EFL studies. Here are few examples from Grabowski (2008: 161) and Bachman and Palmer (1982: 456-7):
<table>
<thead>
<tr>
<th>Grammatical control</th>
<th>Sociolinguistic appropriateness</th>
<th>Sociocultural appropriateness</th>
<th>Psychological appropriateness</th>
</tr>
</thead>
<tbody>
<tr>
<td>* phonological/graphological forms and meanings; lexical forms and meanings;</td>
<td>* cultural identity markers of age, gender, status, and</td>
<td>* cultural meanings (reference, metaphor,</td>
<td>* attitude (sarcastic, irony, understandings, humor,</td>
</tr>
<tr>
<td>morphosyntactic forms and meaning; collective forms and meaning; information</td>
<td>status, and group;</td>
<td>figurative meanings);</td>
<td>deference, criticism);</td>
</tr>
<tr>
<td>management; forms and meanings; interactional forms and meaning;</td>
<td>* cultural norms, preferences, and expectations (manner-</td>
<td>* cultural norms, preferences, and expectations (manners,</td>
<td>* affect (anger, impatience);</td>
</tr>
<tr>
<td>meaningfulness in terms of the conveyance of literal meaning;</td>
<td>isms, frequency and use of apologies, formulaic</td>
<td>frequency, and use of apologies, formulaic expressions,</td>
<td></td>
</tr>
<tr>
<td>* ability to get their point across.</td>
<td>expressions, collocations);</td>
<td>collocations);</td>
<td></td>
</tr>
<tr>
<td></td>
<td>* register variation (modalities and genres).</td>
<td>* register variation (modalities and genres).</td>
<td></td>
</tr>
<tr>
<td>Response demonstrates:</td>
<td>Response demonstrates:</td>
<td>Response demonstrates:</td>
<td>Response demonstrates:</td>
</tr>
<tr>
<td>4 effective use</td>
<td>effective use</td>
<td>effective use</td>
<td>effective use</td>
</tr>
<tr>
<td>3 generally effective use</td>
<td>generally effective use</td>
<td>generally effective use</td>
<td>generally effective use</td>
</tr>
<tr>
<td>2 somewhat effective use</td>
<td>somewhat effective use</td>
<td>somewhat effective use</td>
<td>somewhat effective use</td>
</tr>
<tr>
<td>1 generally ineffective use</td>
<td>generally ineffective use</td>
<td>generally ineffective use</td>
<td>generally ineffective use</td>
</tr>
<tr>
<td>0 no effective use</td>
<td>no effective use</td>
<td>no effective use</td>
<td>no effective use</td>
</tr>
</tbody>
</table>

Figure 6.1: Role-play Rating Criteria by Grabowski (2008: 161)
Figure 6.2: Interview Rating Criteria by Bachman and Palmer (1982: 456-7)
It remains to be seen if the above criteria are applicable and useful in JSL/JFL context. Inter- and Intra-rater reliability must be addressed as well. A classroom-based assessment instruments (role-play, written DCT, reflective writing, self/peer-assessment, student-teacher collaborated assessment, etc.) have been already developed for JSL/JFL learners (Aoshima, Ishihara, and Akikawa 2008; Noda et al. 2002), but they are meant to be used for an achievement and/or prochievement test. Developing rating criteria of pragmatic knowledge for a proficiency test, as well as standardizing the testing procedures for a production-type test, present a major challenge to test developers.

One interesting research topic in interlanguage pragmatics is the effect of instructions on different domains of communicative competence. The test developed in this dissertation project can be used in a more experimental study to investigate if different instructional materials and practices affect the learners’ progress in developing pragmatic knowledge.

Another popular topic in interlanguage pragmatics is the role of environment (i.e., studying in the target country or in the home country) in the development of pragmatic knowledge. Röver (2005) found even exposure of three months or less leads to a significantly greater knowledge of situational routines, while the knowledge of speech acts seem to increase with proficiency regardless of exposure to the target country environment. Dewey (2005) reports that some study-abroad learners “developed communicative strategies that allowed them to engage in basic communication without using the appropriate linguistic tools. These patterns then fossilized and the learners had
great difficulty adding to or altering their hard-set linguistic patterns.” This report can be further confirmed by a research study exploring the effect of learning environment on knowledge of grammatical and pragmatic knowledge. The practicality of the test developed in this study can be helpful in measuring a large number of learners in the target country to further explore the relationship between the exposure/learning environment and particular aspects of pragmatic knowledge development.

6.5. Conclusion

Tests can never be perfect indicators of ability, as they are always limited in terms of specification, observation, and quantification (Bachman 1990). As seen in Chapter 3, a useful language test can be characterized as a function of construct validity, reliability, authenticity, practicality, interactiveness, and impact (Bachman and Palmer 1996). The web-based multiple-choice test of grammatical and pragmatic knowledge developed in this study shows a high level of usefulness, with its construct definitions informed by current research, careful construction of content specification based on both theoretical argument and expert judgments, acceptable reliability and indications of construct validity, and the practicality of a 60-minute online multiple-choice format.

The test developed for this dissertation study promises to be useful in further research on the nature of communicative competence as well as on the measurement of different competencies of communicative language ability. The test can also be used in a larger-scale study in the future to promote instructional effectiveness for JFL/JSL learners.
A test to measure learners’ understanding of spoken language can take several formats or designs. Brown (2001) lists six measures of pragmatic ability (a written discourse completion task, multiple-choice discourse completion task, oral discourse completion task, discourse role-play task, discourse self-assessment task, and role-play self-assessment task), each of which has its own strengths and weaknesses. It would be best to administer the test in the present study along with other types of tests to compensate for weaknesses of the different test formats.

This test is also expected to have positive washback with learners and teachers, because it adds to the traditional tests that focus on grammar, vocabulary and reading, and, unlike them, emphasizes the importance of situated social interactions and understanding intentions in a culturally grounded manner.
References


Stansfield, Charles W. 2008. *Lecture: 'Where we have been and where we should go'*. *Language Testing* 25 (3):311-326.


———. 2007. JFL learners' pragmatic development and classroom interaction examined from a language socialization perspective. Paper read at Selected Papers from Pragmatics in the CJK Classroom: The State of the Art, June 5 to June 7, 2006, at University of Hawaii‘i-Manoa, Honolulu, Hawai‘i.


Appendix A: Content Specifications for the Grammar Section

TOP 211 most commonly introduced items in five Japanese Language Textbooks widely used in U.S. colleges and universities, in order of appearance, from Itomitsu (2005):

(NOTE: N=noun/nominal, A=adjective/adjectival, V=verb/verbal, NA=adjectival noun/na-adjectives, X=predicate or other elements)

1 desu, masu (polite/distal)
2 ka (question) X desu ka?
3 go (languages)
4 N desu / N de wa arimasen
5 ne?
6 wa (N wa N desu)
7 no (kind, possession): Nihongo no sensee, sensee no kasa
8 mo (addition)
9 kore, sore, are, dore
10 V-masu / V-masen
11 numbers 1-99
12 yo
13 kono, sono, ano, dono
14 Adj-i desu
15 Adj-ku nai desu/ Adj-ku arimasen
16 numbers 100-999
17 NA desu
18 NA de wa arimasen / NA zya nai desu
19 N to N
20 ni (location of existence): koko ni aru
21 numbers 10000-
22 V-ru (plain, direct)
23 numbers 10000-9999
24 en
25 N wa X ni arimasu (existence)
26 ga: N ga V masu/masen (subject, doer of activity)
27 V-masita / V-masen desita
28 koko, soko, asoko, doko
29 N o V masu (direct object, operand)
X ni N ga arimasu (existence)
N desita / N de wa arimasen desita
to (with person): tomodati to simasu
NA deshita / NA de wa arimasen deshita
Adj-katta / Adj-ku nakatta
zi
Adj-i N
hun
de (location of action)
ni (destination): tokyoo ni ikimasu
imasu/arimasu
NA-na N desu
sai
yoobi
mai
V-ru (plain/dictionary form)
kara (time): 3:00 kara
made (place): tokyoo made ikimasu
V-masen ka (invitation)
V-te, … (reason/cause): osoku natte, sumimasen
goro
ni (time): 3:00 ni
zikan
tu
V-masyoo (invitation)
family terms (my family vs. other's family)
V-nai/ V-nakatta
particle + wa/mo (ni wa, de mo, etc.)
V-te kudasai
ni (action toward whom): Tanaka-san ni iu
gurai
hon
da (plain/direct)
o (contracted): Tanaka-san no desu
e: gakkoo e ikimasu
gatu
made (time) 3:00 made
NA de, …
de (means): basu de iku
ga: (object/affect): nihongo ga suki desu
nin
X (da) kara (reason/cause)
N/NA da
V-masyoo ka (volunteering)
N/NA datta
dai
ka/kedo
Adj-ku V-masu
X no/n desu (extended predicate): iku n desu
no (noun/nominal): akai no desu
ya: X ya Y
V-te, V-te, … (sequence of activity)
A wa N de, B wa Y desu (connecting two sentences)
X desyoo (kuru desyoo, gakusee desyoo)
V-ta
satu
ni (ratio/frequency) itiniti ni ikkai
V-te imasu (result)
o (originating point) uti o deru
X to iimasu/yomimasu, etc. (It is called X)
niti
Adj-kute,
N-de, … (cause/reason)
kagetu
N ni narimasu
V-te kudasaimasen ka?
ni (action originated from): Tanaka-san ni morau
N to N to, doti ra ga… (comparing two)
yori: X yori Y no hoo ga… (comparing two)
dake
de itiban… (comparing three or more)
V-te kimasu/ikimasu
ka naa
V-tai desu
mada + aff.
N daroo / N de wa nai daroo
o (area for action) kooen o aruku, kore o iku
ni (purpose) benkyoo si ni iku
Adj-ku narimasu
V-ru no (nominalizing clause): tenisu o suru no ga suki desu
V-te imasu (ongoing activity)
V-te imasu (repeated activity)
dai (number of times)
X (na) no de (reason/cause)
X to omoimasu
N ni simasu
V-ta ato de/ni
V-ru N (sentence modifying noun)
irassyaimasu (come, honorific)
moo + aff.
mada + neg.
V-tagarimasu /V-tagatte imasu
ka sira
X to iimasita/itte imasita (quote)
N desyoo / N de wa nai desyoo
V-ru koto (nominalizing clause): suru koto ga suki desu
V-te mo ii desu (permission)
ka (with Wh-word): nani ka arimasu
Wh + mo: dare mo
Wh + de mo: dare de mo
made ni: 3:00 made ni
NA ni V-masu
V-te wa (tya/zya) ikemasen (prohibition)
V-yasui/nikui
V-ta N (sentence modifying noun)
o-V kudasai (humble)
X ka mo siremasen
V-ru koto ga dekimasu (potential)
X soo desu (hearsay)
o-V simasu (humble)
V-nai de kudasai
ka: X ka Y (X or B)
agemasu / yarimasu
X wa Y ga …: Tokyo wa hito ga ooi, zoo wa hana ga nagai
X de gozaimasu (polite)
moraimasu / itadakimasu
X ka kiku (embedded Wh-question)
V-ta hoo ga ii (suggestion, advice)
V-te kara
V-(ra)remasu (potential)
X-soo desu (appearance)
V-te mimasu (trial)
Adj-ku simasu
V-nakute wa ikemasen (necessity)
X tumori desu / tumori wa nai desu
N to iu N (naming)
V-nai de/V-zu ni…
I-(do) mo (neg)
V-ta koto ga arimasu (experience)
V-ru mae ni
V-te itadakemasen ka?
V-nasai (order)
X tame
X yoo desu (inference)
X-sugiru
V-nakereba ikemasen / narimasen (necessity, duty)
V-tari V-tari simasu
V-te kuremasu / kudasaimasu
V-te moraimasu / itadakimasu
X nara, X ga ii
irassyaimasu (be, honorific)
si: A mo X da si, B mo…
kuremasu / kudasaimasu
V-te okimasu
irassyaimasu (go, honorific)
V-tara,…
mairimasu (go, come; humble)
intransitive vs. transitive
V-tara, doo desu ka? (suggestion)
V-ru/ta/te iru toki (ni)
X mitai desu (appearance)
V-ru koto ni suru / site iru
o-V ni narimasu (honorific)
X ka (doo ka) kiku (embedded Yes-NO question)
ossyaimasu
(number) mo: 100-kai mo (as many as)
sika (+negative)
nasaimasu
V-te agemasu / yarimasu
V-kata: kakikata
X ga hosii
V-ba,…
V-ru yoo ni (suru)
(X to) moosimasu (humble)
messagarimasu
X rasii (conjecture)
V-ru yoo ni naru
V-ru to,…
V-te arimasu (result remaining)
V-te simaimasita (completion)
V-yoo: Ikoo to omou (volitional)
orimasu (humble)
V-nagara
itasimasu (humble)
goran ni narimasu
X hazu desu /hazu wa arimasen
V-nakute mo ii desu (negative permission)
V-te hosii
V-(ra)remasita (passive)
V-(sa)semasita (causative), V-sasete kudasai
no ni: itta no ni…
Appendix B: Content Specifications For The Routines Section

The following expressions have been selected from Sanseido (2002), with consultation of the JLPT content spec あいさつ語等表現 for JLPT Levels 4 and 3 (* indicates Level 4, ** indicates Level 3)

明けましておめでとうございます。
遊びにきてください。
ありがとうございました／ました。*
（お）忙しいところをすみません／悪いんですが...
いただきます。*
いって（い）らっしゃい。**
行ってまいります。**
いつもお世話になります／なっております。
いらっしゃい・ませ。*
いろいろお世話になりました。
おかえりなさい。**
おかげさまで。**
お気の毒様（でした）。
お気をつけて。
お変わりありませんか。
お口に合いますかどうか（分かりませんが、どうぞ）。
お元気で。*
お元気そうで（、何よりです）。
お大事に。**
お出かけですか。
お願いします。*
おはようございます。*
お言葉ではありますか
お言葉に甘えて
お先に失礼します。
お騒がせしました。
お邪魔しました。
お世話様（でした）
お粗末様（でした）。
恐れ入ります。
お疲れ様（です／でした）
お手数をおかけして（、すみません）。
おはようございます。*
お久しぶりです。
お待たせしました。**
お待ちしておりました。
お待ちどう様（です／でした）。
おめでとうございます。**
お役に立てなくて、すみません。
おやすみなさい。*

かしこまりました。**
乾杯！
旧年中はお世話になりました。
恐縮です。
具合はどうですか。
ご安心ください。
ごきげんいかが。
ごきげんよう。
ご苦労様。
このたびはどうも（御挨拶様です）。
ご親切にどうも。
ご心配なく。
ごちそうさま（でした）。*
ごちそうします。
ごちそうにります。
こちらこそ。*
ご丁寧にどうも。
今年もよろしくお願いします。
ご無沙汰しました／しております。
ご迷惑をおかけしました。
ごめんください。*
ごめんなさい。*
こんにちは。*
こんばんは。*

さよ（う）なら。*
失礼します／しました。*
すみません。*
折角ですか／けど...。
折角ですから...。
先日はどうも。
そこを何とか...
それはいけませんね。**
そろそろおいとまを／失礼します。

ただいま。**
ただいま席を外しております。
頂戴いたします。
ちょっとすみません／失礼します。
ちょっといいですか。
ちょっとお時間をいただけますか。
ちょっとそこで。
つまらない物ですが。
では、また。*
どういたしまして。*
どうぞお構いなく。
どうぞお気遣いなく。
どうぞご遠慮なく。
どうぞお楽になさってください。
どうぞごゆっくり。
どうも。*
どちら様ですか。
とんでもない（です）。

何もございませんが（、どうぞ）。
何のお構いもしませんで。
何名様ですか。

はじめまして。*
ぱんざい！

またいらして／いらっしゃってください。

夜分遅くすみません。
よいお年を（お迎えください）。
よくいらっしゃいました。**
（どうぞ、）よろしく。*
喜んで。
Appendix C: Content Specifications For The Speech Acts Section  
(Based on Itomitsu 2008)

Content specification list  
suggesting a course of action (including encouraging and inviting)

どうぞ。
X、どうぞ。
～へ、どうぞ。
どうぞ、～へ。

～てみよう。
～てみて。
～てみますか。
～てみましょう。
～てみてください。

どうぞお～ください。
お～しましょう。
～てもいいんだろう。
X、どうですか。
X は？
X はどう？
X はどうですか。
X はどうでしょう。
（X）、いかがですか。
X はいかがですか。
X はいかがでしょう（か）。
X でもいかがですか。

X がいいでしょうか。
X がいいんじゃないでしょうか。
X がいいな、X が。
（どうぞ）～てください。
（どうぞ）おVください。
（一緒に）Vる？
Vますか。
（一緒に）～ましょう。
～ましょうね。
～ましょうか。
（一緒に）～ませんか。
Xにしませんか。
Xにしましょう。
Vない？
Vではしんだけど。
Qにする？
Q～ましょうか。
Qがいいですか。
V（よ）う（よ）。
V（よ）うか。
Vすることにしょうか。
Vすることにしましょう。
いいじゃありませんか。
Vたほうがいいでしょうか。
もし～たければ～ますけど。

concluding
じゃ、～に（～で）（ね）。
では、またいずれ。

requesting someone to do something (including instructing/directing):
こちらへ、どうぞ。

Xね。

お願い。
X（を）お願いします。
よろしくお願いします。
よろしくお願いいたします。
よろしくお願い申し上げます。
X、お願いできますか。
お願いできないでしょうか。

V（ら）れますか。

お～ください。
～て（よ）。
～ないで（ね／よ）。
～ってくる？
～てください（ね）。
～てくださいませんか。
～いでください（ね）。
～てくれます／ませんか。
～てもらえます／ませんか。
～てもらえいない？
～ていただけます／ませんか。
～ていただけないでしょうか。
～させてくれないかなあ。
～させていただけないでしょうか。

V るようにしてね。
～ておきましょう。
～てもらいましょう。
～たほうがいいですね。
～ておられるようにですねえ。

～てほしいんですけど。
～てほしいということなんです。
～てもらいたいんです（が）。
～ていただきたいんですが。
～たかったんですが。

（X）、頼むよ。
（よろしく）頼んだよ。
～て。
～てみて。
～てほしいんですけど。
～てほしいんだ。
Xをお願いします。
お～ください。
～てくれないかな。
～しておいて。
～なさい。
～てください。
～みてください。
～んだ（よ）。
～るの。

～するものですよ。
～ば、～るから。
～るべきよ。

お願いします。
～てくれない？
～てくれないかな。
～もらえないので。
～ていただけませんか。
～ていただけますか。
～ももらいたいんですけど
～たいんですけど／が。
X、（例：どう行ったらいいか）分からないんだけど。
～んです。

そこを何とか頼むよ。
～そうもなくてさ。な。
悪いな。～からさ。

**offering assistance/service**

～は？
Xはどうでしょうか。
どのXがいいでしょうか。
どんなXがよろしいですか。
X、いかがですか。
Xはいかがですか。
Xはいかがでしょうか。
X はAになさいですか、Bになさいですか。
X になさいますか。
外に何か。
X はどこですか。
何かありますか。

～てもいいですよ。
～てみます。
～ておくわね。
～るよ。
～ましょう。
～ましょうか。
お～いたしますようか。

私でお役に立つことがあったら、おっしゃってください。
～（さ）せてください。

～てあげましょう。
～てあげますから
～てあげますよ。
～てあげるわ。

X がいいでしょう。

よろしかったら～ますが。
お～します。
お～しましょうか。
お～いたします。
お～ですか。（どんな家をお探しですか。）

Advising (this category has been abandoned after the pilot test)

～はどう？
～はどうですか。
これなんか、どうかな。

X でしょうか？
X だそうですよ。
～るべきですよ。
～ろんですよ。
～がいい（わ）ね。
～はいいよね。
～がいいですよ。
～したほうが。
～たほうがいい（わ）よ。
～たほうがいいですね。
～たほうがいいんじゃないかしら。
～たほうが便利よ。
～たほうが確か／楽ですよ。
～たほうがいいと思います。
～たほうがいいと思いますけどね。
～たほうがいいかもしれませんね。
～たほうがよさそうだねえ。
～ないほうがいいですよ。
～ないほうがいいでしょう。
～たら。
～たらどうですか。
～たいいですよ。
～てみたら。
～てみたらどう？
～てみたらどうでしょう。
～ことはないだろう。
～てもらえるんじゃないかしら。
～ちかもしれないんです。
～れば、～らわよ。
～れば、～ますよ。

（命令形）よ。（断っちゃえよ、そんなの）（謝ってらっしゃいよ）。
（無理すんなよ）
Appendix D: Content Specifications For The Speech Styles Section
(Based on the JLPT Test Specification, Japan Foundation 1994)

HONORIFIC (29):

lexical: (14)
JLPT 3 (11): いらっしゃる、おいでになる、おいでください、お越しください
いい、おっしゃる、くださる、ご存知、ご覧になる、ご覧ください、なさる、
みえる
JLPT 2 (3): あがる、お越しになる、ご覧ください
productive (15):
JLPT 3 (7): ていらっしゃる、お/ご/になる、てくださる、（さ）
てください、お/ご/ください、（なさる、（ら）れる
JLPT 2 (8): ていらっしゃる、くっていらっしゃる、ておいでになる、
お/ご/だ/です、お/ご/になれる、（さ）てくださる、お/ご/く
くださる、お/ご/なさる

HUMBLE (31):

lexical (18):
JLPT 3 (9): いただす、いただく、伺う、おる、さしあげる、拝見する、参る、
申し上げる
JLPT 2 (9): あがる、承る、お目にかかる、お目にかける、ご覧に入れる、
存じる/存ずる、存じ上げる、頂戴する、拝借する、
productive (13):
JLPT 3 (6): いたす、お/ご/いたす、ていただく、お/ご/する、
ておる、てさしあげる、
JLPT 2 (7): て（き）ていただく、お/ご/いただく、お/ご/願う、お
/ご/できる、拝～、てまいる、お/ご/申し上げる

NEUTRAL (3):
ございません、～でございます、ADJございます（お寒うございます）
Appendix E: Solicitation E-Mail Message  
(OSU IRB Exempt #2008E0706)

We are conducting a research on language testing, and currently seeking volunteers (over 18 years old) who are studying or has studied Japanese as a foreign language at a college level. We would like to gain a better understanding of the test items we have created. Your participation is strictly voluntary. You may refuse to participate or discontinue participation at any time without penalty. If you decide to participate in the survey, your feedback will give us invaluable information about the test items.

The online survey will take approximately 60 minutes of your time. The survey consists of two parts. The first part asks about your learning backgrounds. The second part contains 48 multiple-choice test items. As a small token of appreciation, $10 will be provided upon completion of your survey.

We state that you will be at no psychological, social, physical, or legal risk except that you may experience the amount of stress you may normally feel reading or listening to Japanese. We also state that any and all information we receive will be kept strictly confidential and will only be seen by authorized members of the investigation team. Data gathered from the survey will be coded and summarized in the aggregate. The aggregated results of our analysis will be shared with departments at the Ohio State University and others interested in providing services to education, statistics, and language testing.

You can start the survey from this link: <URL provided by a commercial survey tool> Please contact Masayuki Itomitsu at itomitsu.1@osu.edu if you have any questions about the survey.

Thank you for taking the time to help us with the project.
Sincerely,

Principal Investigator: Mari Noda, Associate Professor  
Co-investigator: Masayuki Itomitsu, Ph.D. candidate  
The Ohio State University
Appendix F: The Operationalized Test (Sample Test Items and Eliminated Items)

=============================================  
Part 1: About You  
=============================================  

Please indicate your consent to participation below:  
1. ( ) I have read and understood the terms above, and consent to participate in this survey.

Instruction  
This survey consists of the following parts:  
Part 1: About You  
Part 2: The Test  
   Section 1: Grammar (12 items)  
   Section 2: Routines (12 items)  
   Section 3: Speech Acts (12 items)  
   Section 4: Speech Styles (12 items)  

The survey takes approximately 60 minutes to complete. Instructions regarding the payment will be provided after you complete the test.

2. Please type in your last name. ________________________________  

3. Please type in your first name: ________________________________  

4. Your e-mail address: ________________________________  

5. May I contact you by e-mail after the survey if we have questions?  
   ( ) Yes ( ) No  

6. Age: ________________________________  

7. Hometown (e.g., Okinawa, Japan): ________________________________
8. Gender
   ( ) Male ( ) Female

9. Your native language: ____________________________________________

10. Name of the institution where you are currently studying (or have studied) Japanese:
    ____________________________________________

11. Your academic standing: ____________________________________________

12. Your major: ____________________________________________

13. Which year/level of Japanese are you currently in?
    ____________________________________________

14. Have you taken OPI (Oral Proficiency Interview) or similar interview test within the last month? What was the rating?
    ____________________________________________

15. How long have you been studying Japanese?
    ( ) 1 year or less ( ) 2 years ( ) 3 years ( ) 4 years ( ) 5 years or more

16. How do you rate your ability in Japanese? Choose the ones that best describes you:
    ( ) Can ask questions or make statements only with memorized words or phrases (ILR 0+)
    ( ) Able to satisfy minimum courtesy requirements and maintain very simple face-to-face conversations on familiar topics. (ILR 1)
    ( ) Can initiate and maintain predictable face-to-face conversations and satisfy limited social demands. (ILR 1+)
    ( ) Able to satisfy routine social demands and limited work requirements. (ILR 2)
    ( ) Able to satisfy most work requirements with language usage that is acceptable and effective. (ILR 2+)
    ( ) Able to participate effectively in most formal and informal conversations in practical, social and professional topics. (ILR 3)
    ( ) Able to satisfy professional needs in a wide range of sophisticated and demanding tasks. (ILR 3+)
    ( ) Able to successfully satisfy all levels normally pertinent to professional needs, with well-organized discourse, appropriate rhetorical speech devices, and native cultural references. (ILR 4, 4+)
    ( ) Functionally equivalent to a highly articulate well-educated native speaker. (ILR 5)

17. Have you lived in Japan for more than 6 months?
    ____________________________________________

Part 1: About You Completed
Thank you for answering the questions in Part 1: About You.
Go on to the next page to start the Test sections.

============================================= 
Part 2: The Test 
============================================= 

Please answer all questions even when you are not sure (there is no penalty for guessing). Also, please answer the questions on your own: do not consult textbooks or dictionaries, and never ask your teacher or a native speaker of Japanese for help. Also, think of the conversations in the test items as normal, everyday talk between normal everyday people (not people being extremely rude, sarcastic, funny, etc.).

============================================= 
Grammar: directions and example 
============================================= 

DIRECTIONS: 
This section tests how much you know about accurate usage of various structural patterns. There are 12 items in this section. 
In each item, a brief description in English and an illustration of the context are provided. These are followed by an utterance with a blank, and four possible options. The utterance and options are provided in Japanese, both in audio and in writing. You will be asked what one of the people would probably say in the situation. Select the best answer from the four choices to fill in the blank.

Example: 
Tetsuya, a college student, is talking with his friend. He is asking where she went yesterday.

What would Tetsuya probably say to his friend?

Read and/or listen to the Japanese:
The correct answer is A. Go on to the next page to begin this section.

GRAMMAR: SAMPLE ITEMS

10: Ms. Noguchi, an employee, is talking with her coworker about the DVD player.

What would Ms. Noguchi probably say to her coworker?

Read and/or listen to the Japanese:

あ、DVD プレーヤーが______いますよ。
11: Ms. Sato, an employee, is talking with her coworker. They are trying to identify who the new manager is.

What would Ms. Sato probably say to her coworker?

Read and/or listen to the Japanese:

A. そう
B. よう
C. みたい
D. らしい

GRAMMAR: ELIMINATED ITEM

GR-D: Ms. Koda, an employee, is talking with her coworker. She is complaining during a training session.

What would she probably say to her coworker?
Routines: directions and example

DIRECTIONS:
This section tests how much you know about language use in specific situations. There are 12 items in this section. In each item, a brief description in English and an illustration of the context are provided. These are followed by an utterance with a blank, and four possible options. The utterance and options are provided in Japanese, both in audio and in writing. You will be asked what one of the people would probably say in the situation. Select the best answer from the four choices to fill in the blank.

Example:
Takashi, a college student, is thanking his friend for a cup of tea.

What would Takashi probably say to his friend?

Read and/or listen to the Japanese:
A. どうぞ
B. どうも
C. いいえ
D. よろしく

The correct answer is B. Go on to the next page to begin this section.

**ROUTINES: SAMPLE ITEMS**

1: Mr. Abe, an employee, is talking with his wife, as he leaves his house for work.

What would Mr. Abe probably say to his wife?

Read and/or listen to the Japanese:
「行ってらっしゃい。」

A. ただいま
   かえ
B. お帰りなさい
   い
C. 行ってきます
   い
D. 行ってらっしゃい

11: Ms. Nakamura, a housewife, is serving dinner to a guest at her house.

What would Ms. Nakamura probably say to the guest?

Read and/or listen to the Japanese:

「_____、どうぞ。」
「あ、どうもすみません。」

おそ い
A. 恐れ入りますが
B. せっかくですが
   なに
C. 何もございませんが
   かま
D. お構いなく

ROUTINES: ELIMINATED ITEM
RT-B: Mr. Nomoto, an employee, is talking with his neighbor. Her grandmother has recently passed away due to illness.

What would he probably say to his neighbor?

Read and/or listen to the Japanese:

```
たび ほんとう
「この度は、本当にどうも______。」
```

しょうしよう
A. ご愁傷さまでした
さわ
B. お騒がせしました
もう わけ
C. 申し訳ございません
ぶたた
D. ご無沙汰しております

Speech Acts: directions and example

DIRECTIONS:
This section tests how much you know about how people express themselves in everyday conversations in Japanese. There are 12 items in this section. In each item, a brief description in English and an illustration of the context are provided. These are followed by an utterance with a blank, and four possible options. The utterance and options are provided in Japanese, both in audio and in writing. You will be asked
what one of the people would probably say in the situation. Select the best answer from the four choices to fill in the blank.

**Example:**
Mr. Yamamoto, an employee, is talking with his coworkers after playing golf on a weekend. One of them complimented him on how well he plays. He is **reacting to the compliment**.

What would Mr. Yamamoto probably say to his coworkers?

Read and/or listen to the Japanese:

え、______ そんなことないですよ。

A. そうですね？
B. そうですか？
C. そうですよ。
D. そうですね。

The correct answer is B. Go on to the next page to begin this section.

**SPEECH ACTS: SAMPLE ITEMS**

1: Mr. Sato, an employee, is having trouble with some documents. His supervisor stops by and asks if he needs assistance. He is **requesting** that he receive some instructions.

What would Mr. Sato probably say to his supervisor?

Read and/or listen to the Japanese:
Mr. Takeshita, an employee, is talking with his coworker on the street. He is suggesting that they have something cold to drink.

What would Mr. Takeshita probably say to his coworker?

Read and/or listen to the Japanese:
SPEECH ACTS: ELIMINATED ITEM

SA-A: Ms. Shimizu, an employee, is talking with her coworker. She has learned he has been working for extra hours. She is advising that he not overdo it.

What would she probably say to her coworker?

Read and/or listen to the Japanese:

でも、_____いいですよ。

A. 無理しなければ
B. 無理しなくても
C. 無理しなかったら
D. 無理しないほうが

Speech Styles: directions and example

DIRECTIONS:

This section tests how much you know about the appropriate use of various polite speech styles in Japanese. There are 12 items in this section.
In each item, a brief description in English and an illustration of the context are provided. These are followed by an utterance with a blank, and four possible options. The utterance and options are provided in Japanese, both in audio and in writing. You will be asked what one of the people would probably say in the situation. Select the best answer from the four choices to fill in the blank.

Example:
Ms. Kinoshita, an employee, is informing a visitor where the elevator is located.

What would Ms. Kinoshita probably say to the visitor?

Read and/or listen to the Japanese:

あ、あちらに______。

まい
A. 参ります
B. おります
C. ございます
D. いらっしゃいます

The correct answer is C. Go on to the next page to begin this section.

SPEECH STYLES: SAMPLE ITEMS

2: Mr. Yoshida, father of Hiromi, is answering a phone. He is informing the caller that Hiromi is not at home.

What would Mr. Yoshida probably say to the caller?
Read and/or listen to the Japanese:

いま
あ、ひろみは今ちょっと______。

A. おりませんが
B. おいでになりませんが
C. いらっしゃいませんが
D. お見えになりませんが

10: Ms. Saito, an employee, is making an announcement at a department store.

What would Ms. Saito probably say in her public announcement?

Read and/or listen to the Japanese:

やまぐちさま　やまぐちさところさま　しょうめんげんかん
A. お越しください
B. お越しになって
C. お伺いできますか
D. こちらてちょうだい

SPEECH STYLES: ELIMINATED ITEM

SS-C: Kenji, a college student, is talking with his teacher. He is giving the teacher a gift that his mother sent him.

What would he probably say to his teacher?

Read and/or listen to the Japanese:

いなか はは
これ、田舎の母からです。よろしくと_____。

A. 申し上げました
B. 申ししておりました
C. おっしゃいました
D. 申し上げставил
Thank you for your participation!!
This is the end of the survey. Your input will be highly appreciated.