

**EFFECTS OF L2 AFFECTIVE FACTORS ON
SELF-ASSESSMENT OF SPEAKING**

A Dissertation
Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Education

By
Noriko Iwamoto
May, 2015

Examining Committee Members

James Sick, Advisory Chair, New York University
Tomoko Nemoto, Teaching and Learning
David Beglar, Teaching and Learning
James Elwood, External Member, Meiji University
Kazuya Saito, External Member, Waseda University

©

Copyright

2015

by

Noriko Iwamoto

ABSTRACT

This study was an investigation of the validity of students' self-assessment of L2 oral performance, the influences of L2 affective variables on their self-assessment bias, and the degree to which the influences of L2 affective variables differ between high and low proficiency learners. The participants were 389 science majors from two private Japanese universities. A questionnaire was administered using items based on the Attitude/Motivational Test Battery (Gardner, 1985), the Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986), the Rosenberg Self-Esteem Scale (Rosenberg, 1965), Sick and Nagasaka's (2000) Willingness to Communicate Scale, and items designed to measure motivation adapted from Gardner, Tremblay, and Masgoret (1997), Yashima (2002), Irie (2005), and Matsuoka (2006). A factor analysis identified seven factors in the questionnaire data: Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence. The scales were further validated using the Rasch rating scale model.

Student oral interviews were recorded and rated by five English teachers using an oral assessment scale based on the Kanda English Proficiency Test (Bonk & Ockey, 2003). Immediately after the interviews were completed, the participants were asked to consider how they perceived their own speaking performance and they rated their own performance from their memory using the same oral assessment scale that the teacher raters used. The oral assessment scale included the

descriptions of the oral performances that match each level. The participants read the descriptions of each level and chose a level that they thought matched their own performance.

The study produced four main findings. First, a multi-faceted Rasch analysis revealed that the participants rated their own L2 speaking more severely than the teacher raters and that the students' self-assessments were neither reliable nor consistent.

Second, self-assessment bias measures were calculated and used to test a hypothesized structural model of how affective factors influenced self-assessment bias. The hypothesized model showed poor fit to the data, possibly due to the poor reliability of the self-assessment measures. Multiple regression analyses conducted as a follow-up analysis revealed that participants with greater Desire to Learn to Speak English tended to underestimate and those with greater L2 Speaking Self-Confidence tended to overestimate their own speaking performance.

Third, 106 participants whose self-ratings were similar to the teachers' ratings were compared with other students in order to examine their distinctive features. However, no significant differences in L2 oral proficiency or affective variable measures were found between the two. Therefore, those whose self-assessments agreed with teachers could have resulted in some agreements that occurred by chance alone.

Finally, 100 higher proficiency students were compared with 100 lower proficiency students and the results showed that the higher proficiency students

with greater Desire to Learn to Speak English generally underestimated their L2 speaking proficiency, while those with higher Self-Esteem and greater L2 Speaking Self-Confidence tended to overestimate it. Lower proficiency students with greater L2 Speaking Self-Confidence tended to overestimate their L2 speaking proficiency.

The results suggest that the self-assessment of L2 speaking might not be a sufficiently reliable or consistent assessment tool. Therefore, if teachers are considering including self-assessment in a speaking class, self-assessment training should be conducted. Additionally, giving L2 learners more opportunities to speak the L2 can help them notice gaps between their productions and those of proficient speakers, which might lead to more accurate self-assessment. Second, although some studies utilized only one teacher-rater, five teacher raters in this study displayed a great deal of diversity and exhibited unique bias patterns, so multiple raters should be employed and Facets analyses should be employed because the multi-faceted Rasch model provides person ability estimates that are adjusted for rater bias. Finally, the use of multi-faceted Rasch analysis is useful for examining oral data because unlike raw scores, multi-faceted Rasch analysis provides detailed information concerning speaker ability, rater severity, and category difficulty. Moreover, while most researchers have utilized self-assessment raw scores, in this study bias measures of self-assessment were calculated using Facets, which indicated that the bias measures produced different outcomes compared with self-assessment scores.

ACKNOWLEDGMENTS

First, I would like to express my very great appreciation to Dr. David Beglar, who has helped me write my dissertation from the start of my work, including its planning and development. He has given me tremendous advice and suggestions for my dissertation drafts and also helped me analyze the statistical data. I appreciate his vast knowledge and skills in many areas, and his editing assistance has improved my writing.

I would also like to express my sincere gratitude to Dr. James Sick, my main dissertation advisor, who has assisted me with statistical data analysis using Facets and helped me to interpret the findings. I was very much impressed by his deep and profound knowledge of data analysis. Without his support and technical assistance, I would not have been able to obtain the results described in this dissertation.

My great appreciation is also extended to my dissertation defense committee members. Dr. Kazuya Saito shared his distinguished paper with me, and it provided me with new insight into L2 speaking and self-assessment research. Thanks to his advice, I was able to improve my dissertation during the revision. Dr. Jim Elwood read my manuscript very carefully and kindly gave me a list of many helpful comments prior to the defense, which helped me greatly in my preparation. Dr. Tomoko Nemoto not only gave helpful advice and suggestions on my dissertation during the dissertation defense, but has also supported me in my courses since I started at TUJ.

I am thankful to 400 students who voluntarily agreed to participate in my study and allowed me to obtain valuable data. I am indebted to five raters who contributed their precious time during summer vacation to the assessment of speaking performance. I also wish to acknowledge the help provided by my colleagues, Professor Hiroyo Yoshida and Professor Michael Schulman, for gathering data and editing my writing. I also would like to thank my cohort members at TUJ, especially Reiko Yoshihara and Chiyo Hayashi, who gave me warm encouragement during the dissertation defense.

Finally, I would like to thank my parents for their support both mentally and financially throughout my study. Without their help and encouragement, I would not have finished this dissertation.

TABLE OF CONTENTS

	PAGE
ABSTRACT	iii
ACKNOWLEDGMENTS	vi
LIST OF TABLES	xiv
LIST OF FIGURES	xix
CHAPTER	
1. INTRODUCTION.....	1
The Background of the Issue.....	1
Statement of the Problem.....	4
Purposes of the Study.....	7
Significance of the Study	9
The Audience for the Study	10
Delimitations.....	11
The Organization of This Study.....	12
2. REVIEW OF THE LITERATURE	14
Self-Assessment	14
Beneficial Effects of Self-Assessment	14
Self-Assessment Bias	15
Validity of Self-Assessment of L2 Skills	16
Section Summary.....	24
Self-Esteem	25
Culture and Self-Esteem	26

Self-Esteem and L2 Skills	31
Self-Esteem and the Self-Assessment of L2 Skills	32
Section Summary.....	34
Language Anxiety	35
Language Anxiety and L2 Proficiency	37
Language Anxiety and the Self-Assessment of L2 Speaking.....	43
Section Summary.....	46
Willingness to Communicate	47
Willingness to Communicate and Perceived Competence	52
Section Summary.....	56
Motivation	57
Motivation and the Self-Assessment of L2 Proficiency	61
Section Summary.....	64
Self-Confidence.....	64
Section Summary.....	69
Oral Proficiency	69
Oral Proficiency Assessment.....	72
Kanda English Proficiency Test	73
Gaps in the Literature.....	75
Purposes of the Study.....	79
Research Questions	83
3. METHODS.....	84
Participants.....	84
Raters.....	86

Instrumentation	87
Self-Esteem Scale	88
L2 Speaking Anxiety Scale	89
L2 Willingness to Communicate Scale	90
L2 Speaking Motivation Scale	91
L2 Speaking Self-Confidence Scale	92
Oral Assessment Scale.....	93
Rasch Model	95
Rasch Rating Scale Model.....	95
Item Fit Analysis	96
Principal Component Analysis (PCA) of Item Residuals	98
Rasch Reliability Estimates.....	98
Rasch Separation Index	99
The Many-Facet Rasch Model (MFRM).....	100
Procedures	102
Data Collection	102
Data Analysis.....	103
4. PRELIMINARY ANALYSIS	107
Factor Analysis Results.....	107
Rasch Analysis.....	115
Self-Esteem.....	116
L2 Speaking Anxiety.....	121
L2 Willingness to Communicate.....	128
L2 Speaking Self-Confidence.....	133

L2 Speaking Motivation Items	139
L2 Speaking Motivational Intensity	139
Attitude Toward Learning to Speak English and Desire to Learn to Speak English.....	145
Attitude Toward Learning to Speak English	146
Desire to Learn to Speak English	156
Comparison of the SPSS Factor Analysis and the Rasch Analysis Results	163
Facets Analysis of the Speaking Data	165
5. RESULTS.....	172
Facets Analysis of the Speaking Data.....	172
Research Question 1: The Validity of Self-Assessment	174
Research Question 2: The Influence of L2 Affective Variables on Self-Assessment	181
Research Question 3: Characteristics of Those Who Conducted Accurate Self-Assessment.....	189
Research Question 4: The Influence of L2 Affective Variables on Self-Assessment of High and Low Proficiency Participants	190
High Proficiency Group.....	192
Low Proficiency Group	193
Summary	195
6. DISCUSSION.....	196
Research Question 1: The Validity of Self-Assessment	196
Research Question 2: The Influence of L2 Affective Variables on Self-Assessment	208
Research Question 3: Characteristics of Those Who Conducted Accurate Self-Assessment.....	216

Research Question 4: The Influence of L2 Affective Variables on Self-Assessment of High and Low Proficiency Participants	218
Theoretical Implications.....	223
Pedagogical Implications	225
7. CONCLUSION	228
Summary of the Findings.....	228
Limitations of the Study.....	229
Suggestions for Future Study.....	231
Final Conclusions.....	232
REFERENCES.....	234
APPENDICES	
A. CONSENT FORM	249
B. ORAL INTERVIEW FRAMEWORK	250
C. SPEAKING ASSESSMENT RUBRIC (ENGLISH VERSION).....	251
D. SPEAKING ASSESSMENT RUBRIC (JAPANESE VERSION)	252
E. SELF-ESTEEM SCALE (ENGLISH VERSION)	253
F. SELF-ESTEEM SCALE (JAPANESE VERSION).....	254
G. L2 SPEAKING ANXIETY SCALE (ENGLISH VERSION)	255
H. L2 SPEAKING ANXIETY SCALE (JAPANESE VERSION).....	256
I. L2 WILLINGNESS TO COMMUNICATE SCALE (ENGLISHVERSION).....	257
J. L2 WILLINGNESS TO COMMUNICATE SCALE (JAPANESE VERSION).....	258
K. L2 SPEAKING MOTIVATION SCALE (ENGLISH VERSION).....	259
L. L2 SPEAKING MOTIVATION SCALE (JAPANESE VERSION)	261

M. L2 SPEAKING SELF-CONFIDENCE SCALE (ENGLISH VERSION)	263
N. L2 SPEAKING SELF-CONFIDENCE SCALE (JAPANESE VERSION)	264
O. ORAL INTERVIEW TRANSCRIPTS FOR EACH LEVEL PARTICIPANT (ENGLISHVERSION).....	265

LIST OF TABLES

Table	Page
1. Criteria for Unidimensionality (Linacre, 2007).....	98
2. Pattern Matrix of the Questionnaire Items	112
3. Category Separation Criteria (Wolfe & Smith, 2007).....	115
4. Six-Point Rating Scale Functioning for Self-Esteem	116
5. Rasch Item Statistics for the Self-Esteem Items.....	117
6. Rasch Item Statistics for the Self-Esteem Items Excluding Item SE7	118
7. Rasch Principal Component Analysis for the Self-Esteem Items	120
8. Six-Point Rating Scale Functioning for L2 Speaking Anxiety.....	121
9. Five-Point Rating Scale Functioning for L2 Speaking Anxiety.....	121
10. Four-Point Rating Scale Functioning for L2 Speaking Anxiety	122
11. Three-Point Rating Scale Functioning for L2 Speaking Anxiety.....	122
12. Rasch Item Statistics for the L2 Speaking Anxiety Items	124
13. Rasch Item Statistics for the L2 Speaking Anxiety Items Excluding Item ANX10.....	124
14. Rasch Item Statistics for the L2 Speaking Anxiety Items Excluding Items ANX10 and ANX2	125
15. Rasch Principal Component Analysis for the L2 Speaking Anxiety Items....	127
16. Six-Point Rating Scale Functioning for L2 WTC	128
17. Rasch Item Statistics for the L2 WTC Items.....	129
18. Rasch Item Statistics for the L2 WTC Items Excluding Item WTC12	130
19. Rasch Item Statistics for the L2 WTC Items Excluding Items WTC12 and WTC1.....	130
20. Rasch Principal Component Analysis for the L2 WTC Items.....	133

21. Six-Point Rating Scale Functioning for L2 Speaking Self-Confidence	133
22. Five-Point Rating Scale Functioning for L2 Speaking Self-Confidence	134
23. Rasch Item Statistics for the L2 Speaking Self-Confidence Items.....	135
24. Rasch Item Statistics for the L2 Speaking Self-Confidence Items Excluding Item SC11.....	136
25. Rasch Principal Component Analysis for the L2 Speaking Self- Confidence Items	138
26. Rasch Principal Component Analysis for the MI, ALSE, and DLSE Items	140
27. Six-Point Rating Scale Functioning for L2 Speaking Motivational Intensity	140
28. Rasch Item Statistics for the L2 Speaking Motivational Intensity Items	141
29. Rasch Item Statistics for the L2 Speaking Motivational Intensity Items Excluding Item MI5.....	142
30. Rasch Principal Component Analysis for the L2 Speaking Motivational Intensity Items	145
31. Rasch Principal Component Analysis for the ALSE and DLSE Items	146
32. Six-Point Rating Scale Functioning for Attitude Toward Learning to Speak English	147
33. Five-Point Rating Scale Functioning for Attitude Toward Learning to Speak English	147
34. Four-Point Rating Scale Functioning for Attitude Toward Learning to Speak English	148
35. Three-Point Rating Scale Functioning for Attitude Toward Learning to Speak English	148
36. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items	150
37. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Item ALSE5.....	150

38. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5 and ALSE11	151
39. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5, ALSE11, and ALSE12	152
40. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5, ALSE11, ALSE12, and ALSE13	153
41. Rasch Principal Component Analysis for the Attitude Toward Learning to Speak English Items	155
42. Six-Point Rating Scale Functioning for Desire to Learn to Speak English....	156
43. Five-Point Rating Scale Functioning for Desire to Learn to Speak English	156
44. Four-Point Rating Scale Functioning for Desire to Learn to Speak English	157
45. Rasch Item Statistics for the Desire to Learn to Speak English Items	158
46. Rasch Item Statistics for the Desire to Learn to Speak English Excluding Item DLSE10	158
47. Rasch Item Statistics for the Desire to Learn to Speak English Including Items ALSE5, ALSE12, and ALSE13	159
48. Rasch Item Statistics for the Desire to Learn to Speak English Including Items ALSE12 and ALSE13	160
49. Rasch Item Statistics for the Desire to Learn to Speak English Including Item ALSE13	160
50. Rasch Principal Component Analysis for the Desire to Learn to Speak Items	163
51. A Comparison of the Constructs Identified by Rasch Analysis and SPSS Factor Analysis	164
52. Category Statistics for the Step Difficulties of the Rating Scale	167
53. Calibration Report for Two Teacher-Raters	170
54. Calibration Report for Six Students	170
55. Calibration Report for Categories	171

56. Category Statistics for the Step Difficulties of the Nine-Point Rating Scale.....	173
57. Calibration Report for Teacher- and Self-Raters.....	177
58. Calibration Report for Categories.....	178
59. Pearson Correlations Among Self-Assessment and Teacher-Assessment Measures.....	180
60. Pearson Correlations Among Five Teacher-Raters	180
61. Descriptive Statistic for Each Affective Variable	181
62. Pearson Correlations Between Oral Performance Measures (SA and TA) and Affective Variables	182
63. Pearson Correlations Between Oral Performance Measures (Bias) and Affective Variables.....	183
64. Multiple Regression Predicting Teacher-Assessment from Affective Variables.....	183
65. Multiple Regression Predicting Teacher-Assessment of Grammar, Vocabulary, Fluency, and Pronunciation with Affective Variables.....	184
66. Selected Rasch Fit Statistics with Bias Size.....	186
67. Results of the Paths with Bias Size	186
68. Multiple Regression Predicting Bias Size from Affective Variables.....	187
69. Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables.....	188
70. Comparison of Accurate and Inaccurate Groups for Speaking Proficiency...	189
71. Comparison of Accurate and Inaccurate Groups for Affective Variables.....	190
72. Comparison of High and Low Proficiency Groups for Affective Variables.....	191
73. Multiple Regression Predicting Bias Size of Self-Assessment from Affective Variables with High Proficiency Group.....	192

74. Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables with High Proficiency Group.....	193
75. Multiple Regression Predicting Bias Size of Self-Assessment from Affective Variables with Low Proficiency Group.....	194
76. Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables with Low Proficiency Group.....	194

LIST OF FIGURES

Figure	Page
1. A portion of MacIntyre’s (1994) willingness to communicate model.....	48
2. MacIntyre and Charos’ (1996) L2 WTC model	49
3. MacIntyre et al.’s (1998) L2 WTC pyramid model	50
4. A portion of Gardner’s (1985) socio-educational model	59
5. Tremblay and Gardner’s (1995) model of L2 motivation.....	60
6. Schematic representation of individual meditational processes (Clément, 1980).	67
7. Aspects of production (Hughes, 2002)	70
8. Hypothesized structural model for the self-assessment bias size of L2 speaking.	105
9. Category probability curves for the six-point rating scale for Self- Esteem.....	116
10. Item-person map for Self-Esteem	119
11. Category probability curves for the five-point rating scale for L2 Speaking Anxiety	123
12. Item-person map for L2 Speaking Anxiety	126
13. Category probability curves for the six-point rating scale for L2 WTC.....	128
14. Item-person map for L2 WTC	132
15. Category probability curves for the five-point rating scale for L2 Speaking Self-Confidence.....	134
16. Item-person map for L2 Speaking Self-Confidence	137
17. Category probability curves for the six-point rating scale for L2 Speaking Motivational Intensity	141
18. Item-person map for L2 Speaking Motivational Intensity.....	144

19.	Category probability curves for the three-point rating scale for Attitude Toward Learning to Speak English.....	149
20.	Item-person map for Attitude Toward Learning to Speak English.....	154
21.	Category probability curves for the four-point rating scale for Desire to Learn to Speak English	157
22.	Item-person map for Desire to Learn to Speak English.....	162
23.	Rating scale category probability for the rating scale.....	168
24.	Facets map for two teacher-raters	169
25.	Rating scale category probability for the nine-point rating scale	174
26.	Facets map for teacher-raters and self-raters	176
27.	Bias interaction (Average observation).....	178
28.	Bias interaction (Relative to overall measure).....	179
29.	Path analysis results with bias size	185

CHAPTER 1

INTRODUCTION

The Background of the Issue

Although teacher assessment has long been considered the primary way of assessing students, student self-assessment has been the focus of research attention for a number of years. Self-assessment has been used in many fields, such as psychology, sociology, and business, but it was only during the 1980s that its potential use in English as a Second Language (ESL) and English as a foreign Language (EFL) contexts became the focus of attention (Oscarson, 1997). Brown (2005) defined self-assessment in second or foreign language contexts as “any items wherein students are asked to rate their own knowledge, skills, or performances. Thus, self-assessments provide the teacher with some idea of how the students view their own language abilities and development” (p. 58). Many researchers have emphasized the importance and usefulness of self-assessment in language learning, as it is widely believed to increase learners’ autonomy and motivation for learning (Brown & Hudson, 1998; Kusnic & Finley, 1993; Oscarson, 1989; Todd, 2002).

Some teachers have successfully introduced self-assessment into foreign language classrooms. Halbach’s (2000) students kept learner diaries from which she could see how her students viewed their in-class L2 performance. Myers (2001), who asked students to self-assess their journal writing, stated that this task

“involved students in an increased self-monitoring of their writing skills which led them to increased insights into their strengths and weaknesses as writers” (p. 487).

Some Japanese teachers have also used self-assessment in their classrooms. Abiko (1997), who was teaching English at Inawa Junior High School in Gifu, Japan, conducted a three-year self-assessment project. Students wrote comments about how much they had learned at the end of every class. This activity allowed the teacher to better understand which areas the students had difficulty understanding and to better reflect the students’ voices in his teaching. Consequently, more students were able to follow the lessons. Similar to Abiko, Inoue (1997) asked the university students in an Educational Psychology course he was teaching to self-assess what they had learned in that day’s class. Through the students’ comments, he was able to better understand the degree of the students’ comprehension of each lesson, so in the next class he could explain difficult aspects of the course more clearly.

LeBlanc and Painchaud (1985) reported that the University of Ottawa successfully utilized self-assessment questionnaires instead of placement tests in order to place students in classes based on their L2 proficiency. The researchers maintained that self-assessment questionnaires have several advantages over standard proficiency tests. First, students can answer self-assessment questionnaires more easily and in less time compared with completing proficiency tests. Second, it is not necessary to prepare a room, set up a testing schedule, or recruit proctors in

order to conduct self-assessments; students can even complete a self-assessment questionnaire at home.

Many studies have been conducted examining the reliability of self-assessments by calculating the correlations between self-assessment questionnaires and proficiency test scores for a variety of L2 skills. Many studies showed significant correlations between them; however, there was always a degree of error that was not entirely due to the self-assessment instrument; rather, it was caused by the student raters themselves as they engaged in self-assessment. Some researchers have suggested that the influence of affective variables causes part of the bias present in self-assessments (AlFallay, 2004; MacIntyre, Noels, & Clément, 1997), possibly because, as Butler and Lee (2006) put it, “self-assessment is a very complex cognitive and metacognitive activity” (p. 514).

The way in which foreign language learners assess themselves might also be influenced by culture because how people view themselves differs in different cultural contexts. For instance, according to Kitayama and Markus (2000), Americans tend to identify positive features in themselves and believe that they are slightly better than their peers. In contrast, many Japanese are self-critical rather than self-enhancing because they place considerable value on harmony and therefore do not wish to stand out too much. This possibility has received empirical support, as a number of researchers have reported that Japanese self-evaluate their traits and abilities lower than North Americans (Brown, 2005, 2006; Heine, Kitayama, & Lehman, 2001; Matsumoto & Kitayama, 1998).

Statement of the Problem

How learners view their own abilities is an important issue in language education; however, because the number of studies of the self-assessment of L2 speaking skills is still limited, a number of issues in this area have not yet been researched thoroughly. Three problems are addressed in this study.

The first problem is that many studies investigating the validity of self-assessment of L2 speaking ability have not measured the participants' L2 speaking skills accurately. For example, some researchers have utilized indirect testing such as proficiency tests (Le Blanc & Painchaud, 1985; Yashima, 2002), course grades (Chen, Horwitz, & Schallert, 1999; Clément et al. 1994), and cloze tests (Gardner & MacIntyre, 1993; Le Blanc & Painchaud, 1985). However, direct testing is considered better than indirect testing because it evaluates speaking skills in actual performance (Ginther, 2012).

Other researchers tested participants' L2 speaking skills directly by having them speak the L2; however, there is sometimes a mismatch between speaking tests and self-assessment of speaking. For example, Peirce et al. (1993) asked the participants to comment on the strictness of the parents during the speaking test, while for self-assessment the participants answered if they can explain the plot of a mystery book or movie using the scale from 1 (*Not at all*) to 5 (*Without any difficulty*). Moreover, in MacIntyre et al. (1997), the raters assessed the number of ideas expressed and the quality of the French for the participants' L2 performance, while the participants answered can-do questions as self-assessment. Therefore, the

discrepancy caused by the use of different assessment rubric between teacher-assessment and self-assessment might bring the bias.

Additionally, many researchers did not include multiple raters but utilized only one teacher rater (e.g., Chen, 2008; Jafarpur, 1991; MacIntyre et al., 1997; Peirce et al., 1993), so no interrater reliability was reported and thus it is doubtful that teacher-assessment in those studies was accurate enough to be compared with self-assessment.

Even when multiple raters were used, some researchers often used raw scores for the analyses (e.g., Alfalay, 2004; Hewitt & Stephenson, 2012; Phillips, 1992); however, in order to assess speaking skills more accurately, it is important to utilize multi-faceted Rasch measurement, as it can provide ability estimates that are adjusted for rater bias. Although the past decade has seen more studies on L2 speaking testing utilize multi-faceted Rasch analysis (Bonk & Ockey, 2003; Moere, 2006; Sato, 2011), no researchers have utilized this form of analysis for the validation of L2 speaking self-assessment. Instead, they have used raw scores for self- and teacher-assessments of L2 speaking.

The second problem is that most studies of the self-assessment of L2 skills have been primarily focused on the validity of self-assessment questionnaires (e.g., Bachman & Palmer, 1989; Le Blanc & Painchaud, 1985; Ross, 1998). As a result, the potential influences of affective variables that might bias the accuracy of self-assessments have generally been ignored. The researchers that have investigated this issue have focused on only one or two affective variables. For example,

MacIntyre, Noels, and Clément (1997) examined the effects of anxiety only, while Liu and Jackson (2008) investigated anxiety and unwillingness to communicate. Including more affective variables might provide more information about the sources of bias demonstrated by some EFL students when they engage in self-assessment.

Moreover, researchers investigating the relationship between L2 affective variables and self-assessment often used raw scores and calculated only correlation coefficients between them (e.g., AlFalla, 2004; Chen, 2008; MacIntyre et al., 1997); however, calculating bias using multi-faceted Rasch analysis and testing structural equation models can provide a more rigorous test of the causal relationships among those variables and the bias that is an inherent aspect of self-assessments.

The final problem addressed in this study is that although there might be important differences in the influence of affective variables between high and low proficiency L2 learners, few researchers have investigated the degree to which L2 affective variables influence the self-assessment of L2 proficiency with learners at different proficiency levels. Baker and MacIntyre (2000) conducted the only study I am aware in which self-assessments by students with different proficiency levels were investigated. However, Baker and MacIntyre (2000) did not administer L2 speaking tests to distinguish the different proficiency levels; instead, they considered immersion language students as having high proficiency and nonimmersion language students as having low proficiency.

Purposes of the Study

The present study is motivated by three purposes. The first purpose is to examine how Japanese students rate their own L2 speaking performance. I have chosen to focus on speaking proficiency because among the four skills, speaking is most frequently self-assessed in formal as well as informal settings. As Underhill (1987) stated, when people talk with others, consciously or unconsciously, they are constantly assessing themselves in terms of how successfully they are communicating. In addition, compared with other language skills, the self-assessment of speaking proficiency has the strongest relationship with affective variables such as anxiety (MacIntyre et al., 1997) and unwillingness to communicate (Liu & Jackson, 2008). In this study, the participants' L2 oral performances are measured by conducting oral interviews instead of using indirect measures, such as final course grades or cloze tests, and their oral performances were assessed by raters using the same rating scale with self-assessment in order to minimize bias from sources other than affective factors. The participants' self-assessments of L2 speaking are conducted immediately after they finish oral interviews, when they assess their own performance from their memory using the assessment rubric. This approach can provide a more realistic view about the act of self-assessment than when assessing their own L2 speaking skills without speaking the L2 at all. Additionally, multiple raters are used and multi-faceted Rasch measurement is utilized to calculate person ability measures that are adjusted for

rater bias. Moreover, the analytic scale used in this study is made up of four subcomponents of speaking ability: grammar, vocabulary, fluency, and pronunciation, which can provide separate measures for specific features of the learners' performances and thereby provide more detailed information about their performances by describing specific strengths and weaknesses (Luoma, 2004).

The second purpose is to investigate the degree to which five affective variables influence the act of self-assessment of L2 speaking and which lead to the underestimation, overestimation, and accurate evaluation of L2 oral skills. In order to investigate the influences of affective variables on self-assessment bias, Rasch bias measures are calculated instead of using self-assessment raw scores. Causal relationships among seven affective variables and self-assessment were tested using structural equation modeling. The affective variables are Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, L2 Speaking Motivation, and L2 Speaking Self-Confidence. The L2 Speaking Motivation construct consists of three subconstructs: Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, and Desire to Learn to Speak English. Self-Esteem is an important factor that influences the act of self-assessment (Heine et al., 2001; Todd, 2002). L2 Speaking Anxiety is included because it can exert negative influences on the self-assessment of L2 skills (Chen, Horwitz, & Schallert, 1999; Gardner & MacIntyre, 1993). L2 Willingness to Communicate is reported to influence the self-assessment of the all four language skills (Liu & Jackson, 2008). L2 Speaking Motivation can also influence the self-ratings of L2 skills, including L2 oral

performance (AlFallay, 2004; Masgoret & Gardner, 2003). L2 Speaking Self-Confidence can influence the self-assessment of L2 proficiency (Clément, 1980).

The third purpose is to investigate the degree to which the influences of L2 affective variables on the self-assessment of L2 oral performance differ between high and low proficiency learners. In this study, the participants' L2 oral performances are measured by conducting individual oral interviews. Rasch person ability measures are calculated using Facets and these are used to distinguish high and low proficiency speaking groups.

Significance of the Study

Examining the self-assessment of L2 oral performance in conjunction with affective factors that might influence self-assessment is important for three reasons. First, the findings of this study can help explain how Japanese university students view their own L2 speaking performance. Unlike ESL learners who can test their own English speaking ability by using English in everyday situations, EFL learners encounter few opportunities to speak English in their daily life, so it can be difficult for them to judge their own speaking skills accurately. Thus, it is worth investigating how accurately EFL learners can self-assess their L2 speaking ability.

Second, identifying variables that influence the ways in which learners view their own L2 speaking performance is a significant issue worth investigating. Many researchers have demonstrated the important influence of affective variables on the development of L2 proficiency, but looking at learners' self-perceptions of their

own L2 speaking proficiency is also important because it directly influences their use of the L2 as well as their future achievement. According to Giles and Byrne (1982), self-confidence in an L2 leads to the predisposition to participate in L2 communication, which potentially allows learners to improve their L2 proficiency. Tsui (1996) also stated that the self-perception of being an unskilled communicator in the L2 can adversely affect students' willingness to participate in L2 communicative events, and ultimately, this can limit the development of communicative competence.

Third, understanding how the influences of affective variables on self-assessment differ between students at two proficiency levels is important. The findings can allow teachers to better understand how students differ in terms of their perceptions of their own L2 performance and how those perceptions differ because of their proficiency levels. This knowledge can allow instructors to select more effective pedagogy when teaching high and low proficiency learners.

The Audience for the Study

The first audience for the present study is researchers investigating the relationship between affective variables and L2 speaking proficiency as well as those conducting research investigating the validity of self-assessment. As little research on the relationship between the self-assessment of speaking ability and affective factors has been conducted, this study can provide new insights into those relationships.

The second audience is English teachers at Japanese universities who want to better understand how their students view their own English speaking skills. By recognizing which subcomponents of speaking many Japanese students self-assess too leniently or too severely, teachers can teach speaking classes more effectively by informing students about potential positive and negative biases. Moreover, teachers who are considering introducing the self-assessment of speaking skills into their classrooms will benefit from this study, as they can inform their students about how Japanese learners of English tend to view their L2 speaking skills. This information might help them to guide the students to self-evaluate their L2 speaking skills more accurately.

The third audience of this study is Japanese students. As many Japanese students rely too much on teacher evaluation, the results of this study can provide them with an opportunity to learn about the importance of self-assessment and the autonomy that is an integral part of it. Additionally, learning about how Japanese students tend to view their own L2 speaking abilities might help some of those students engage in self-assessment more accurately.

Delimitations

This study is confined to examining Japanese science majors studying at two universities in eastern Japan. The purposive sampling procedure decreases the generalizability of the findings. As the participants received a standard English education at Japanese secondary schools, the results can best be generalized to

other Japanese students majoring in science. However, the results of this study can be generalized only with caution to other Japanese students, especially those majoring in English because those students are usually more motivated to learn English and have more confidence in English than many non-English majors.

The participants are drawn from two universities: One is highly competitive and the students have a high-intermediate level of English speaking proficiency (*hensachi*¹ rating is 66), while the other is an average level university (*hensachi* rating is 45), so most of the students have a low-intermediate level of English speaking proficiency. Although this study involves participants from two universities with different academic levels, the results should not be generalized to students who are much lower or much higher in terms of their speaking proficiency. Finally, because of the potential effects of culture on self-assessment, the results should be generalized to learners from other cultures with caution.

The Organization of This Study

In Chapter 2, I look at the self-assessment literature and review the research on the affective variables that are investigated in this study; Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, L2 Speaking Self-Confidence, and L2 Speaking Motivation. At the end of this chapter, I present the gaps in the

¹*hensachi* is standard score for academic ability determined by trial examinations offered by major preparatory schools in Japan. It is calculated by $(\text{scores} - \text{average scores}) \div \text{standard deviation} \times 10 + 50$. Schools with a *hensachi* of 50 are at an average level of difficulty. Schools with a *hensachi* of 60 belong to the upper 15%, while those with a *hensachi* of 40 belong to lower 15%.

literature, the purposes of the study, and the research questions that guide this study. In Chapter 3, I describe the methodology that is used in this study in terms of the participants, instruments, procedures, and the analyses used to answer the research questions. Chapter 4, Preliminary Analyses, shows the results for instrument validation, in which the validity and reliability of the constructs measured by the questionnaire are examined. Following the Preliminary Analysis, the results for each research question are reported in Chapter 5, Results. Key findings and the implications of the results are discussed in Chapter 6, Discussion. The limitations of the study, and suggestions for future research are provided in Chapter 7, Conclusion.

CHAPTER 2

REVIEW OF THE LITERATURE

Self-Assessment

Beneficial Effects of Self-Assessment

The usefulness of self-assessment is widely recognized. For example, Todd (2002) described five reasons for using self-assessment: Self-assessment is a prerequisite for a person to become a self-directed learner, it can raise learners' awareness of the foreign language, it increases motivation and goal orientation in learning, some aspects of language learning, such as effort and learner beliefs, can only be assessed through self-assessment, and it can reduce the teacher's assessment burden. In addition to these benefits, Oscarson (1989) stated that self-assessment gives learners training in learning, which is itself beneficial to learning. Moreover, asking questions such as "How well have I done?" potentially stimulates learners to consider course content and assessment criteria more carefully, and this type of activity can foster an important evaluative attitude in learners and help them consider their own progress in acquiring the foreign language. Otherwise, many learners focus their attention entirely on the results of tests. Brown and Hudson (1998) listed an additional advantage of self-assessment by pointing out that learner involvement in the evaluation process is likely to bring about a broadened perspective of classroom assessment, which can help students to be more autonomous language learners. In addition, Kusnic and Finley (1993) stated "self-

evaluation helps students make meaning, derive relevance, and build coherence through their educational experience” (p.13). They concluded that active engagement is a potentially important antidote to a passive approach to learning. Jassen-van Dieten (1989) also noted that “reflection on one’s proficiency and insight into evaluation criteria will stimulate self-management, motivation and goal orientation” (p.31).

Self-Assessment Bias

Kruger and Dunning (1999) investigated the “Dunning-Kruger Effect” of self-assessment act on American university students, who compared their own abilities of recognizing jokes, logical reasoning, and grammar to those of other students. On average, they overestimated their own abilities and considered themselves above average. Those in the bottom quartile tended to overestimate their abilities because their incompetency prevented them from accurately evaluating their abilities, which is called the Dunning-Kruger Effect. On the other hand, top quartile students underestimated their abilities probably because they tended to believe that their proficiency was similar to that of their peers. However, their self-assessments became more accurate when competent participants learned that their peers actually performed poorly and when incompetent participants recognized their incompetency.

The Dunning-Kruger Effect can be seen in studies on self-assessment of L2 skills; low proficient learners overestimated their L2 skills (e.g., Jassen-van Dieten,

1998; Patri, 2002; Trofimovich, Issacs, Kennedy, Saito, and Crowther, forthcoming) and high proficiency learners underestimated their abilities (e.g., Matsuno, 2007).

Validity of Self-Assessment of L2 Skills

Many experimental studies have been conducted in which the validity of self-assessment has been determined by calculating and interpreting correlation coefficients between self-assessment and L2 proficiency as measured by objective tests. One of the early studies was conducted by Le Blanc and Painchaud (1985), who investigated whether their participants had the ability to evaluate their own second language proficiency and whether self-assessment was valid enough to be used to place students in a foreign language program. The participants were 200 first-year students at the University of Ottawa, whose second language was either English or French. They took English or French standard proficiency tests with listening and reading subsections, and answered a self-assessment questionnaire about their second language proficiency. For example, two self-assessment questions in the reading section were “I can understand short and simple written communications (posters, schedules, announcements)” and “I read specialized articles concerning my fields of study fluently.” The participants rated themselves using a five-point Likert scale ranging from 1 (*I cannot do this at all*) to 5 (*I can do this all the time*). The Pearson correlation between the test and questionnaire was .53, which indicated the presence of a large mismatch between the self-

assessment and objective measures. Next, the authors modified the self-assessment questionnaire to be more closely tied to the students' situations as potential second language users. For instance, the above-mentioned question was changed to "I would be able to read and understand the following in French when encountered on campus" and the participants choose from the following; 1. signs on doors, etc. indicating entrance, exit, danger, etc., 2. posters, announcements and advertisements, 3. written instructions for the use of various equipment, 4. information in a university calendar such as policies, course descriptions, etc., and 5. articles in the student newspaper. This change resulted in an improved correlation of .80. Therefore, the results suggested that the characteristics of the self-assessment questionnaire can strongly influence the strength of the correlations and that when self-assessment is closely related to situations students are familiar with, it can be used as a placement tool in a university's second language program. In fact, Le Blanc and Painchaud reported that the University of Ottawa used self-assessment instead of proficiency tests for placing students and that it worked well.

Bachman and Palmer (1989) investigated the construct validity of a self-rating test of communicative language ability (i.e., linguistic, pragmatic, and sociolinguistic competences) through the use of a multitrait-multimethod (MTMM) design and confirmatory factor analysis. They used three types of self-assessment questionnaires: ability to use trait, difficulty in using trait, and recognition of input. The participants were 116 non-native English speakers in the Salt Lake City area ranging in age from 17 to 67, who were from 36 countries. The results showed that

self-ratings could act as a reliable and valid measure of communicative language abilities. What was unique about this study was that although can-do questions were used frequently, questions eliciting the participants' perceived difficulties using the language appeared to be the most effective. This result suggested that L2 users might be more aware of the areas in which they encounter difficulty than those they find easier to perform.

Jassen-van Dieten (1989) is another early study in which the validity of self-assessment of the four skills was investigated. The participants were 730 adult learners of Dutch as a second language who came from 73 mainly nonwestern countries. She compared reading, listening, speaking, and writing tests of Dutch with a parallel version of each test in a self-assessment format. Students also took a C-test. The Pearson correlations between the self-assessment ratings and the criterion tests ranged from .33 to .69, while the correlations between the C-test and the criterion tests were .44 to .76. These results indicated that although self-assessment correlated significantly with the criterion tests, the C-test was a better predictor of the criterion tests than self-assessment. Next, because the self-assessment formats and the criterion formats consisted of the same items, the researcher calculated the proportions of correct and incorrect estimates. The answer combination of "Yes, I can" and correct answer outnumbered the combination of "No, I can't" and incorrect answer. The proportion of the overestimation of L2 proficiency was, on average, six times the proportion of underestimation.

Ross (1998) conducted a self-assessment validity study for a language-training program at a Japanese electronics company. The participants, 254 employees at beginning and elementary proficiency levels of English, completed a self-assessment questionnaire of their English skills. Eight teachers also provided assessments of each learner. The learners, then, took an achievement test that included two formats; one of them utilized the same format as that found in the course book the participants had studied, while the other used a different format. Only the listening section of the test was utilized in the analysis. A contrastive multiple correlation analysis was conducted and the results showed that the teacher evaluations of the participants' performances generally displayed a higher correlation with the test results compared to the correlation of the participants' self-assessment with the test results. The correlation of the self-assessment and the modified format test was $r = .39, p < .05$, while the correlation of the self-assessment and the exact match format test was $r = .50, p < .0001$. Consequently, these results revealed that the accuracy of self-assessment was greater with an achievement criterion than when self-assessment was based on proficiency criteria.

Although many researchers have reported medium to high correlations between self-assessment and a criterion test, some teachers and researchers have questioned whether self-assessment is an accurate measure of learners' abilities (Lewkowicz & Moon, 1985). Indeed, some researchers have reported a weak relationship between the two as shown in the following four studies. First, Peirce, Swain, and Hart (1993) investigated 500 Grade 8 early and middle immersion

students studying in French immersion programs in Canada. The difference between the two groups was that the total hours of instruction in French at the end of Grade 8 for the middle immersion students were less than half of that for the early immersion students. Both groups took French tests covering the four skills and answered a self-assessment questionnaire. The early immersion students outperformed the middle immersion students on the test and they assessed themselves more highly than the students in the middle immersion group. The Pearson correlations between self-assessments of language proficiency and tested proficiency were low ($r = .01$ to $.25$); therefore, self-assessed ratings were a very weak indicator of tested proficiency for both French immersion groups.

Second, Brantmeier (2006) did not find the significant relationship between self-assessment and reading test. She explored the effectiveness of self-assessment for 34 university learners of Spanish who were enrolled in an advanced Spanish class in the United States. At the beginning of the semester, they took the reading section of the Online Placement Exam (OPLE) in which their comprehension was measured with multiple-choice items. They also completed a self-assessment questionnaire concerning Spanish reading skills. In order to assess subsequent reading performance, in the third week of the course, the same students completed a self-assessment questionnaire and took three reading tests: the written recall of a reading passage, sentence completion items, and multiple-choice questions. The results of a regression analysis revealed no statistically significant relationships between the pre-self-assessment and OPLE results or between pre-self-assessment

and subsequent reading performance as measured by written recall, sentence completion, and multiple-choice section results. Likewise, no statistically significant association was found between post-self-assessment and the OPLE, or between post-self-assessment and the three tests. On the other hand, there was a statistically significant association between the OPLE and recall ($\beta = .20$), and the OPLE and the multiple-choice section ($\beta = .32$). However, no statistically significant association was found between the OPLE and the sentence completion section. There was no statistically significant relationship between the four self-assessment items and reading achievement (OPLE scores and all three subsequent comprehension tasks). This study showed that the OPLE was a better predictor of actual reading performance in the course than the self-assessment.

Third, Matsuno (2007) investigated 91 Japanese university students and four native Japanese teachers of English, and compared self-, peer-, and teacher-assessments of English essays using multi-faceted Rasch measurement. With self-assessment, many of the writers' ability estimates were below .00 logits, with peer-assessment, most writer ability estimates were above 1.00 logits, and with teacher-assessment, more writers were assessed as below 1.00 logits. Therefore, it was found that students tended to evaluate their own essays severely, but they evaluated their peers' essays leniently. Matsuno also reported that lower proficiency students did not display a common tendency toward overestimation or underestimation; however, those with higher ability tended to rate themselves severely.

Fourth, Trofimovich et al. (forthcoming) examined the Dunning-Kruger Effect by investigating the relationships between self- and other-assessments of accentedness and comprehensibility in L2 speech. The participants were 134 non-native English speakers living in Canada. They completed the picture story task and rated how well they performed in accentedness and comprehensibility from the memory. Their self-assessments were compared with three native-speaker judges. No significant correlation was found between self- and other-assessments for accent ($r = .06, p = .50$), and there was only a weak correlation for comprehensibility ($r = .18, p = .03$). However, there were significantly negative correlations between speakers' overconfidence scores and other-assessment for accent ($r = -.67, p < .0001$) and comprehensibility ($r = -.56, p < .0001$), indicating that most accented and least comprehensive speakers overestimated their abilities, while speakers at the top of each scale underestimated. Next, 60 speakers were drawn from the participants and a stepwise multiple analysis was conducted to examine the contribution of their phonology and lexico-grammar scores to their overconfidence scores. Their overestimations were associated with phonology scores (segmental accuracy, word stress, rhythm, intonation, and speech rate) for accent ($\beta = -.12, p = .001$) and for comprehensibility ($\beta = -.11, p = .001$), but not with lexico-grammar scores. By investigating their L1 background, it was found that the more accented speakers such as Chinese and Hindi/Uru and less comprehensible speakers such as Chinese tended to overestimate their abilities compared with the speakers in other L1 backgrounds.

Some researchers have conducted self-assessment training and examined the validity of self-assessment. Kruger and Dunning (1999) discovered that by learning other students' performances, high proficiency participants were able to evaluate their own skills more accurately because they recognized how poorly their peers performed, and low proficiency learners recognized their incompetency and evaluated themselves more accurately than others who did not receive the training. Therefore, Kruger and Dunning (1999) commented that self-assessment training was successful because the participants who received the training gained the metacognitive skills by engaging in the mediational analysis.

As for the L2 studies, Chen (2008) examined 28 Chinese university students in Taiwan, who took part in a self-assessment training program about L2 oral presentations. It involved the process of internalizing a set of standards for good oral performance through observation and discussions of peer performances and reflection on self-performance. Before the training, the correlation between self- and teacher-assessments of L2 oral presentations was moderate ($r = .55, p < .05$), but it improved after the training ($r = .79, p < .05$). Thus, Chen (2008) concluded that self-assessment training helps students assess their own L2 oral performance more accurately.

On the contrary, the training of self-assessment in Patri (2002) did not improve the accuracy of the self-assessment. She examined the importance of peer feedback on self- and peer-assessments. In this study, 56 Chinese students in a university in Hong Kong were given two-hour assessment training that introduced

the students to the important elements of a good presentation. The students were divided into a control group ($N = 30$) and an experimental group ($N = 26$), and only the students in the latter group received peer feedback after each presentation. Correlations among self-, peer-, and teacher-assessments of the oral presentations revealed that the correlation between peer-assessment and teacher-assessment was greater for the experimental group than for the control group (Experimental Group, $r = .85, p < .01$; Control Group, $r = .50, p < .01$), but the correlation between self-assessment and teacher-assessment for the experimental group was slightly lower than that of the control group (Experimental Group, $r = .46, p < .01$; Control Group, $r = .50, p < .01$). Therefore, peer feedback on L2 oral presentation rather than self-assessment training greatly helped the students to perform peer-assessment more accurately, but did not significantly affect self-assessment. Additionally, Patri found the tendency that high achievers underestimated their performances, while low achievers overestimated their performances.

Section Summary

Many researchers reviewed above have maintained the beneficial effects of self-assessment for L2 learning. For example, self-assessment helps learners to become autonomous learners who can focus on their own progress and have greater motivation to improve L2 skills (e.g., Brown & Hudson, 1998; Oscarson, 1989; Todd, 2002).

Many empirical studies have been conducted investigating the validity of the self-assessment. Some researchers found significantly positive relationships between self-assessment and L2 proficiency as measured by objective tests (e.g., Jassen-van Dieten, 1989; Le Blanc & Painchaud, 1985), while other researchers found weak correlations between them (e.g., Peirce et al., 1993; Trofimovich et al., forthcoming).

Kruger and Dunning (1999) maintained that the self-assessment bias is caused by the Dunning-Kruger Effect; low proficiency test-takers tend to overestimate their own abilities because their incompetency prevents them from evaluating themselves accurately, while high proficiency test-takers tend to underestimate themselves because they believe that their proficiency is shared by others.

The studies on self-assessment reviewed above did not consider the effect of affective factors on self-assessment. In the following sections, I review five affective variables that plausibly affect language learners' self-assessment of their oral performances.

Self-Esteem

Self-esteem can be defined in many ways. The Longman Dictionary of Language Teaching and Applied Linguistics (2002) defines self-esteem as “a person's judgment of their own worth or value, based on a feeling of ‘efficacy’, a sense of interacting effectively with one's own environment” (p. 475). Baumeister, Campbell, Krueger, and Vohs (2003) defined self-esteem as “how much value

people place on themselves. It is the evaluative component of self-knowledge. High self-esteem refers to a highly favorable global evaluation of the self. Low self-esteem, by definition, refers to “an unfavorable definition of the self” (p. 2). Brown (1998) stated that “self-esteem and self-evaluation are related—people with high self-esteem think they have many more positive qualities than do people with low self-esteem” (p.192). However, Baumeister et al. (2003) pointed out that self-esteem is a perception rather than reality, so a person’s belief about his or her qualities is not necessarily accurate.

Culture and Self-Esteem

According to Shokraii (1996), from the late 1970s to the early 1990s, many educators in the United States assumed that students’ self-esteem played an important role in their scholastic achievement and their later success in life. Therefore, programs designed to increase students’ self-esteem were created in order to increase students’ academic performance while decreasing interpersonal conflicts.

While self-esteem is considered important in the United States, Japanese society puts relatively little value on it. Although self-esteem is usually translated as *jisonjin*, 自尊心, in Japanese, R. A. Brown (2005) argued that self-esteem and *jisonjin* do not mean the same thing and that the Japanese language does not possess a translation equivalent of self-esteem. In Japanese, *jisonshin* connotes *pride* or *vanity*; thus, *jisonshin* suggests that the person lacks modesty and humility,

both of which are considered virtues in Japan. Therefore, many Japanese feel that it is childish to boldly assert that “I feel that I have many good qualities” or “I am able to do things as well as most other people” (p. 3). Yamaguchi (2006) also pointed out the difference in the meaning of self-esteem between individuals from Western cultures and Japanese by noting that “self-esteem represents Western values such as self-confidence, self-reliance, and self-assertiveness. In the East, on the other hand, those characters associated with self-esteem are not necessarily desirable, because they are often detrimental to interpersonal harmony” (p. 1).

According to Markus and Kitayama (1991), in European-American cultural contexts, high self-esteem is often seen as a characteristic that is a prerequisite for participating in independent and mutually approving relationships. On the other hand, in East Asian cultural contexts, self-criticism is indispensable for engaging in independent and mutually sympathetic relationships. Likewise, Heine, Markus, Lehman, and Kitayama (1999) maintained that unlike North Americans, who regard themselves positively, Japanese self-perception tends to be critical and self-effacing. Heine et al. argued that this is partly because most Japanese believe that self-criticism serves as the basis for future improvement and achievement.

This view is supported empirically by several studies. For example, according to Farh, Dobbins, and Cheng (1991), a positive view of the self brings about leniency effects in self-assessment, which is related to the notion that “individuals are motivated to view themselves as positively as possible” (p.130). Farh et al. investigated whether the received doctrine of leniency is supported outside of

Western culture with 982 leader-subordinate pairs who were drawn from eight organizations in Taiwan, including an oil company, a hospital, and an automobile manufacturer. The self-ratings of their job performance were compared with their supervisor's ratings. The results showed that Taiwanese workers evaluated themselves less favorably than their supervisors did. These findings were contrary to the typical finding in the United States that self-ratings of performance were higher than the supervisor's ratings. Therefore, the authors concluded that Taiwanese workers exhibited a modesty bias, and that culture played an important role in shaping workers' perceptions of their own work performance.

Heine, Kitayama, and Lehman (2001) also examined the difference between Japanese and Canadian students, and 77 Japanese students were compared with 60 Canadian students using a self-enhancement measure that asked the students to evaluate the percentage of students who were better than themselves in 11 abilities and traits such as attractiveness, cooperation, and work ethic. The results showed that the Canadian students all rated themselves above the 50th percentile, while most of the Japanese students rated themselves below the 50th percentile. This study revealed that the Canadian students had higher self-esteem and stronger self-other bias than the Japanese students.

Likewise, Brown (2006) compared 72 Japanese students and 110 American students by investigating their self-esteem and their self-evaluation of their traits and abilities. For both the American and Japanese participants, students with higher self-esteem produced higher self-evaluations. On the whole, the American students

had higher self-esteem than the Japanese students. Although some Japanese students rated themselves higher than average, most of them rated themselves as average, and some rated themselves as below average; a histogram indicated a reasonably normal distribution. On the other hand, no American students rated themselves below average, so the distribution was negatively skewed. Based on this result, Brown concluded that the American students rejected the possibility that they were below average, while the Japanese students produced self-competence assessments that placed them near the middle of the distribution.

Unlike the three studies reviewed above, some researchers have argued against the view that East Asians have low self-esteem that leads to self-criticism and a lack of self-enhancement. Kobayashi and Brown (2003) asked 59 American students and 54 Japanese students to rate themselves, their best friends, and most students at their university on eight attributes such as being competent, friendly, and responsible. Like their American counterparts, the Japanese participants believed that they and their best friends were better than most other students at their university. However, unlike the American students, who judged themselves higher than their friends, the Japanese students judged their best friends more positively than they judged themselves. The authors suggested that because best friends form part of the extended self, the Japanese students in the study displayed indirect forms of self-enhancement.

Another example is Kudo and Numazaki (2003) who investigated the social sensitivity of 39 Japanese female college students in a condition where their

anonymity was assured. The students were asked to take the Social Sensitivity Test, which consisted of two sets of questions. The first set asked the students to indicate which of four alternatives (e.g., *cloud*, *blue*, *bird*, and *the sun*) was most frequently evoked by the target word (e.g., *the sky*). The second set asked them what a protagonist in a troublesome situation would say. About 20 minutes later, success or failure feedback was randomly assigned to each student. The students then completed the questionnaire asking how their ability, effort, task difficulty, and luck affected the results of the Social Sensitivity Test. A MANOVA was conducted in which the independent variable was the type of feedback (success vs. failure), and the dependent variables were four ability measures (ability, effort, task difficulty, and luck). The main effect for feedback was significant, $F(4, 32) = 2.77$, $p < .05$, so univariate one-way ANOVAs for the effect of feedback on the four attribution measures were conducted. The main effect for feedback was significant for ability, $F(1, 35) = 5.34$, $p < .05$, indicating that the students who received feedback indicating that they were successful were more likely to consider the Social Sensitivity Test as reflecting their true ability than those who received feedback indicating failure ($M_{\text{success}} = 4.68$, $M_{\text{failure}} = 4.11$). A significant difference between the two feedback groups was also found for task difficulty, $F(1, 35) = 4.47$, $p < .05$. This finding indicated that the students in the failure feedback group rated the Social Sensitivity Test as more difficult than the success feedback group ($M_{\text{success}} = 2.95$, $M_{\text{failure}} = 3.94$). Unlike previous research indicating a self-critical bias rather than a self-serving bias in Japanese participants, Kudo and

Numazaki reported that the Japanese participants in their study showed a self-serving attributional bias when they were assured that their responses were anonymous and confidential.

Some researchers have argued that Japanese have self-enhancing motives but that many Japanese display such motives in indirect ways. Moreover, even when Japanese have shown high self-esteem indirectly, their self-esteem was not as high as Americans. Thus, although these researchers have argued against the common view, I do not believe that they were completely successful; it is reasonable to believe that many Japanese have lower self-esteem than individuals from western countries.

Self-Esteem and L2 Skills

With regard to foreign language learning, learners with high self-esteem are generally said to achieve higher L2 proficiency than learners with low self-esteem because those who possess high self-esteem have self-confidence in their abilities, which motivates them to participate in L2 communication, and this leads to improvements in their L2 skills (Ortega, 2007). Shirahata, Tomita, Muranoi, and Wakabayashi (1999) maintained that students with high self-esteem are better language learners, especially where oral skills are concerned. In contrast, low self-esteem has a detrimental effect on language learning because it often generates anxiety (Ávila, 2007). Therefore, students with low self-esteem are often afraid of making mistakes and refrain from using the target language (Ellis, 1994; Horwitz et

al., 1991; Shirahata et al., 1999). Andrés (1999) reported a case study of a six-year-old boy who showed low self-esteem. In English class, he avoided work and was reluctant to express opinions due to his lack of confidence in himself. However, by engaging in classroom activities in which he received compliments from others, his English proficiency improved significantly.

Liu (2012) investigated 934 Chinese university EFL learners and explored the relationship between students' performance in English (scores on a mid-term exam) and four individual difference variables (self-esteem, personal traits, language class risk-taking, and sociability). The results of a regression analysis revealed that among these variables self-esteem was the most powerful predictor of students' performance in English ($\beta = .14, p = .000$). Therefore, Liu (2012) maintained that enhancing students' self-esteem is important because it results in students taking more risks when using English in class and becoming more socially active.

Self-Esteem and the Self-Assessment of L2 Skills

Very few researchers investigated the relationship between self-esteem and self-assessment of L2 skills. For example, Anderson (1982) investigated 22 university students learning English as a second language. The participants completed a 15-item questionnaire that evaluated their own English abilities, and the questionnaire results were compared with teacher-evaluations that were based on the same questionnaire. The students from the Far East, including Japan, rated themselves the lowest, with a mean of 2.91, while the teachers rated them much

higher, with a mean of 4.76. The students from the Middle East rated themselves higher ($M = 4.48$) than the teachers ($M = 3.92$), while the students from South America rated themselves lower ($M = 4.51$) than the teachers ($M = 5.13$). The author also reported the correlations between the self-assessment and TOEFL scores and between the teacher-assessment and TOEFL scores. While the former showed a non-significant relationship, the latter displayed a significantly positive relationship, indicating that the teachers' ratings of the students' English abilities resembled their actual abilities as measured by TOEFL, whereas the self-assessment did not. Anderson (1982) concluded that students' self-assessment of L2 abilities were not reliable possibly because of self-esteem. The students from the Far East who are said to have lower self-esteem tended to rate their English abilities lower than the teachers, while the students from the Middle East tended to rate their English abilities higher than the teachers. However, Anderson did not measure the students' self-esteem, so it is not possible to know whether the unreliability in self-assessment was actually due to self-esteem.

AlFallay (2004) conducted the only study I am aware of investigating the relationship between self-esteem and the self-assessment of L2 ability. The participants were 78 Arabic university students in Saudi Arabia studying in an intensive EFL program. The researcher examined the relationship of self-and teacher-assessment with four psychological variables: self-esteem, motivation types, anxiety, and motivational intensity. Because of the focus of this study, I only report the findings for self-assessment. The questionnaire responses for the affective

variables were compared with the scores of L2 oral presentations rated by the students themselves (self-assessment) and the teachers (teacher-assessment). The presentations were assessed with respect to organization and content, use of grammar, manner, and interaction with the audience. The same criteria were used for self-assessment and teacher-assessment. The results showed that the students with low self-esteem were most accurate in self-assessing their L2 oral ability; their self-assessment and teacher-assessment had the largest correlation ($r = .85, p = .0001$). On the other hand, the correlation between self- and teacher-assessments for the high self-esteem group was lower ($r = .55, p = .013$) because these students overestimated their L2 oral ability. The author stated that this result was unexpected because learners possessing low self-esteem were more accurate in their self-assessment than those who had high self-esteem.

Section Summary

Self-esteem is an affective variable that can influence self-evaluation. As self-esteem is a cultural construct, people in different cultures have somewhat different degrees of self-esteem. For example, North Americans generally value high self-esteem a great deal, while Japanese consider it to be a lack of the traditional virtue of modesty. Most studies dealing with self-assessment and self-esteem have been conducted by psychologists interested in identifying cultural differences between individuals from western and eastern cultures. In these studies, the participants were often asked to evaluate their own traits such as friendliness, attractiveness,

and work ethic, as well as their abilities. Most of the results indicated that Asians showed lower self-esteem than individuals from western countries. Only AlFallay (2004) conducted a study examining the relationship between self-esteem and L2 self-assessment. He found that the participants with low self-esteem evaluated their L2 oral presentation most accurately.

Language Anxiety

Language anxiety is one of the most important variables influencing learners' L2 learning and performance. Horwitz, Horwitz, and Cope (1986) defined foreign language anxiety as "a distinct complex of self-perceptions, beliefs, feelings, and behaviors related to classroom language learning, arising from the uniqueness of the language learning process" (p. 31) and described three components of language anxiety: communication apprehension, test anxiety, and fear of negative evaluation. Communication apprehension is "an attitude characterized by fear of or anxiety about communicating with people," which includes oral communication anxiety, stage fright, and receiver anxiety (p. 30). Test anxiety refers to "a type of performance anxiety stemming from a fear of failure" (p. 30). Students can have test anxiety and oral communication anxiety simultaneously when they take L2 oral tests (Horwitz et al., 1986). Fear of negative evaluation, which is not limited to test-taking situations, has been defined as an "apprehension about others' evaluations, avoidance of evaluative situations, and the expectation that others would evaluate oneself negatively" (Watson & Friend, 1969, p. 449). Watson and Friend

differentiated fear of negative evaluation from test anxiety by pointing out that the former is not specific to testing situations because it can operate in socially evaluative situations. Taking account of the three characteristics of anxiety, the Foreign Language Classroom Anxiety (FLCAS) was developed by Horwitz et al. (1986) as a measure of anxiety specific to foreign language learning.

Although many researchers have pointed out the detrimental effects of anxiety (Horwitz et al., 1986; MacIntyre & Gardner, 1989), Scovel (1991) stated that there are two kinds of anxiety: facilitative anxiety and debilitating anxiety. He defined the difference of the two as follows:

Facilitative anxiety motivates the learner to “fight” the new learning task; it gears the learner emotionally for approach behavior. Debilitating anxiety, in contrast, motivates the learner to “flee” the new learning task; it stimulates the individual emotionally to adopt avoidance behavior (p. 22).

Therefore, it is important for researchers and teachers to consider which type of anxiety that language learners have because they have the opposite effects on L2 learning.

For example, Brown, Robson, and Rosenkjar (2001) reported on facilitating anxiety in their examination of the relationships among personality, motivation, anxiety, learning strategies, and language proficiency with 320 Japanese students at Temple University Japan in Tokyo. The students were divided into three proficiency groups: high ($N = 107$), middle ($N = 106$), and low ($N = 107$) based on the results of a cloze test. The results showed that English class anxiety and the

FLCAS correlated positively with Social Extraversion ($r = .32$, $r = .44$) and with Ascendance ($r = .37$, $r = .48$), which are the extraversion traits in the Yatabe/Gilford Personality Inventory, and they correlated negatively with Inferiority Feelings ($r = -.43$, $r = -.55$) and Depression ($r = -.30$, $r = -.43$), which are the neuroticism traits in the Personality Inventory. Thus, anxiety was associated with positive personality traits. The scores on the FLCAS for each group were 110.10 (high proficiency), 105.35 (middle proficiency), and 105.64 (low proficiency), and the scores on the English Class Anxiety Scale were 21.01 (high proficiency), 20.36 (middle proficiency), and 19.60 (low proficiency). These figures indicated that the high proficiency group was more anxious than the other groups. Therefore, anxiety in this study was viewed as “beneficial anxiety, or anxiety that pushes students to perform better” (p. 392).

Language Anxiety and L2 Proficiency

Language anxiety has often been investigated in the context of its relationship to L2 proficiency. Most studies have indicated that language anxiety has a negative correlation with L2 proficiency. For example, in Horwitz (1991), the correlation between FLCAS and final course grades was $r = -.49$, $p = .003$ for 35 students in Spanish classes, and $r = -.54$, $p = .001$ for 32 students in French classes.

Aida (1994) examined the relationship between anxiety and Japanese language learning with 96 students (64 native speakers of English and 32 non-native speakers of English) in beginning Japanese classes at the University of

Texas. The students' FLCAS scores were compared with their final course grades. The correlation coefficient between anxiety and course grades was $r = -.38, p < .01$, which indicated that higher anxiety was moderately associated with lower course grades. In addition, the author reported that a high-anxiety group received significantly lower grades than a low-anxiety group. Thus, the author concluded that anxiety affected the students' language learning negatively.

Saito and Samimy (1996) explored the role of anxiety and students' performance at three instructional levels with 257 students (134 beginning, 79 intermediate, and 44 advanced-level learners of Japanese) at the University of Texas. Students' anxiety and other affective variables, including risk-taking, sociability, motivation, attitude, and concern for grade, were compared with final grades. A stepwise regression was conducted for each instructional level (beginning, intermediate, and advanced) to determine which affective variables predicted the students' final grades. The results revealed that Year in College predicted the final grades of the beginning students ($R^2 = .047, p < .02$), while Language Class Anxiety predicted the final grades of both the intermediate and advanced level students (intermediate: $R^2 = .17, p < .001$; advanced: $R^2 = .22, p < .004$). The results supported earlier anxiety studies in which it was found that foreign language anxiety can exert a negative influence on learners' academic performances.

The three studies described above dealt with the relationship between anxiety and course grades. The following four studies collected speaking data from the participants and examined the relationship between Anxiety and L2 oral

performance and found negative relationships between them. First, Phillips (1992) investigated the effects of anxiety on oral examination grades (free-talk and role-play) and on oral performance variables (percent of total words in communication units (CU) and average length of CU) with 44 students enrolled in an intermediate French course in an American university. The results showed that the students' FLCAS scores correlated negatively with oral performance ($r = -.40, p < .01$), indicating that students with higher language anxiety tended to have lower oral performance scores than those with lower anxiety. Moreover, a statistically significant inverse relationship was found between language anxiety and oral performance variables ($r = -.34$ to $-.39, p < .01$ to $.02$). Thus, compared with less anxious students, students who had higher language anxiety tended to say less, produce shorter utterances, and use fewer dependent clauses and target structures.

Second, Oya, Manalo, and Greenwood (2004) investigated 73 Japanese students studying English in New Zealand. The participants' English speaking data were collected from a story-retelling task. It was found that accuracy, as measured by accurately used clauses, had significant negative correlations with the participants' degree of anxiety, which was measured with the Spielberger State and Trait Anxiety Inventory ($r = -.23, p < .05$). This finding implies that the more anxious the participants, the less accurate their clause constructions.

Third, Woodrow (2006) investigated the relationship between in-class and out-of-class anxiety and L2 oral performance. The participants were 275 students in English for Academic Purposes (EAP) courses at intensive language centers in

Australia. Data from the Second Language Speaking Anxiety Scale (SLSAS) were compared with the students' performance on an IELTS type oral assessment, which consisted of an introduction and general interview, individual long turn, and two-way discussion. The results showed that anxiety had a negative relationship with L2 oral communication. Negative correlations were found between in-class anxiety and oral performance, $r = -.23, p < .01$, and out-of-class anxiety and oral performance, $r = -.24, p < .01$. However, Woodrow commented that as the negative correlation was not very strong, anxiety should be viewed as just one of many variables having relationship with L2 oral communication.

Fourth, Hewitt and Stephenson (2012) conducted a replication study of the Phillips study (1992) with 40 Spanish university students taking an English course. They reported that the Pearson correlation between FLCAS scores and oral exam scores was $r = .49, p < .001$, suggesting that students with higher levels of language anxiety spoke Spanish more poorly. Hewitt and Stephenson (2012) also found that the FLCAS scores correlated negatively with the number of total words produced ($r = -.38, p < .015$), which means that more anxious students spoke less.

While these four studies found that anxiety was negatively correlated with L2 oral performance, Young (1986) found somewhat different results in an investigation of the relationship between anxiety and foreign language oral performance. Sixty university students or prospective teachers of French, German, or Spanish completed the Self-Appraisal of Speaking Proficiency (SASP) questionnaire and one anxiety measure, the State Anxiety Inventory (SAI), and

took the Oral Proficiency Interview (OPI). After taking the OPI, the students completed three other anxiety measures, the Cognitive Interference Questionnaire (CIQ), the Self-Report of Anxiety (SRA), and the Foreign Language Anxiety Scale of Reactions (FLASR). The participants' mean score on the SASP was 2.4 and their mean on the OPI was 2.1. A Pearson correlation indicated that the OPI and SASP were significantly correlated ($r = .60, p < .001$). These results suggested that the participants were able to self-assess their own oral performance fairly accurately, although they tended to evaluate their oral performance slightly higher than the OPI raters. Significant correlations between the OPI and four anxiety measures were also reported: SAI, $r = -.32, p = .01$; CIQ, $r = -.15, p = .13$; SRA, $r = -.32, p = .01$; FLASR, $r = -.38, p = .01$. These findings suggested that lower levels of anxiety are associated with higher oral proficiency. However, when the effect of the four types of language anxiety was accounted for, significant correlations were no longer found between the OPI and the four anxiety measures. Therefore, the author concluded that language proficiency exerts a greater influence on the OPI score than anxiety.

In the studies reviewed above, anxiety was found to have a significant relationship with L2 oral performance. The effect of anxiety on oral performance has also been investigated qualitatively. First, Price (1991) investigated foreign language anxiety through interviews with 10 students (8 female and 2 male students) with high levels of foreign language anxiety studying at the University of Texas. Each participant was interviewed for one hour, and the resulting data were

transcribed and analyzed. First, three sources of anxiety were mentioned by the students: The participants were afraid of making pronunciation errors, and they felt frustrated when they could not communicate effectively in the target language, while the greatest source of anxiety was speaking the foreign language in front of peers. Second, the causes of anxiety were examined. Many of the students felt that foreign language courses were more demanding and more difficult than other courses and that they were not as good as other students because they had no aptitude for foreign language learning. The researcher also identified two personality variables that caused anxiety: perfectionism and fear of public speaking. Based on the interviews, the researcher suggested that instructors should attempt to alleviate students' anxiety by letting them know that making mistakes while speaking is natural and by providing positive reinforcement.

Second, Gregersen and Horwitz (2002) conducted an interview study investigating the relationship between foreign language anxiety and perfectionism with four anxious and four non-anxious EFL students at the Universidad de Atacama in Chile. The students had a one-on-one English oral interview for about five minutes, and they were then interviewed about their oral performance. The researchers found that anxious learners showed perfectionist characteristics: They had a higher level of concern for their errors and worried a great deal about negative evaluations by others. On the other hand, non-anxious learners did not place such high demands on themselves where accuracy was concerned; rather, they set realistic personal standards and felt satisfied when they met them.

Language Anxiety and the Self-Assessment of L2 Speaking

Other researchers have investigated not only the relationship between anxiety and oral proficiency, but also the relationship between anxiety and the examinees' perception of oral proficiency. The number of the studies on self-assessment is still limited, but the following four studies were focused on the relationship between anxiety and self-assessment of L2 speaking.

First, Gardner and MacIntyre (1993) investigated six affective variables, Integrativeness, Attitude Toward the Learning Situation, Motivation, Language Anxiety, Integrative Motivation, and Attitude/Motivation, with 92 students taking introductory French courses at a Canadian university. Language anxiety displayed the strongest correlation with all objective measures of achievement ($r = -.23$ to $-.55$, $p < .01$ to $.05$) except for grades (the Likert and Guilford assessments: $r = -.14$, $r = -.22$, respectively) and word production (Guilford assessments: $r = -.18$). In addition, language anxiety correlated more highly with self-rated proficiency than objective measures ($r = -.45$ to $-.65$, $p < .01$), and self-ratings of speaking proficiency had the highest correlation among the four skills. Therefore, this study indicated that anxiety correlated with self-perceptions of L2 ability more highly than actual achievement.

Second, Clément, Dörnyei, and Noels (1994) examined the relationship between social psychological factors and L2 achievement in the unicultural Hungarian setting with 301 secondary school students in Budapest. The participants

answered a questionnaire designed to measure orientations, attitudes, and anxiety, as well as the evaluation of their teacher, the course, and their own English language proficiency. Data concerning the students' L2 achievement measures were obtained from their teachers. First, when these variables were factor analyzed, five factors were extracted. One factor, Self-Confidence with English, included items on Anxiety in Class (loading = .81) and English Use Anxiety (loading = .77), both of which were reverse coded, as well as Self-Evaluation of English Proficiency (loading = .66). Therefore, the researchers concluded that students who had little speaking anxiety tended to evaluate their English proficiency positively. Another finding was that students' self-evaluation of L2 proficiency correlated significantly with the teachers' ratings of the students' L2 proficiency, including communication skills, passive skills, and most recent grades ($r = .16$ to $.49$, $p < .001$ to $.05$); communicative skills showed the strongest relationship with students' self-evaluation ($r = .49$, $p < .001$). Additionally, the self-confidence index, which was arrived at by aggregating two anxiety indices, and self-evaluation of L2 proficiency, was more strongly correlated with the teachers' ratings of students' L2 proficiency ($r = .18$ to $.52$, $p < .001$ to $.05$); again, communicative skills had the strongest relationship with self-confidence ($r = .52$, $p < .001$). These results indicated that the students were able to self-evaluate their L2 abilities, and especially their own communicative skills more accurately than other skills.

Third, MacIntyre, Noels, and Clément (1997) examined the accuracy of perceived L2 competence by considering how language anxiety creates bias in a

self-assessment task. The participants were 37 Anglophone students studying French at a bilingual university. They completed a language anxiety questionnaire and a can-do test in which they rated their L2 speaking, reading, writing, and listening proficiency. The participants then engaged in each of those skills, and bilingual judges rated their performances. The speaking tasks were rated for both the number of ideas expressed and for the quality of the French. The writing tasks were rated by counting the number of ideas expressed and rating output quality. The reading tasks were rated for the number of times the students expressed the correct translation of ideas. As for the listening tasks, the students responded in English after listening to the French passage and the judge rated them only for the number of ideas correctly identified. The results showed that the ratings of actual competence were significantly and positively correlated with perceived competence ($r = .51$ to $.72, p < .001$). However, the shared variance (r^2) between the two sets of ratings was below 50%. This result can be explained by the finding that language anxiety correlated negatively with perceived competence ($r = -.52$ to $-.60, p < .001$). Therefore, it was suggested that more anxious students tended to underestimate their ability, while less anxious students tended to overestimate their ability.

Fourth, Chen, Horwitz, and Schallert (1999) investigated the associations of two anxiety constructs, the FLCAS and the second language version of the Daly-Miller Writing Apprehension Test (SLWAT), with second language speaking and writing achievement. The participants, 433 Taiwanese university students, completed the FLCAS, the SLWAT, a background questionnaire, and a self-

perceived proficiency measure in English speaking and writing. The participants' final course grades for their speaking and writing classes were used as achievement measures. The results indicated that the FLCAS' relationship with speaking and writing was $r = -.28, p < .001$ and $r = -.25, p < .001$, respectively, and the SLWAT's relationship with speaking and writing was $r = -.14, p < .01$ and $r = -.27, p < .001$, respectively. Therefore, the authors concluded that the FLCAS is primarily a measure of speaking anxiety, while the SLWAT primarily measures writing anxiety. What is intriguing about this study is that the students' self-rated proficiency levels had higher correlations with the FLCAS and SLWAT than with their actual achievement. The FLCAS' relationship with self-rated speaking and self-rated writing was $r = -.53$ and $r = -.31$, respectively, while the SLWAT's correlation with self-rated speaking and self-rated writing was $r = -.26$ and $r = -.55$, respectively. This study indicated that anxiety has more influence on the self-perception of L2 proficiency than actual L2 achievement.

Section Summary

Language Anxiety is considered one of the important variables influencing L2 learning. Since the FLCAS was developed by Horwitz et al. (1986), numerous researchers investigated L2 learners' language anxiety using the FLCAS. Many researchers calculated correlations of the FLCAS scores with L2 proficiency or other affective variables such as L2 motivation, and found negative relationships between them. Therefore, it is often considered that language anxiety has a

detrimental influence on the development of L2 proficiency. On the other hand, fewer studies such as Brown et al. (2001) found facilitative anxiety, which had a positive relationship with L2 learners' proficiency.

Fewer studies have been focused on the relationship between anxiety and the self-assessment of L2 skills. Mostly, language anxiety was negatively correlated with self-assessment, indicating that L2 learners with greater language anxiety tended to evaluate their own L2 skill lower.

Willingness to Communicate

The concept of willingness to communicate (WTC), which was first developed in L1 communication (McCroskey & Richmond, 1987), evolved from "unwillingness to communicate" (Burgoon, 1976), "predisposition toward verbal behavior" (Mortensen, Arnston, & Lusting, 1997), and "shyness" (McCroskey & Richmond, 1982). L1 WTC was defined as a "personality orientation which explains why one person talks and another does not under identical, or virtually identical, situational constraints" (McCroskey & Richmond, 1987, p. 130).

MacIntyre (1994) developed an L1 WTC model (Figure 1) in which WTC is most directly influenced by communication apprehension ($\beta = -.15$) and perceived competence ($\beta = .58$). MacIntyre explained his model as follows: "It would appear that people are willing to communicate to the extent that they are not apprehensive about it and perceive themselves to be capable (competent) of effective

communication. The person least willing to speak up would be the apprehensive individual who feels incompetent as a communicator” (pp.137-138).

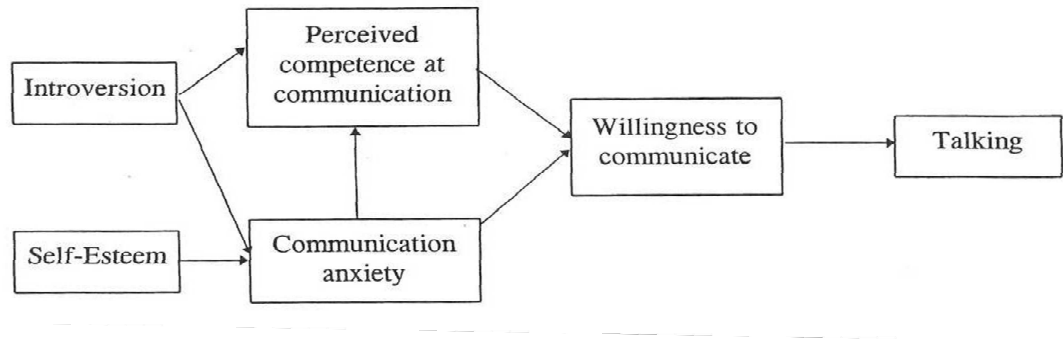


Figure 1. A portion of MacIntyre’s (1994) L1 willingness to communicate model.

Later, L1 WTC was applied to L2 communication by MacIntyre and Charos (1996); they presented the first path model of L2 WTC (Figure 2), which was developed by combining MacIntyre’s (1994) L1 WTC model and Gardner’s (1985) socio-educational model. In order to develop the L2 WTC model, MacIntyre and Charos (1996) investigated 92 Anglophone students taking introductory conversational French adult evening classes offered by local school boards in Ottawa. The participants completed a questionnaire that included self-report measures of the Big-Five personality traits, frequency of communication, willingness to communicate, perceived competence, attitudes, motivation, and the amount of French used in the work and home context. The results revealed that students who have greater motivation use the L2 more frequently ($\beta = .24$), and students who are more willing to communicate tend to communicate willingly

($\beta = .16$). A context variable, which represented the opportunity to converse with Francophones, also influenced the use of the L2 ($\beta = .13$). Because perceived competence had the largest effect on L2 use ($\beta = .60$), the authors suggested that “simply perceiving that one has the ability to communicate, regardless of one’s actual proficiency, can affect the rate of participation in L2 conversation” (p. 18).

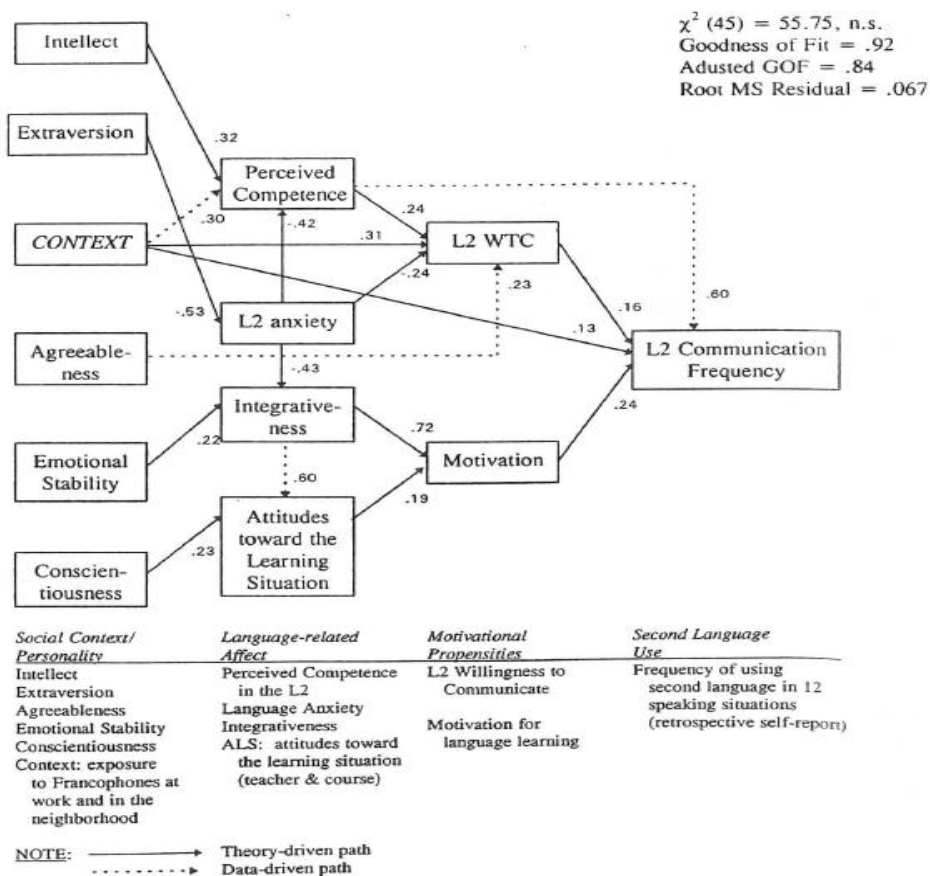


Figure 2. MacIntyre and Charos' (1996) L2 WTC model.

MacIntyre and Charos' (1996) L2 WTC model was further expanded into a heuristic model of L2 WTC by MacIntyre, Clément, Dörnyei, and Noels (1998) (Figure 3). The model consists of twelve variables in six layers. L2 WTC is located

in the Layer II, which is immediately under L2 Use in the Layer I, a top of the pyramid, implying that L2 WTC has more direct impact on L2 Use than other variables such as Self-Confidence and Motivation.

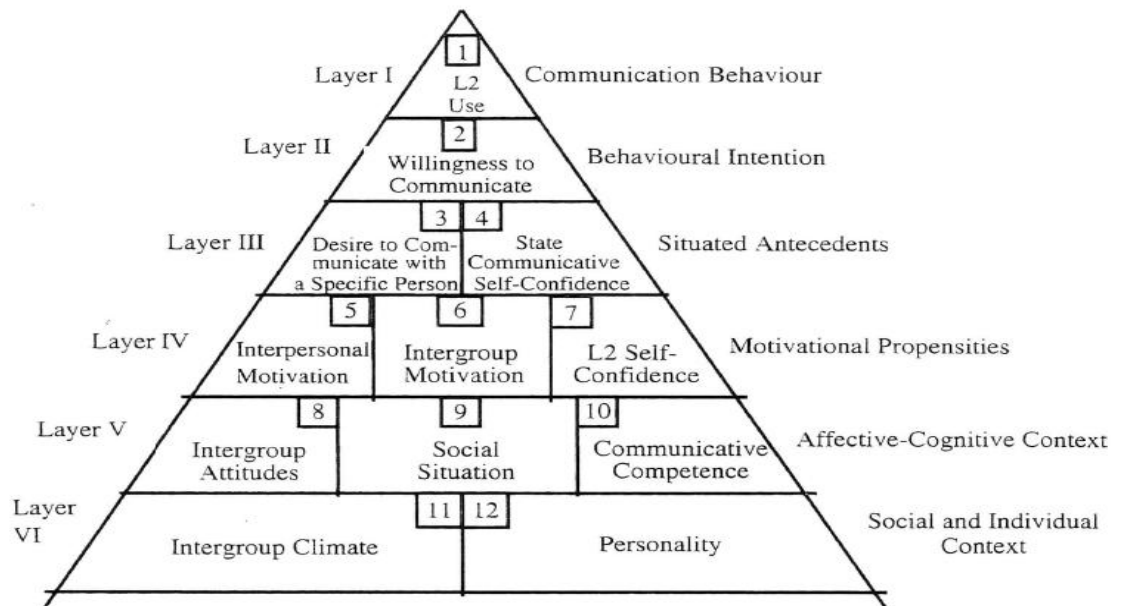


Figure 3. MacIntyre et al.'s (1998) L2 WTC pyramid model.

Both L1 and L2 researchers have empirically confirmed the validity of the concept of willingness to communicate. With regard to L1 WTC, McCroskey (1987) reported that the proportion of the students with high WTC who participated in classroom interaction was greater than the proportion of students who scored low on the WTC scale for each of the three observation sessions ($z = 3.12, p < .002$; $z = 2.56, p < .005$; and $z = 3.19, p < .001$). Zakahi and McCroskey (1989) reported that 92% of the students who scored high on the WTC scale were willing to participate in a laboratory study. However, only 24% of the students with low WTC

scores were willing. MacIntyre, Babin, and Clément (1999) replicated this study and found the same results; students who volunteered for the laboratory study had significantly higher WTC than those who did not, $t(209) = 1.79, p < .05$. Therefore, MacIntyre et al. (1999) concluded that students with high WTC were more willing to engage in oral and written communication tasks than those who had low WTC.

With regard to L2 WTC for Japanese students, Yashima, Zenuk-Nishide, and Shimizu (2004) examined the degree to which WTC predicts voluntary communication behavior in the L2 with 60 Japanese high school students who participated in a study-abroad program in the United States. The students completed the WTC questionnaire both before their departure and after spending three weeks in the United States. Pearson correlations indicated that WTC was significantly correlated with the frequency of communication with a host family ($r = .27, p < .05$), the frequency of communication in class ($r = .28, p < .05$), and the amount of time the students spent talking with the host family ($r = .37, p < .05$). In sum, the students with higher WTC tended to communicate with Americans more frequently and for a longer time than those who assessed themselves as having lower WTC.

In the Japanese EFL context, Yoshikawa (2005) investigated 35 Japanese university students majoring in English Education and explored the relationships between English speaking proficiency and affective variables. A path analysis indicated that Willingness to Communicate contributed strongly to basic interpersonal communication ability in English ($\beta = .40$), but it did not have a

significant path to academic language communication ability. Thus, Japanese students who were more willing to communicate in English tended to have basic English speaking skills, but they were not necessarily good at speaking academic English.

Willingness to Communicate and Perceived Competence

As reported above, L1 and L2 WTC models include perceived competence as an important predictor. Indeed, McCroskey and Richmond (1991) emphasized the strong relationship between WTC and perceived competence by commenting as follows:

Since the choice of whether to communicate is a cognitive one, it is likely to be more influenced by one's perception of competence (of which one usually is aware) than one's actual competence (of which one might be totally unaware). (p. 27)

The following five studies are the empirical studies that have been conducted examining how perceived competence predicts WTC. First, MacIntyre, Babin, and Clément (1999) examined L1 WTC, in which measures of WTC, perceived competence, communication apprehension, extraversion, emotional stability, and self-esteem were gathered from 226 university students. The results showed that perceived competence was the best predictor of WTC ($\beta = .84$), while apprehension ($\beta = -.33$) and extraversion ($\beta = .35$) showed statistically significant relationships with perceived competence.

Second, Baker and MacIntyre (2000) examined how immersion and nonimmersion foreign language learners differ in their willingness to communicate, perceived competence, communication apprehension, frequency of communication, and attitude/motivation in both English (L1) and French (L2). The participants were 195 students (71 immersion students and 124 nonimmersion students) from Grades 10, 11, and 12 at schools located in a predominantly Anglophone community in Canada. The results showed that while there was no significant difference between the communication variables in English (L1), there were significant differences between the two groups in French (L2). Compared with nonimmersion students, the immersion students had lower communication apprehension, higher willingness to communicate, greater perceived competence, and more frequent communication in French. Moreover, the correlation between L2 WTC and L2 perceived competence was not significant for the immersion group ($r = .17, p > .05$), whereas the correlation was strong for the nonimmersion group ($r = .72, p < .01$). Contrary to previous studies, the high proficiency students did not show a statistically significant relationship between WTC and perceived competence. What is more, in the immersion group, L2 perceived competence was significantly correlated only with the attitude/motivation index ($r = .28, p < .05$). On the other hand, the nonimmersion students' L2 perceived competence correlated strongly with WTC ($r = .72, p < .01$), frequency of communication ($r = .61, p < .01$), and the attitude/motivation index ($r = .40, p < .01$). Therefore, this study indicated that the correlates of L2 perceived competence differ between high and

low proficiency students. The perceptions of low proficiency students appear to be strongly influenced by affective and communication variables.

Third, Burroughs, Marie, and McCroskey (2003) investigated both L1 and L2 WTC in terms of the relationships among self-perceived communication competence (SPCC), willingness to communicate (WTC), and communication apprehension (CA). The participants were 131 undergraduate students studying at the Community College of Micronesia. They spoke various Micronesian languages as their L1, while English was their L2. The results showed that the Micronesian students felt less willingness to communicate and less communicatively competent in their L2 than in their L1. One interesting finding was that the correlation between SPCC and WTC in an L2 ($r = .80, p < .05$) was higher than that for the L1 ($r = .59, p < .05$). Therefore, this study demonstrated that perceived competence has a strong relationship with WTC and that the relationship is much stronger in an L2 than in the L1.

Fourth, MacIntyre and Doucette (2010) investigated 238 high school students, 97% of whom were English native speakers studying French. They examined the relationship among action control variables and language anxiety, WTC, and perceived communication competence using a path analysis. The results indicated that language anxiety had a negative path to perceived competence ($\beta = -.54$), while perceived competence had a positive path to WTC, both inside and outside the classroom (inside classroom, $\beta = .74$; outside classroom, $\beta = .16$). Thus, students

who perceived their L2 speaking skills to be at a high level tended to have higher willingness to communicate in the L2, especially inside the classroom.

Fifth, Yashima (2002) examined the relationships among L2 learning and L2 communication variables in a Japanese context using the WTC model and socio-educational model as a framework. The participants were 297 Japanese university students. The variables used in the structural equation model were WTC, International Posture, Motivation, L2 Proficiency, and L2 Communication Confidence, which was defined as a combination of low language anxiety and perceived communicative competence. Because my interest is in self-assessment, I focus on the L2 Communication Confidence variable, which predicted L2 WTC most strongly ($\beta = .68$). Moreover, L2 Learning Motivation was the strongest predictor of L2 Communicative Confidence ($\beta = .41$). Yashima (2002) reported that in the Japanese context, like in previous studies, there was a strong relationship between WTC and perceived competence/anxiety. However, there was no statistically significant relationship between L2 Communication Competence and L2 Proficiency ($\beta = .14$). This finding suggested that the students' self-perceived competence was inaccurate and that L2 WTC was directly influenced by L2 perceived competence, not L2 proficiency.

Unlike other studies, Liu and Jackson (2008) dealt with self-assessment of L2 skills rather than the communication construct and investigated the relationships among students' unwillingness to communicate, foreign language anxiety, and self-rated English proficiency in the four skills of listening, speaking, reading and

writing. The participants were 547 Chinese students at Tsinghua University in China. The results showed that the students' unwillingness to communicate was significantly and positively correlated with anxiety ($r = .50, p < .01$). Moreover, unwillingness to communicate had significant negative correlations with self-rated estimates of proficiency (listening, $r = -.26, p < .01$; speaking, $r = -.29, p < .01$; reading, $r = -.20, p < .01$; writing, $r = -.21, p < .01$); the higher the scores for unwillingness to communicate, the lower the students rated their own English proficiency. Among the four skills, speaking ability showed the strongest negative correlations with unwillingness to communicate. Additionally, like many studies reviewed in the Language Anxiety section above, foreign language anxiety correlated significantly with the students' self-rated English proficiency (listening, $r = -.29, p < .01$; speaking, $r = -.36, p < .01$; reading, $r = -.25, p < .01$; writing, $r = -.26, p < .01$). Speaking displayed the strongest negative correlation with anxiety.

Section Summary

The concept of WTC was first created for L1 communication and later applied to L2 by MacIntyre and Charos (1996). L2 learners with greater WTC are more likely to participate in L2 communication, which was confirmed by many empirical studies such as Yashima et al. (2004).

Because L2 WTC model includes perceived competence as a predictor of L2 WTC (MacIntyre & Charos, 1996), many researchers explored the degree to which perceived competence predicts WTC using structural equation modeling. In

addition, many researchers have reported that higher L2 perceived competence leads to higher WTC. However, as L2 perceived competence is not usually compared with actual L2 competence, so the accuracy or bias of the self-assessment of one's L2 skills is rarely considered. Liu and Jackson (2008) compared unwillingness to communicate with L2 self-assessment, but the actual proficiency was not included. Therefore, so far no studies have been conducted investigating the relationships between L2WTC and the self-assessment bias of L2 skills.

Motivation

Gardner and MacIntyre (1993) described the motivated individual as “one who wants to achieve a particular goal, devotes considerable effort to achieve this goal, and experiences satisfaction in the activities associated with achieving this goal” (p. 3). Dörnyei (2005) argued that motivation is so important in second language acquisition that a learner who possesses high aptitude but has insufficient motivation cannot acquire an L2 successfully. On the other hand, “high motivation can make up for considerable deficiencies both in one's language aptitude and learning conditions” (Dörnyei, 2005, p. 65).

Early second language acquisition researchers proposed that language aptitude plays a major role in language acquisition. However, Gardner and Lambert (1959) focused researchers' attention on the role of motivation when they conducted the first study measuring variables that were part of the socio-

educational model of second language acquisition. A factor analysis identified three factors: language aptitude, orientation, and motivation. Motivation in this study was defined as “a willingness to be like valued members of the language community” (Gardner & Lambert, 1959, p. 21), which later became integrative motivation.

In a subsequent study, Gardner and Lambert (1972) developed items to assess integrative and instrumental orientations. At this time, Gardner and Lambert defined integrative orientation as a positive disposition toward the L2 group and the desire to interact with and even integrate into that community, while instrumental orientation pertains to the pragmatic benefits of increased L2 proficiency, such as getting a good job.

Gardner (1985) argued that the key aspect of his socio-educational model (Figure 4) is integrative motivation, which consists of three main components: integrativeness, attitude toward the learning situation, and motivation. Integrativeness represents a genuine interest in learning an L2 in order to come closer to that community. Attitude toward the learning situation involves attitudes toward any aspect of the situation where the language is learned. Motivation comprises three elements, effort, desire, and positive affect, and it is believed to distinguish more motivated learners from less motivated learners (Gardner, 2001, pp. 5-6). Therefore, according to Gardner, the integratively motivated learner is “one who is motivated to learn the second language, has a desire or willingness to identify with the other language community, and tends to evaluate the learning situation positively” (Gardner, 2001, p. 6).

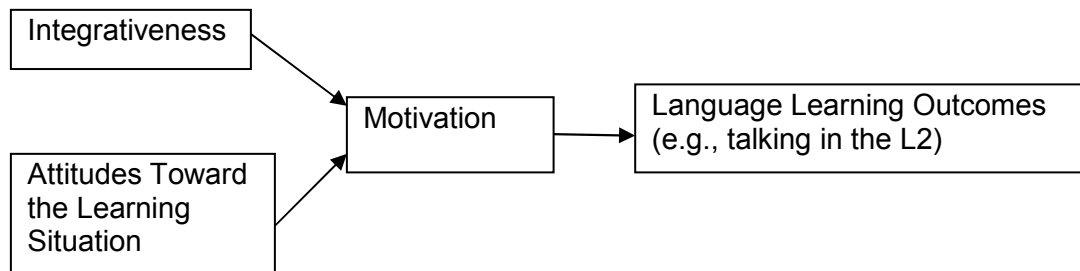


Figure 4. A portion of Gardner's (1985) socio-educational model.

Based on the socio-educational model, Gardner (1985) developed a multicomponential motivation questionnaire called the Attitude/Motivation Test Battery (AMTB). In addition to Integrative Motivation, the AMTB is designed to measure other affective variables such as Instrumental Orientation and Parental Encouragement. The AMTB has been considered the best measure of motivation and has been used by researchers in a number of countries (Dörnyei, 2005).

Tremblay and Gardner (1995) extended their social-educational model in response to calls for the “adaptation for a wider version of motivation” (p. 505). The revised model incorporated three new elements: goal salience, valence, and self-efficacy, and included a “language attitude → motivational behavior → achievement” sequence (Figure 5).

Although Gardner's socio-educational model established a basis for motivational research, some researchers came to doubt the applicability of the model to foreign language contexts. Gardner emphasized the importance of integrative motivation in L2 acquisition (Gardner, 1985), whereas other researchers argued that instrumental motivation is also important, especially in EFL contexts

(Clément, Dörnyei, & Noels, 1994; Dörnyei, 1990). For example, Dörnyei reported that in the Hungarian EFL context, instrumental orientation plays an important role for learners to reach the intermediate level of proficiency. Dörnyei also questioned the concept of integrative orientation in EFL situations because “affective predispositions toward the target language community are unlikely to explain a great proportion of the variance in language attainment” (p. 49).

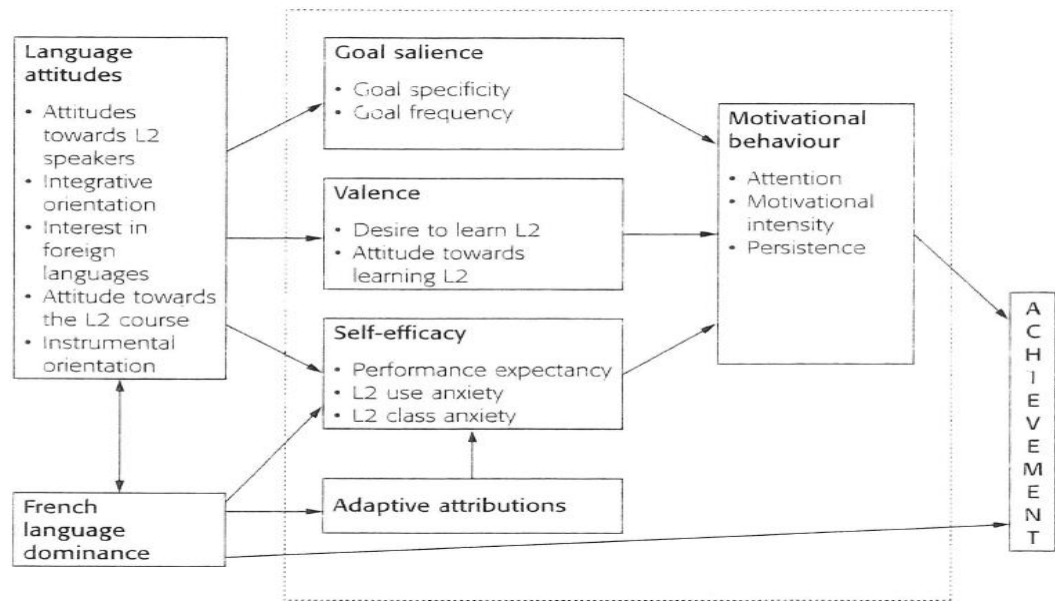


Figure 5. Tremblay and Gardner's (1995) model of L2 motivation.

Yashima (2002) agreed with Dörnyei (1990), as she argued that Japanese usually have little contact with English speakers in their daily lives and they do not have a clear affective response to specific L2 groups. Thus, she proposed a new construct, International Posture, in her model. She defined International Posture as an “interest in foreign or international affairs, willingness to go overseas to stay or

work, readiness to interact with intercultural partners, and, one hopes, openness or a non-ethnocentric attitude toward different cultures, among others” (p. 57).

International Posture had a strong positive path to L2 Learning Motivation ($\beta = .79$, $p < .01$) and a weak path to WTC ($\beta = .22$, $p < .01$) (Yashima, 2002).

Motivation and the Self-Assessment of L2 Proficiency

Most studies that investigated the relationship between motivation and self-ratings of L2 skills have shown a significant relationship between the two. For example, Baker and MacIntyre (2000) reported that perceived competence in French was significantly correlated with an attitude/motivation index that assessed students’ attitudes and motivation toward French and French Canadians ($r = .40$, $p < .01$). Moreover, Yashima (2002), as shown above, found that communicative competence had a significant path to a composite motivation variable made up of motivational intensity and desire to learn English ($\beta = .41$, $p < .01$). AlFallay (2004), whose study was reviewed in the Self-Esteem section, also reported that while students with low self-esteem self-assessed most accurately (the correlation between self-assessment and teacher-assessment was $r = .85$, $p = .0001$), the second most accurate group was the integratively motivated students ($r = .84$, $p = .0001$), while the instrumentally motivated students were the least accurate ($r = .35$, $p = .13$).

Noels, Clément, and Pelletier (1999) investigated the relationships between motivation and educational variables with 78 Anglophone students who registered

in a summer French immersion program in Canada. Noels et al. applied self-determination theory (Deci & Ryan, 1985) to language learning. Motivation was seen to include amotivation, intrinsic motivation, and three subtypes of extrinsic motivation: external regulation, introjected regulation, and identified regulation. The educational variables were the self-assessment of L2 competence in writing, listening, reading, and speaking. The results indicated that self-assessment of these four skills correlated significantly with amotivation ($r = -.22, p < .05$) and intrinsic motivation ($r = .34, p < .01$). Therefore, Noels et al. suggested that feeling amotivated was associated with lower self-evaluation, whereas feeling intrinsically motivated leads to higher self-evaluation.

Motivation was examined through interviews in Ushioda (2001), who analyzed the relationship between motivation and the self-perception of L2 ability in her study of factors that motivate second language learners and their motivational evolution. The participants were 20 students at Trinity College Dublin, Ireland, taking a French (L2) course. The students had studied French for five to six years. The data were collected from two interviews (December 1991 and March–April 1993), which were conducted in English (L1). Their course grades and C-test scores were also obtained. The analysis of the interview data revealed two major motivating factors, Language-related enjoyment/liking and Positive learning history. As for the latter, the students made statements such as “You keep up something you’re good at” or “I always found learning languages very easy at school” (p. 106). Therefore, the students who had the motivational impetus of a

positive learning history seemed to assess their L2 ability highly. Among the eight motivating factors, positive learning history had the highest correlation with course grades ($r = .59, p < .01$) and C-test scores ($r = .46, p < .05$). Therefore, Ushioda (2001) concluded that “the more proficient subjects in the sample attributed greater motivational importance to perceptions of L2 ability and a positive learning history” (p. 108).

Moreover, the relationship between motivation and self-assessment was examined in Masgoret and Gardner’s (2003) large-scale meta-analysis study that investigated the relationship of language achievement (course grades, objective measures such as cloze tests and grammar tests, and self-rated proficiency) with the three primary components of the AMTB (Integrativeness, Attitudes Toward the Learning Situation, and Motivation) as well as its relationship with measures of Integrative Orientation and Instrumental Orientation. The researchers examined 75 independent samples involving 10,489 individuals and reported the following findings. First, the five classes of variables were all positively related to L2 achievement. The mean corrected Pearson correlations of grades with the variables were .24 (Attitudes toward the Learning Situation), .24 (Integrativeness), .37 (Motivation), .20 (Integrative Orientation), and .16 (Instrumental Orientation). The correlations of objective measures with the motivational variables were .17 (Attitude toward the Learning Situation), .21 (Integrativeness), .29 (Motivation), .15 (Integrative Orientation), and .08 (Instrumental Orientation). The correlations of self-ratings with the motivational variables were .26 (Attitudes

Toward the Learning Situation), .26 (Integrativeness), .39 (Motivation), .20 (Integrative Orientation) and .16 (Instrumental Orientation). These figures indicated that among the three achievement measures, self-ratings of achievement had the strongest relationship with the AMTB, with self-rating and Motivation displaying the highest correlation (.39).

Section Summary

Since Gardner and his associates directed their attention to the influence of motivation on L2 learning in 1950s, numerous researchers investigated L2 motivation and many of them focused on its relationship with L2 proficiency or with other affective variables. Generally, L2 motivation was found to have positive relationships with L2 proficiency or with affective variables such as positive attitude toward L2 learning and desire to learn L2, but negative relationship with language anxiety.

Because self-perceived competence is considered as an antecedent of motivation in motivation models, some researchers calculated the correlations between L2 motivation and self-perceived competence of L2 skills and have found that greater motivation is related with higher self-assessment. Masgoret and Gardner (2003) even reported that motivation was correlated greater with self-assessment than with objective measures.

Self-Confidence

Another important affective variable is self-confidence, “the belief that a person has the ability to produce results, accomplish goals, or perform tasks completely” (Dörnyei, 2005, p.73). A number of studies have indicated that self-confidence is important for second language acquisition. For example, Clément and Kruidenier (1985) stated that “self-confidence is the most important determinant of motivation to learn and use the second language” (p. 24). Noels and Clément (1996) maintained that self-confidence leads to increased usage of and communicative competence in the second language.

The concept of self-confidence was first introduced into the L2 literature by Clément, Gardner, and Smythe (1977), who investigated the motivational characteristics of 304 Grade 10 and 11 francophone students in Canada. The results of a factor analysis identified a Self-Confidence with English factor that had high positive loadings from the self-ratings of the four skills (.58 to .84) and high negative loadings from two types of anxiety, English Class Anxiety and English Use Anxiety (-.69 to -.77).

Based on the findings reported in Clément et al. (1977), Clément (1980) proposed a model (Figure 6) in which self-confidence is developed through the frequency and quality of interethnic contact, which leads to motivation and eventually to communicative competence. Clément (1980) maintained that a learner’s degree of self-confidence is influenced by the frequency and quality of interethnic contact; “a high frequency of pleasant contacts will have a more positive

outcome than a low frequency. Conversely, much unpleasant contact will have a more negative effect than a little contact” (p. 151).

Clément’s model was tested by Clément and Kruidenier (1985) with 1,180 francophone students in Grades 7, 9, and 11 in Canada. The students completed the attitude questionnaire, which was designed to measure Integrativeness, Fear of Assimilation, Contact with Anglophones, Self-Confidence in English, and Motivation. Estimates of the participants’ language proficiency measures were obtained from the teachers. The results of a structural equation model supported Clément’s hypotheses by showing the following structural relationships: Integrativeness ($\beta = .45$) and Fear of Assimilation ($\beta = -.44$) predicted Inter-Ethnic Contact, which was a predictor of Self-Confidence ($\beta = .70$), which predicted Motivation ($\beta = .51$). Finally, Motivation predicted two language outcomes (Learning Behavior [$\beta = .50$] and Evaluation by Teacher [$\beta = .20$]).

Clément (1987) tested the same model with 293 francophone university students in Canada, with either a majority ($n = 183$) or a minority ($n = 110$) background. The participants completed an attitude and motivation questionnaire and took a general proficiency test that was made up of cloze, listening, and reading sections, and their oral proficiency was assessed with an English oral interview. A factor analysis identified three factors: Self-Confidence and Proficiency, Integrative Motive, and Ethnolinguistic Vitality. The results of a *t*-test indicated that the minority group had more self-confidence and higher English proficiency than the majority group, but no significant difference was found for motivation. A multiple

regression also showed that the self-confidence index was the best predictor of oral proficiency for both groups (Majority, $\beta = .73, p < .01$; Minority, $\beta = .63, p < .01$). Accordingly, the results of Clément (1987) contradicted Clément's model in that self-confidence rather than motivation had a strong relationship with communicative competence.

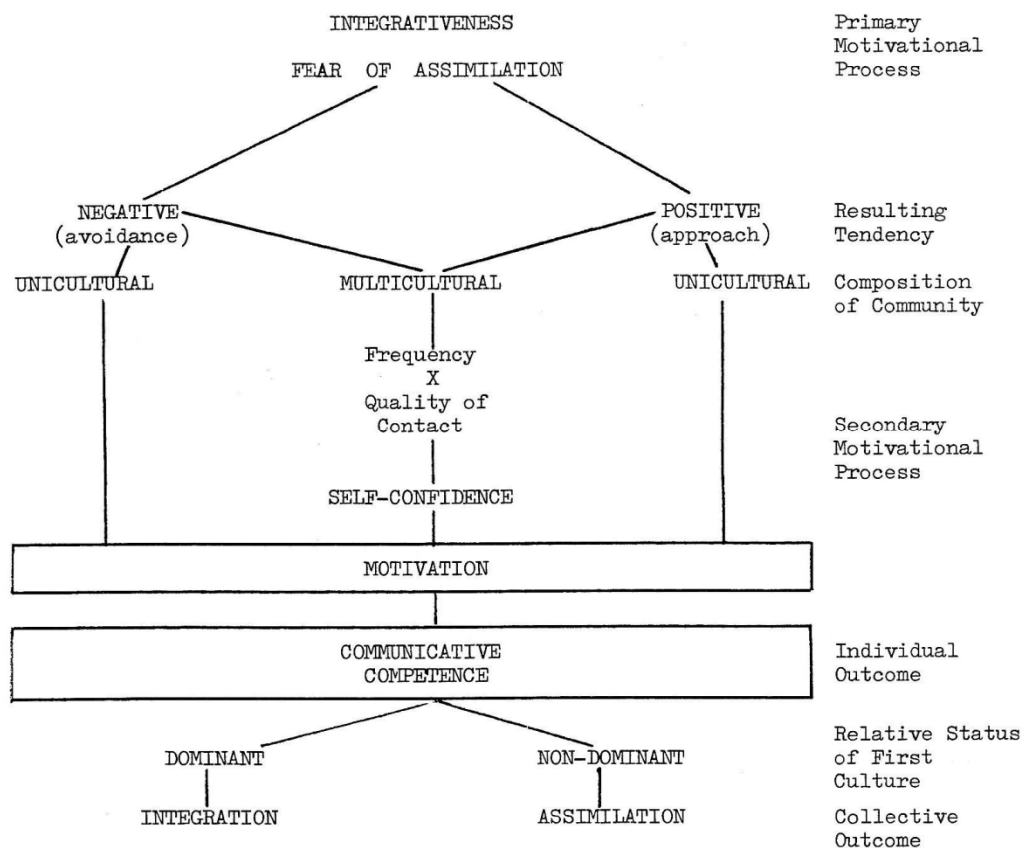


Figure 6. Schematic representation of individual mediational processes (Clément, 1980).

Noels and Clément (1996) examined the influence of ethnolinguistic vitality on the relationships among interethnic contact, self-confidence, identity, and

feelings of adjustment. The participants were 285 native English speakers and 243 native French speakers, all of whom were undergraduate students at a bilingual university in Canada. The ANOVA results showed that the Anglophone students had less contact and less self-confidence in their second language than the Francophone students. The results of a path analysis showed that “contact with the second language community predicts self-confidence in the second language, which in turn predicts feelings of second language group identity and psychological adjustment” (p. 224). Therefore, in Noels and Clément (1996), self-confidence was shown to play an important role in mediating the relationship between interethnic contact and identity and adjustment.

Although Clément’s model emphasized that intercultural contact generates self-confidence, Clément, Dörnyei, and Noels (1994) extended the applicability of self-confidence into an EFL context where learners have little direct contact with native speakers of the target language. Investigating 301 secondary school students in Hungary, Clément et al. found that students’ self-confidence in the L2 correlated significantly with teachers’ ratings of the students’ L2 proficiency; communication proficiency showed the strongest relationship ($r = .49, p < .001$). Therefore, L2 self-confidence was found to be important not only in ESL contexts but also in an EFL context.

Section Summary

Compared with other L2 affective variables, such as language anxiety and motivation, fewer researchers have investigated self-confidence of L2 proficiency; most such studies were conducted by Clément and his associates. Because Clément defined self-confidence as high self-perceptions of communicative competence and low levels of language anxiety, the self-confidence construct used by Clément and his associates is a composite index comprised of anxiety using the second language, self-confidence using the second language, and self-evaluation of second language proficiency. This self-confidence index was shown to be important in L2 acquisition and some studies reviewed above indicate that self-confidence was positively correlated with L2 proficiency. However, no researchers have explored the relationships between self-confidence and self-assessment bias of L2 skills.

Oral Proficiency

Luoma (2004) stated that speaking a foreign language is very difficult because learners “must master the sound system of the language, have almost instant access to appropriate vocabulary and be able to put words together intelligibly with minimum hesitation” (p. ix). Hughes (2003) listed several distinctive features of speaking by contrasting it with writing (Figure 7), which shows that while users can have time to plan, edit, and correct for writing, users need to plan, process, and produce instantly for speaking, and the spoken form of language is fundamentally transient.

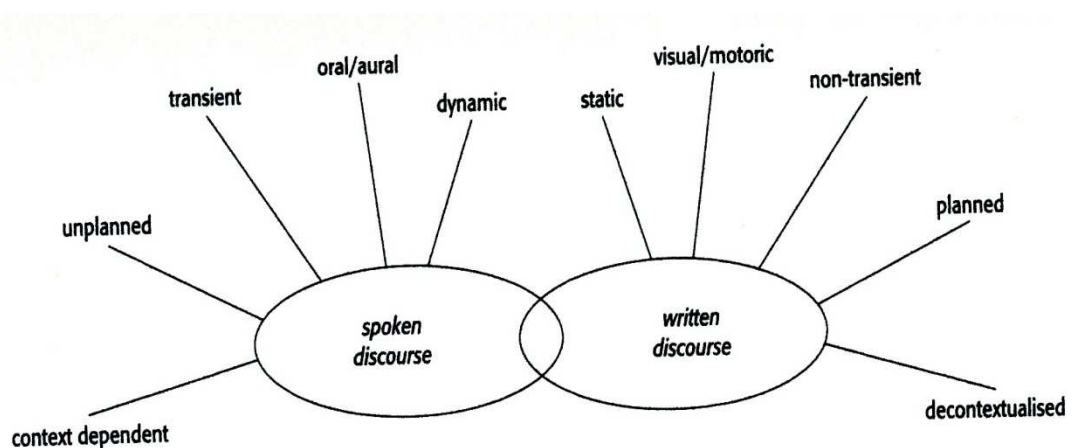


Figure 7.Aspects of production (Hughes, 2003).

Speaking skill includes several different competences. For example, Canale and Swain (1980) defined communicative competence as “a synthesis of knowledge of basic grammatical principles, knowledge of how language is used in social contexts to perform communicative functions, and knowledge of how utterances and communicative functions can be combined according to the principle of discourse” (p.20). Bachman and Palmer (1982) established the framework for communicative competence, which comprises three main components, grammatical competence, pragmatic competence, and sociolinguistic competence. Indeed, many oral assessment rubrics such as the ACTFL and IELTS contain several categories such as function, content, and accuracy.

Some researchers investigated which factors distinguish levels of speaking proficiency by comparing overall scores and analytic scores. Adams (1980) reported that vocabulary and grammar were found to be the main factors to distinguish the levels, while accent and fluency were not. Higgs and Clifford

(1982) found that at lower levels vocabulary and pronunciation mainly contributed to the overall scores, but the contributions of fluency and grammar increased at the higher levels. McNamara (1990) revealed that Resources of Grammar and Expression were the most difficult category as well as the strongest determinant of the overall scores. Iwashita, Brown, McNamara, and O'Hagan (2008) reported that vocabulary knowledge and production features had an impact on the overall scores, and especially vocabulary and fluency had the strongest effect sizes, while grammatical accuracy did not predict the overall scores of L2 speaking proficiency.

Japanese university students' L2 speaking proficiency has been examined by several researchers. Koizumi and In'nami (2013) found that L2 vocabulary knowledge, size, and depth predicted 84% of L2 speaking proficiency of the Japanese students at low to intermediate levels. Ockey (2009) investigated 225 Japanese university students who took a group oral test. Their mean Grammar scores were the lowest (2.47), while Fluency (2.69) and Pronunciation (2.59) were higher. Bonk and Ockey (2003) reported similar results using Rasch logit measures by administering two group discussion tests to 1,324 Japanese university students. They reported that in the first administration, the Rasch measures for each category were Grammar (.57), Vocabulary (.55), Fluency (.03), and Pronunciation (-.87) and in the second administration, the measures for each category were Grammar (-.45), Vocabulary (-.56), Fluency (-.98), and Pronunciation (-1.22). Sato (2011), who examined 30 Japanese university students' English monologues, reported somewhat different results: Fluency (.23 logits) and Vocabulary (.23 logits) were

perceived as the difficult categories, while Grammar (-.05 logits) was easier and Pronunciation was the easiest (-.41 logits).

Oral Proficiency Assessment

Some researchers measured oral proficiency through indirect testing such as class grades (Chen et al., 1999; Clément et al., 1994), cloze tests (Gardner & MacIntyre, 1993; Le Blanc & Painchaud, 1985), and self-assessment (Liu & Jackson, 2008). It is probably easier to collect data using indirect tests than conducting direct tests such as oral interviews, which are laborious and time-consuming. Folsie (2006) also mentioned that scoring an indirect assessment is easier than direct testing because the errors can be easily identified in the indirect tests. However, direct methods are considered better than indirect methods for the assessment of speaking because direct methods evaluate speaking skills in actual performance (Ginther, 2012).

Speaking assessment methods through interviews are called oral proficiency interview (OPI). One well-known OPI is the American Council of Teachers of Foreign Languages Oral Proficiency Interview (ACTFL), which is a holistic scale that includes four factors of the speech, function, content, context, and accuracy, and distinguishes L2 speakers into 11 levels from novice low to distinguished. Another famous example of OPI is the speaking section of the International English Language Testing System (IELTS), which is an analytical scale that includes four categories, fluency and coherence lexical resource, grammatical range and accuracy,

and pronunciation, and the scale ranges from 0 to 9. In the Japanese context, the Kanda English Proficiency Test (KEPT) is used for assessing Japanese university students' English speaking proficiency at Kanda University in Japan. The KEPT is more similar to the IELTS than the ACTFL because it is an analytical scale that provides separate scores for each category. Unlike the IELTS, the KEPT has six levels instead of nine for each category and the descriptions for each category level match Japanese university students' speaking abilities because the assessment was created to assess Japanese university students' oral proficiency. For example, Level 5, the highest level, requires "near native-like," while the ACTFL and the IELTS require "native-like" to meet the highest level. Moreover, the pronunciation category includes the words "katakana-like speech," which describes some Japanese students' poor pronunciation. Therefore, in this study, the KEPT is used to assess the participants' English speaking skills.

Kanda English Proficiency Test

According to Thompson (2009), the KEPT was first used in 1989. It consists of five subtests: reading, grammar, listening, writing, and speaking. The speaking test is a ten-minute oral discussion among four students. The test-takers are given the prompt, which includes a short text about a simple everyday topic such as friends, family, and food. Two raters in the room observe the discussion but do not interact with the students while rating the four students individually. An analytic scoring rubric that consists of five categories, pronunciation, fluency, grammar,

vocabulary, and communication strategies, is used. Pronunciation category measures pronunciation, intonation, and word blending; Fluency category measures automatization, fillers, and speaking speed; Grammar category measures use of morphology, complexity of syntax (relativization, embedded clauses, parallel structures, and connectors); the Vocabulary category measures the range of vocabulary used (Ockey, 2011). Although there are five levels in each category, it is a nine-point interval scale due to the use of half points (i.e., 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0).

Bonk and Ockey (2003) examined four facets modeled in two KEPT test administrations (examinee, prompt, rater, and rating items) with 1,324 examinees for the first administration and 1,103 examinees for the second administration, who were first-, second-, and third-year Japanese university English language majors. In both administrations, the data fit the Rasch model and unidimensionality held. It was found that the Rasch model reliably separated the examinees by ability and that ratings did not differ greatly from prompt to prompt. Great differences in rater severity were found among new raters, but the authors maintained that the raters tended to increase their internal consistency as they gained experience. An analysis of the scales revealed that the gradations of scale steps worked effectively. Items were found to show acceptable fit to the Rasch model; however, the raters had difficulty distinguishing categories at the ends of the scale for pronunciation and communicative skills.

Gaps in the Literature

The review of previous research reveals three major gaps in the literature. First, although many researchers have investigated the validity of self-assessment of L2 speaking, the participants' L2 speaking skills are not measured directly in some studies. Some researchers have utilized class grades, standardized test scores, or cloze tests to measure communicative competence (e.g., Le Blanc & Painchaud, 1985; Chen et al., 1999; Clément et al., 1994). These types of measures are indirect indicators of speaking proficiency at best and invalid indicators in the worst cases. In order to obtain more reliable and valid measures of speaking ability, it is necessary to assess speaking directly by having the participants speak the target language.

In contrast, some researchers have participants speak the L2 to measure their speaking skills directly; however, there was often a mismatch between self- and teacher-assessments because they did not utilize the same assessment rubric. Moreover, teacher-assessment has not always been properly conducted because multiple raters were not used in many studies, and even when multiple raters were used, they used raw scores for self- and teacher-assessments. Recently, more researchers have utilized multi-faceted Rasch measurement for L2 speaking testing because it can produce person ability measure estimates that are adjusted for rater bias and because it can provide detailed information, such as category difficulty, rating criteria fit to the Rasch model, and rater severity. Nevertheless, no

researchers have utilized multi-faceted Rasch measurement to investigate the validity of self-assessment of L2 speaking skills.

The second gap is that causal relationships of how L2 affective variables influence self-assessment of L2 oral performance have not been adequately studied. Most of the research on self-assessment has concerned the validity of self-assessment questionnaires; less attention has been paid to the influence of L2 learners' psychological traits on their L2 self-assessment. It is important to investigate this relationship because how L2 learners perceive their L2 proficiency influences their L2 use and their capacity for future improvement. In the present study, I investigate L2 learners' self-assessed L2 speaking ability and the degree to which seven affective variables influence the self-assessment of L2 oral performance: Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence. Although many previous researchers have calculated correlations between L2 affective variables and self-assessment, in this study I produce bias scores of self-assessment using Facets and investigate causal relationships among L2 affective variables and self-assessment bias of L2 oral performance using structural equation modeling.

The first affective variable, self-esteem, is hypothesized to influence self-assessment (Brown, 2006; Heine et al., 2001); however, in most cases, studies of self-esteem have been conducted by psychologists who are interested in how self-esteem affects the self-assessment of general traits and abilities such as

attractiveness and friendliness. On the other hand, few researchers (e.g., AlFallay, 2004; Anderson, 1982) have investigated its relationship to the self-assessment of L2 skills.

The second variable, language anxiety, has been investigated in terms of its relationship to L2 proficiency and other affective variables; however, few researchers (e.g., Chen et al., 1999; MacIntyre et al., 1997) have investigated its influence on the self-assessment of L2 proficiency. Most of the previous results have indicated that anxiety affected self-assessment negatively and that among the four L2 skills, speaking had the strongest negative association with anxiety.

The third variable, willingness to communicate, has been investigated in terms of the degree to which self-perceived competence predicts WTC in structural equation models(e.g., MacIntyre et al., 1999; Yashima, 2002).However, because L2 perceived competence is not usually compared with actual L2 competence, the accuracy or bias of the self-assessment of one's L2 skills is rarely considered.

The fourth affective variable, motivation, has primarily been investigated in terms of the degree to which self-perceived competence predicts motivation (Baker &MacIntyre, 2000; Yashima, 2002).To date, only one researcher has examined its relationship to the accuracy of L2 self-assessment: AlFallay (2004) investigated this issue and found that integratively motivated learners showed relatively high accuracy, while instrumentally motivated learners were the least accurate.

The fifth affective variable is self-confidence. As stated above, self-confidence has usually been viewed as a composite index that is made up of

anxiety using the second language, self-confidence using the second language, and self-evaluation of second language proficiency. While the relationship between anxiety and the self-assessment of L2 proficiency has been examined by researchers such as MacIntyre, Noels, and Clément (1997), little research has been conducted investigating the relationship between self-confidence using the second language and the self-assessment of second language proficiency.

The third gap is that some researchers have investigated the difference in the self-assessment of L2 skills between high and low proficiency learners. According to the Dunning-Kruger Effect proposed by Kruger and Dunning (1999), low proficiency individuals tend to overestimate their abilities because their incompetency prevents them from assessing themselves accurately, while high proficiency individuals tend to underestimate their own abilities because they are likely to consider their high proficiency is shared by others. The Dunning-Kruger Effect can be applied to L2 learners, for some studies reported that high achievers were more self-critical and low achievers were less self-critical (e.g., Jassen-van Dieten, 1989; Patri, 2002). On the other hand, Matsuno (2007) found that lower proficiency students did not display a common tendency toward overestimation or underestimation. However, no researchers have investigated to what degree L2 affective variables influence self-assessment of L2 proficiency differently between high and low proficiency learners that are measured with multi-faceted Rasch analysis.

Purposes of the Study

The first purpose of the study is to investigate how Japanese students rate their own L2 oral performance in comparison to teachers' assessment in order to determine whether Japanese students tend to underestimate or overestimate four aspects of their L2 oral performance. In this study, in order to minimize bias from sources other than affective factors, the students' L2 oral performances are measured using oral interviews, rather than indirect measures, such as course grades or cloze tests, and the teacher- and self-assessments are conducted using the same assessment rubric. Moreover, L2 speaking subskills are investigated by utilizing an analytic scale that consists of four subcomponents: grammar, vocabulary, fluency, and pronunciation. These subcomponents make it possible to investigate multiple aspects of the participants' oral performance and demonstrate in which categories the students' self-assessments and teachers' ratings differ. In addition, to accurately evaluate the participants' speaking performance, multiple raters are used in this study. Instead of using raw scores for student- and teacher-assessments, I utilize multi-faceted Rasch analysis, which can minimize teacher-rater bias and provide detailed information about issues such as speaker ability, rater severity, task difficulty, and category difficulty.

This first purpose is important because how learners evaluate their own L2 oral skill has a potentially strong effect on their current L2 use as well as their future achievement. Determining how accurately Japanese students perceive their L2 oral performance can help them improve their oral skills by informing them of

which subskills they overestimate and underestimate. This information is also useful for teachers, as it can help them better understand their students' perceptions of their oral proficiency.

The second purpose is to investigate the degree to which Japanese students' self-assessment is influenced by L2 affective variables and which affective variables lead to accurate evaluation, underestimation, and overestimation of their L2 oral skills. Although some researchers have investigated the influences of L2 affective variables on the self-assessment of L2 proficiency (AlFallay, 2004; Gardner & MacIntyre, 1993), they mainly focused on the correlations between L2 variables and self-assessment. In this study, I calculate the bias size² of self-assessments and investigate causal relationships of how L2 affective variables influence self-assessment using structural equation modeling.

The L2 affective variables included in this study are Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, L2 Speaking Motivation, and L2 Self-Confidence. L2 Speaking Motivation consists of three subconstructs: Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, and Desire to Learn to Speak English.

A measure of Self-Esteem is included because it is considered an important factor influencing the act of self-assessment (Heine et al., 2001; Todd, 2002). Moreover, some Japanese might have lower self-esteem than persons from western

²Raters may display particular patterns of severity or lenience in relation to some candidates or some tasks. The multi-faceted Rasch measurement identifies bias size through bias analysis. If bias size is positive, the self-rater is more lenient than the teacher rater. If it is negative, the self-rater is more severe than the teacher-rater (Linacre, 2014).

countries because self-criticism is a traditional virtue in Japan (Heine, Kitayama, & Lehman, 2001; Markus & Kitayama, 2000).

L2 Speaking Anxiety is included because anxiety is the most frequently researched affective variable in studies of the self-assessment of L2 skills and many researchers have reported negative correlations between anxiety and the self-assessment of L2 skills (Chen, Horwitz, & Schallert, 1999; Gardner & MacIntyre, 1993; MacIntyre et al., 1997).

L2 Willingness to Communicate is considered to be related with the frequent engagement in L1 and L2 communication for L1 and L2 speakers (McCrosky, 1987; Yashima, 2002), and it is believed to be a trait communication construct that exerts a substantial impact on communicative behavior (Baumeister, Campbell, Krueger, & Vohs, 2003; McCroskey, 1997). Therefore, it is a potentially important variable to include in a study of the self-assessment of L2 oral skills.

Fourth, empirical research indicates that L2 Speaking Motivation can influence self-assessment. In a meta-analysis conducted by Masgoret and Gardner (2003), self-ratings of L2 ability were most highly correlated with motivation as measured by the Attitude/Motivation Test Battery (Gardner, 1985). Because that study was a meta-analysis, self-ratings involved all four L2 skills; however, in this study, the focus is on speaking and the degree to which Japanese students' motivation for L2 speaking affects their self-assessment of L2 oral performance.

Finally, L2 Speaking Self-Confidence plausibly influences the self-assessment of speaking proficiency. For example, Clément's (1980) model includes

the concept of self-confidence, in which language anxiety and the self-evaluation of language proficiency are subsumed. This model was tested by Clément (1987), who reported that self-confidence was the best predictor of oral proficiency.

This second purpose is important because previous researchers have investigated the influence of these affective variables on L2 oral performance; in this study I focus on their influence on students' self-assessment of L2 oral performance. Moreover, providing empirical findings regarding this issue is important because by determining which affective variables influence students' self-evaluation, teachers can identify better ways to instruct Japanese students.

The third purpose is to investigate to what degree the influences of L2 affective factors on self-assessment bias of L2 oral performance differ between higher and lower proficiency learners. The existence of the Dunning and Kruger Effect is examined with Japanese university students. Instead of using correlations of the raw scores, the bias sizes of self-assessment are used to determine how the influences of L2 variables on self-assessment differ between the two proficiency levels.

Investigating this issue is important because the influences of psychological traits on language learning are complex and not yet fully understood. By shedding light on the relationships among those variables and how they differ between learners at two proficiency levels, teachers can better instruct learners at different proficiency levels and choose the most effective strategies for teaching according to learners' proficiency.

Research Questions

The following research questions are the focus of this study.

1. To what degree do Japanese university students' self-assessments of their L2 oral performance differ from teacher-assessments?
2. Which affective variables, Self-Esteem, L2Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence, predict the students' self-assessment of L2 oral performance?
3. What are the characteristics of the students who could assess their own L2 oral performance accurately?
4. To what degrees do the seven affective variables affect the self-assessment of high and low proficiency students differently?

CHAPTER 3

METHODS

Participants

The participants in this study were 400 university students. Among them there were two exchange students from China and one from South Korea, so these three students were eliminated from the study because their native language was not Japanese. Moreover, five students had lived in English-speaking countries more than three years when they were young. One spoke native-like English and was therefore eliminated from the study. The other four students did not speak English very fluently, so they were included in the study. Therefore, data from 396 students were analyzed in the main study.

Three hundred forty-eight of the students (249 male and 99 female students) were enrolled in a private university in eastern Japan (University A), and among them, 293 students (218 male and 82 female students) were majoring in engineering, 24 students (21 male and 3 female students) in information science, and 36 students (12 male and 14 female students) in sports science. Forty-eight students (39 male and 9 female students) were enrolled in a second private university in eastern Japan (University B) and they were all engineering majors. The participants' ages ranged from 18 to 22.

University B is considered one of the most prestigious and highly competitive universities in Japan (*hensachi* rating is 66), so most of the students have high-

intermediate proficiency in English reading and grammar. University A is less competitive (*hensachi* rating is 45), so many of the students have low-intermediate proficiency in English reading and grammar. The participants have mainly studied reading and grammar in their secondary school English courses and have had limited opportunities to develop oral skills. Because they have few opportunities to speak English in their daily lives, their speaking proficiency is not very high. The students at University B generally have higher oral proficiency than those at University A because they have larger English lexicons and more knowledge of English grammar. The students at University A take two 90-minute English classes per week (one speaking/listening class and one reading/writing class) as first-year students, and take one 90-minute English class per week (one reading/writing class) when in their second year. The students at University B take one 90-minute English class per week (a general four-skills English class) as first- and second-year students.

The participants were informed that participation in the study is voluntary; each participant was paid 500 yen. I obtained written consent from the participants (Appendix A). They were informed that (non)participation in the study had no effect on their course grades, that they could withdraw from the study at any time, and that they would be informed of the results, if they wished to know them.

Raters

In addition to the student participants, five raters took part in the study. All the raters were English teachers at Japanese universities. Three raters were native English speakers and two raters were native Japanese speakers. Rater 1 is a female American in her early 30s. She received B.A. and M.A. in English Literature at Columbia University in New York. She had three years of experience teaching English at a Japanese university as an adjunct professor. Rater 2 is a male Japanese professor in his mid-60s. He received B.A. in English at a Japanese university, and M.A. and Ph. D in Philosophy at an American university. He taught Philosophy, and Japanese Language and Culture at an American university as a full-time professor for 20 years. Then, he became a professor at a Japanese university and has taught English and Philosophy classes for 15 years at a Japanese university. Rater 3 is a male Canadian in his early 50s. He received M.A. in TESOL in a Canadian university. He has taught English for twelve years at several Japanese universities as an adjunct professor. Rater 4 is a male Canadian professor in his late 50s. He received M.A. in TESOL at British Columbia University in Canada. He has taught English more than 20 years at a Japanese university. Rater 5 is a female Japanese associate professor in her early 40s. She received B.A. and M.A. in English Literature at a Japanese university, and M.A. in TESOL at an American university. She has taught English at Japanese universities as an adjunct professor for eight years and as a full-time professor for five years.

All the raters were English teachers at Japanese universities and had experience teaching speaking English to Japanese university students, so they had experience testing the Japanese students' speaking abilities in class in order to give the marks. Therefore, all the raters had sufficient education and teaching backgrounds to assess Japanese university students' speaking performance.

As for rater training, I met each rater individually and explained about speaking assessment for about 20 minutes. First, I explained the organization of the interview by showing Appendix B. All the teachers understood the questions and the cartoon. Second, the rating scale (Appendix C for the English version and Appendix D for the Japanese version) was shown to them. I asked the raters to look at the scale and read the descriptions. Some of them did not understand the half point at first, so I explained it. No raters had problems understanding the analytic scale and distinguishing the categories in part because the descriptions for each category at each level were adequate.

As a preliminary analysis, Raters 3 and 4 rated six students' speaking performance in order to examine if the raters are able to assess L2 speaking performance using the speaking assessment rubric. The results are reported in the Preliminary Analysis section.

Instrumentation

A questionnaire and an oral assessment scale were used in this study. The questionnaire was used to measure seven L2 affective factors: Self-Esteem, L2

Speaking Anxiety, L2 Willingness to Communicate, L2 Speaking Motivation (composed of Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, and Desire to Learn to Speak English), and L2 Speaking Self-Confidence. The oral assessment instrument was used by the teacher raters to assess the participants' L2 oral performance and by the participants to self-assess their own L2 oral performance.

Self-Esteem Scale

The Self-Esteem scale was based on the Rosenberg (1965) self-esteem scale, one of the most widely used instruments for measuring global self-esteem (See Appendix E for the English version and Appendix F for the Japanese version). Global self-esteem is defined as “a favorable or unfavorable attitude toward the self” (Rosenberg, 1965, p. 15). The scale is made up of ten questions asking people's general feelings toward themselves without specifying any qualities or attributes. Although five items were worded negatively in the original instrument, they were rewritten, so only positively worded items were used in this study. A sample item is “I believe that I have a number of good qualities.” A six-point Likert scale was used: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*. I translated the English questionnaire into Japanese, and the Japanese translation was back-translated into English by a bilingual professor, who was the same professor as Rater 2. His native language was Japanese and after he graduated from a Japanese university, he studied in

Master's and doctor's courses at an American university and since then he has lived in the United States for more than 30 years. He had experience teaching classes using English at an American university for 20 years. Therefore, he had high proficiency in both Japanese and English, so I asked him to conduct the back-translation. The results were compared with the original. Even though the word *positive* was at first translated as *sekkyokuteki* (積極的) in Japanese, I decided to use the word *positive* (ポジティブ) because Japanese use the word frequently and know its meaning. No other problem was found in the translation.

L2 Speaking Anxiety Scale

The L2 Speaking Anxiety scale was created based on the Foreign Language Classroom Anxiety Scale (FLCAS) (Horwitz et al., 1986) (See Appendix G for the English version and Appendix H for the Japanese version). In this study, Speaking Anxiety includes communication apprehension, which Horwitz et al. (1986) describes as “an attitude characterized by fear of or anxiety about communicating with people” (p. 30) and a fear of negative evaluation, which Watson and Friend (1969) described as an “apprehension about others’ evaluations, avoidance of evaluative situations, and the expectation that others would evaluate oneself negatively” (p. 449). The scale consists of 10 items designed to measure anxiety caused by speaking English. A sample item is “I worry that other students think my English speaking ability is low.” The participants answer each question using a six-point Likert scale: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 =

Slightly agree, 5 = *Agree*, and 6 = *Strongly agree*. I translated the questionnaire into Japanese, the Japanese translation was back-translated into English by a bilingual professor, who was the same professor as Rater 2, and the results were compared with the original. In item 4, the word *activity* was originally translated as the katakana form of the word *activity* (アクティビティー); however, because the Japanese word *katsudo* (活動) is more comprehensible for Japanese students than the katakana form, I used the Japanese word.

L2 Willingness to Communicate Scale

The L2 Willingness to Communicate scale was designed based on a questionnaire developed by Sick and Nagasaka (2000). L2 Willingness to Communicate is defined as “a readiness to enter into discourse at a particular time with a specific person or persons, using a L2” (MacIntyre et al., 1998, p. 547). The scale consists of 12 items describing specific situations in which the participants might use English, and they are asked to rate their willingness to speak English in each situation (See Appendix I for the English version and Appendix J for the Japanese version). An example item is “I would be willing to answer a question from my teacher in English class.” The participants indicate their degree of L2 WTC using a six-point Likert scale: 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*. I translated the questionnaire into Japanese, the Japanese translation was back-translated into

English by a bilingual professor, Rater 2, and the results were compared with the original. No problem was found in the translation.

L2 Speaking Motivation Scale

The L2 Speaking Motivation scale was developed based on the Attitude/Motivation Test Battery (AMTB) (Gardner, 1985), and some modification were made based on questionnaires used by Gardner, Tremblay, and Masgoret (1997), Yashima (2002), Irie (2005), and Matsuoka (2006). Because this study is an investigation of English speaking ability, this motivation scale includes three categories with ten items each that are related to speaking English: Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, and Desire to Learn to Speak English (See Appendix K for the English version and Appendix L for the Japanese version). Attitude Toward Learning to Speak English measures the degree to which students hold positive attitudes toward speaking English. An example item is “I enjoy speaking English.” L2 Speaking Motivational Intensity measures the amount of effort students make to improve their English speaking ability. A sample item is “I think I try to speak English more than other students.” Desire to Learn to Speak English measures students’ willingness to learn to speak English. An example item is “I would take an English conversation course in school, even if it were not required.” The participants answered each item using a six-point Likert scale, where 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*. I translated the

questionnaire into Japanese, the Japanese translation was back-translated into English by a bilingual professor, Rater 2, and the results were compared with the original. Because it was pointed out that the word *would* in item ALSE6 (I would enjoy talking with native English teachers.) was not translated properly, I changed the translation from *hanasuto* to *hanasetara* (話せたら). Moreover, I did not translate the word *really* in item DLSE3 (I really want to learn to speak English better.); instead, I added the word *tsuyoku* (強く), which means *strongly*, to the Japanese translation.

L2 Speaking Self-Confidence Scale

The L2 Speaking Self-Confidence scale was designed based on Sick and Nagasaka's (2000) questionnaire (See Appendix M for the English version and Appendix N for the Japanese version). The scale consists of 16 can-do items describing specific situations in which the participants might use English. L2 Speaking Self-Confidence is defined as "the belief that a person has the ability to produce results, accomplish goals, or perform tasks completely" (Dörnyei, 2005, p.73). A sample item is "I can hold a 5-minute conversation with my teacher in English." The participants answer each item using a six-point Likert scale, where 1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, and 6 = *Strongly agree*. I translated the questionnaire into Japanese, the Japanese translation was back-translated into English by a bilingual professor,

Rater 2, and the results were compared with the original. No problem was found in the translation.

Oral Assessment Scale

The participants' speaking ability was assessed through a 10-minute two-part oral proficiency test made up of an interview and a picture task (See Appendix B). The participants' performances were assessed with an analytical scale based on the speaking assessment instrument used at Kanda University in Japan, the Kanda English Proficiency Test (KEPT) (Bonk & Ockey, 2003) (See Appendix C for the English version and Appendix D for the Japanese version). This scale was chosen because it was developed for assessing speaking skills of Japanese university students and it was used successfully as a diagnostic testing and a placement testing for Japanese university students (Ockey, 2009, 2011). The KEPT analytical categories are made up of Grammar, Vocabulary, Fluency, and Pronunciation, which are considered to have an impact on the overall assigned scores of L2 speaking performances (Iwashita et al., 2008).

The original KEPT scale consists of five categories but one category, communicative strategies, was not used in this study because KEPT was originally designed to assess oral discussion skills by a group of four students; individual oral interviews are used in this study, so the remaining four categories, grammar, vocabulary, fluency, and pronunciation, are used in this study. Although the original KEPT has six levels, *Does not discuss*, was deleted and the

remaining five levels were used: 1 = *Very weak*, 2 = *Weak*, 3 = *Fair*, 4 = *Good*, and 5 = *Very good*. Each level is accompanied by a description of the performance for that level. Because raters can use half points from levels 1 to 4, it is a nine-point scale. For example, the description for Level 5 of grammar use is as follows: Level 5 (*Very good*) *Uses high level of discourse structures with near native-like accuracy.*

The same scale was used for both the students' self-assessment and the teachers' assessment. I translated the rubric into Japanese, and the Japanese translation was back-translated into English by the bilingual professor, who was the same professor as Rater 2, and the results were compared with the original. In the pronunciation category, *accent* was not initially translated, so I first used the katakana form of the word *accent* (アクセント). However, I came to the conclusion that Japanese word *namari* (なまり) is easier for most Japanese to understand, so *accent* was translated as *namari*, which means that English is spoken with a foreign accent. Moreover, *foreign accent* was first translated as *gaikokutekina accent* (外国的なアクセント), but as the participants are all Japanese students, I translated it as *nihongo namari*, (日本語なまり), or Japanese accent.

Rasch Model

Rasch Rating Scale Model

The Rasch rating scale model can measure item difficulty and person ability, and thus is more advantageous than classical test theory. McNamara (1996) represents the equation for the Rasch rating scale model known as Andrich's Rating Scale model as follows:

$$P = B_n - D_i - F_k$$

where

P = a mathematical expression of the probability of achieving a score within a particular score category on a particular item,

B_n = the ability (B) of a particular person (n) and

D_i = the overall difficulty (D) of a particular item (i), and

F_k = the difficulty (F) of achieving a score within a particular score category (k) on any item.

The questionnaire data in this study are analyzed using the Rasch software, Winsteps 3.63 (Linacre, 2007). The Rasch rating scale model provides three major advantages in analyzing Likert-scale data over using raw scores. First, the Rasch model converts ordinal raw scores to interval measures known as logits (i.e., log-odd units). While classical test theory treats the relative value of all items as having the same values and assumes that the units increase across the rating scale equally, the Rasch model indicates the relative difficulty of each item in comparison with other items in the questionnaire and places both persons and items on a single logit

scale. Therefore, the Rasch model allows us to learn which item is more difficult or easier to endorse with and who is the most or least able person. In the person-item map, the Rasch model set at 50% probability of success for any person on the item located at the same logit scale. Thus, researchers can see the relationships between the person ability and item difficulty estimates visually. Second, the Rasch model provides fit statistics that show how well the items adhere to the expectations of the Rasch model and also allow for the identification of poorly performing items. Thus, by using fit criteria for Infit and Outfit MNSQ, such as .50 to 1.50 (Linacre, 2007), researchers are able to identify and eliminate misfitting items. Third, researchers can check the dimensionality of the items hypothesized to measure the same trait using the Rasch PCA of item residuals analysis (Bond & Fox, 2007; McNamara, 1996). Estimates of person ability and item difficulty are difficult to interpret when the unidimensionality assumption is violated.

Item Fit Analysis

Item fit statistics indicate how well the items meet the requirements of the Rasch model. Outfit statistics are unweighted estimates of the degree of fit of responses. They give more value to off-target observations, so they are more sensitive to the influence of outlying scores (Bond & Fox, 2007). The Outfit mean-square (MNSQ) statistic is simply an average of the standardized residual (Z_{ni}) variance, across both persons and items. The formula for Outfit MNSQ statistic is as follows (Bond & Fox, 2007, p. 285):

$$\text{Outfit MNSQ} = \frac{\sum Z_{ni}^2}{N}$$

On the other hand, Infit MNSQ statistics are a weighted standardized residual that gives more weight to on-target observations; thus, it is more sensitive to unexpected responses close to a person or item's measure. Residuals are weighted by their individual variance (W_{ni}) to decrease the impact of unexpected responses far from the measure. The formula for the Infit MNSQ statistics is as follows (Bond & Fox, 2007, p. 286):

$$\text{Infit MNSQ} = \frac{\sum Z_{ni}^2 W_{ni}}{N W_{ni}}$$

Rasch analysis reports Infit and Outfit MNSQ statistics as two chi-square ratios. Its expected value is 1.0, which indicates perfect fit to the Rasch model. This study adopts Linacre's (2007) Infit and Outfit MNSQ criteria of .50-1.50. Therefore, when MNSQ values are between .50 and 1.50, the items do not greatly diverge from the Rasch model expectations. According to Bond and Fox (2007), an Infit or Outfit MNSQ value of $1 + X$ represents 100X% more variation in the observed data than the Rasch model expects. For example, an Infit MNSQ value of 1.20 indicates 20% ($100 \times .20$) more variation in the observed response pattern than expected. An Outfit MNSQ value of .80 indicates 20% [$100 \times (1-.80)$] less variation than expected.

Rasch analysis also provides standardized (t statistics) forms of fit statistics: Infit ZSTD and Outfit ZSTD. Infit and Outfit t values greater than +2 or less than -2 are considered having less compatibility with the Rasch model (Bond & Fox,

2007). Because the standardized fit statistics is said to be useful for small sample size and sensitive with large sample (Linacre, 2007), unstandardized fit statistics are used in the main study because it deals with about 400 participants.

Principal Component Analysis (PCA) of Item Residuals

While good fit to the Rasch model is an important indicator of item quality, unidimensionality is generally determined through the PCA of item residuals analysis. According to Linacre (2007), the criteria for determining unidimensionality is that variance explained by measures should be over 50%, and unexplained variance by the first contrast should account for either less than 10% of the variance and/or have an eigenvalue less than 3.0 of the variance, with ideal eigenvalues being approximately 2.0. In sum, the criteria for unidimensionality I follow in this study are shown in Table 1.

Table 1. *Criteria for Unidimensionality (Linacre, 2007)*

Criteria	Value
Infit and Outfit MNSQ	.50–1.50
Explained variance	Variance > 50%
Unexplained variance by the first contrast	Variance < 10% or Eigenvalue < 3.0

Rasch Reliability Estimates

The Rasch model provides reliability estimates both for persons and items. The person reliability estimates the replicability of person placement across other items measuring the same construct, and the person reliability index (R_p) is

calculated from total person variability adjusted for measurement error(SA^2_p) divided by unadjusted person variability (SD^2_p) (Bond & Fox, 2007, p.284):

$$R_p = \frac{SA^2_p}{SD^2_p}$$

The item reliability estimates the replicability of item placement among the items measuring the same construct if the same abilities are administered to participants with compatible abilities. A reliability estimate for items (R_I) is calculated from total item variability adjusted for measurement error (SA^2_I) divided by unadjusted item variability (SD^2_I) (Bond & Fox, 2007, p.284):

$$R_I = \frac{SA^2_I}{SD^2_I}$$

Rasch person and item reliability estimates are analogous to Cronbach's alpha. They range from 0 to 1.00 (Bond & Fox, 2007).

Rasch Separation Index

The Rasch model provides a separation index both for persons and items. Person separation is an estimate of the spread or separation of persons on the measured variable, and person separation index (G_p) is calculated from adjusted person standard deviation (SA_p) divided by average measurement error (SE_p) (Bond & Fox, 2007, p.286):

$$G_p = \frac{SA_p}{SE_p}$$

Similarly, item separation is an estimate of the spread of or separation of items on the measured variable, and item separation index is calculated from adjusted item

standard deviation (SA_I) divided by average measurement error (SE_I) (Bond & Fox, 2007, p.286):

$$G_I = \frac{SA_I}{SE_I}$$

Unlike the separation reliability, the separation index is not bound by 0 and 1 (Bond & Fox, 2007).

The Many-Faceted Rasch Model (MFRM)

The many-faceted Rasch model is an extension of the Rasch model. In addition to person ability and item difficulty, the MFRM can assess other variables such as tasks and raters. Therefore, it is useful for assessing language performances such as essays, presentations, and interviews, in which judges or raters are used for assessment.

The mathematical model of the MFRM is as follows (McNamara, 1996):

$$P = B_n - D_i - C_j - F_k$$

Where

P = a mathematical expression of the probability of achieving a score within a particular score category on a particular item from a particular judge,

B_n = the ability of examinee n ,

D_i = the difficulty of item I ,

C_j = the severity of judge j , and

F_k = the difficulty (F) of achieving a score within a particular score category (k) averaged across all judges and all items.

MFRM has several advantages in L2 speaking assessment over a conventional approach. First, the MFRM can provide estimates of ability that are adjusted for rater bias, while speaking scores in a conventional approach using raw scores are likely to be degraded due to differences in rater severity/leniency. Often the interrater reliability is examined by calculating intercorrelations between judge ratings; however, in that case, only consistence among the rank orders of candidates is indicated. However, using the MFRM, we are able to learn the severity or leniency differences between judges (Bond & Fox, 2007).

Second, the MFRM is robust for missing data; in other words, it does not require a complete data set for calculations (Bond & Fox, 2007). In a conventional method, raters need to assess all the examinees to determine interrater reliability, whereas the MFRM does not require a complete data set as long as there is enough overlap, which can reduce time-consuming act of judging (Linacre & Wright, 2002).

Third, the MFRM can provide more information than traditional method. Due to the joint calibration of facets, rater severity can be placed on the same scale as ratee performance and task difficulty, so researchers can “draw useful, diagnostically informative comparisons among the various facets” (Myford & Wolfe, 2003, p.404). For example, if a candidate is located on the same scale with a rating category, the probability of the candidate to be rated in the rating scale is 50%, and the probability of being rated in the higher rating is less than 50%, and the probability of being rated in the lower category is greater than 50%.

Finally, the MFRM allows researchers to conduct bias analyses. Often raters are unlikely to be consistent across all items and all candidates because they may display some patterns of severity or leniency to only some candidates, but not others, or to particular items, but not others. Bias analysis identifies such interactions involving a rater and the systematic patterns of behavior (McNamara, 1996). Bias analysis calculates a bias measure in logits. Observed score, or raw score of the estimable responses, minus expected score, or no bias score based on the measures from the analysis, divided by observed count, or the number of estimable responses, equals observed-expected score average. Therefore, it is the average difference between the observed and the expected ratings (Linacre, 2014). Larger observed score corresponds to higher bias size, which is the size of bias measure in logits, relative to overall measures. If the bias size is positive, the rater is more lenient than expected. If this is negative, then the rater is more severe than expected (Linacre, 2014).

Procedures

Data Collection

The data were gathered from May to August in 2011. The participants completed the 81-item questionnaire, which was designed to measure Self-Esteem, L2 Speaking Anxiety, L2 Willing to Communicate, L2 Speaking Motivation, and L2 Speaking Self-Confidence (see Appendices F, H, I, J, and K, respectively). This took approximately 15 minutes. They completed the questionnaire either at school

or at home and they submitted it when they had the oral interview. Each participant took part in a two-part 5-minute oral interview (Appendix B). During the first part, the participant was asked to respond to five questions and follow-up questions, while in the latter part, the participant told a story in English while looking at cartoons. The interviews were tape-recorded with a Sony IC recorder ICD-SX813, and after completing the interview and story-telling task, from their memory of their performance each participant self-assessed their own oral performance using an analytical scoring rubric (Appendix D). Using the tape-recorded data, five raters independently assessed each participant's oral proficiency using the same scoring rubric (Appendices C and D).

Data Analysis

The first research question, "To what degree does Japanese students' self-assessments of their L2 oral performance differ from teacher-assessments?," was answered by looking at the Facets results. The Facets indicates how severe/lenient the students' self-assessments were compared to the teacher-rater assessments. In addition, the Pearson correlations were calculated between the participants' self-assessment and teacher-assessment Rasch measures for the total, grammar, vocabulary, fluency, and pronunciation categories.

The second research question, "Which affective variables, Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak

English, and L2 Speaking Self-Confidence, predict the students' self-assessment of L2 oral performance?," was first answered by calculating the Pearson correlations between the students' self-assessment measures and each affective variable (i.e., Self-Esteem, L2 Speaking Anxiety, and L2 WTC). Second, I investigated the degree to which the affective variables predicted self-assessment bias by testing a hypothesized structural model. Figure 8 shows the hypothesized structural model for the self-assessment bias size of L2 speaking. Self-assessment of L2 speaking was hypothesized to be influenced positively by Self-Esteem (AlFallay, 2004) (H₁). L2 Speaking Anxiety is hypothesized to predict Self-Assessment Bias Size negatively (H₂) (MacIntyre, Noels, & Clément, 1997). Attitude Toward Learning to Speak English (ALSE), L2 Speaking Motivational Intensity (MI), and Desire to Learn to Speak English (DLSE) are components of L2 Speaking Motivation based on Gardner's (1985) Attitude/Motivation Test Battery (AMTB), which is hypothesized to predict Self-Assessment Bias Size positively (H₃, H₄, and H₅, respectively) (Masgoret & Gardner, 2003). Because L2 Speaking Self-Confidence is defined as high self-evaluation of L2 communicative competence and a low level of anxiety (Noels & Clément, 1996), Self-Assessment Bias Size was hypothesized to be influenced positively by L2 Speaking Self-Confidence (H₆). Finally, Self-Assessment Bias Size is hypothesized to predict L2 WTC (MacIntyre, Babin, & Clément, 1999; MacIntyre & Doucette, 2010). Before running the structural model, I conducted a confirmatory factor analysis and Rasch analysis to determine the dimensionality of the questionnaire items.

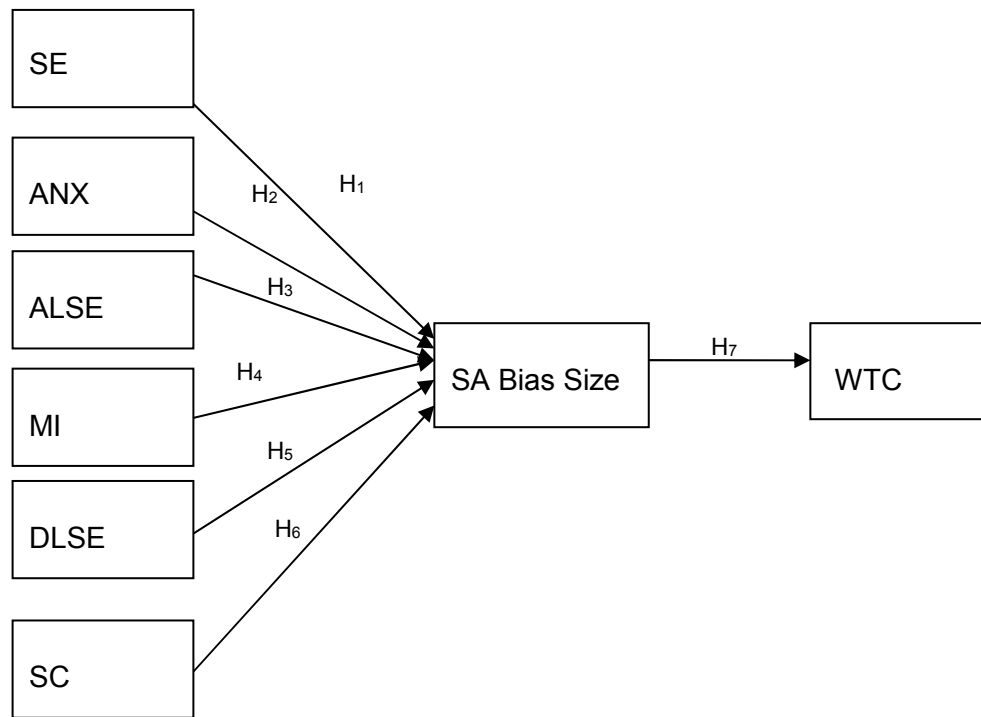


Figure 8. Hypothesized structural model for the self-assessment bias size of L2 speaking. SE = Self-Esteem; ANX = L2 Speaking Anxiety; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence; SA = Self-Assessment; WTC = L2 Willingness to Communicate.

The third research question asked, “What are the characteristics of the students who could assess their own L2 oral performance accurately?” This question was answered by selecting the participants whose ratings are similar to the teachers and their speaking measures and affective variable measures were compared with those of other participants using *t*-tests in order to examine if there are some distinctive characteristics of those who could assess their own L2 performance accurately.

The fourth research question is, “To what degree do the seven affective variables affect the self-assessment of high and low proficiency students differently?” This question was answered by dividing the students into high and low proficiency groups and conducting stepwise multiple regressions for each group using the seven affective variables (i.e., Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence) as predictors and L2 speaking self-assessment bias measures (i.e., Grammar, Vocabulary, Fluency, and Pronunciation) as dependent variables.

CHAPTER 4

PRELIMINARY ANALYSIS

The purpose of this chapter is to examine the validity and reliability of the questionnaire through SPSS factor analysis and a Rasch analysis. The results of the factor analysis are reported first and they are followed by the results of the Rasch analysis.

Factor Analysis Results

First, missing values were checked for the 396 students who responded to the 81-item questionnaire. Six persons (ID numbers 40, 63, 79, 89, 246, and 269) were deleted because they had six or more missing responses; the new *N*-size was 390. Four persons (ID numbers 174, 276, 283, and 350) had 1 or 2 missing responses. The missing values were replaced with the mean for the item, which is one of the ways to deal with missing data for a large sample suggested by Field (2005).

Second, univariate outliers were checked using SPSS. Persons 66 and 146 were found to have a raw score of self-assessment greater than 3.29 standard deviations above or below the mean, which was significant at $p < .001$. Outliers were not included in SPSS factor analysis, but they were included in the Rasch analysis because the Rasch analysis transforms raw scores to Rasch measures.

Third, the questionnaire data were checked for multivariate outliers using Mahalanobis distance in SPSS Regression. Person 46 was identified as a

multivariate outlier. Therefore, person 46 was deleted from the factor analysis, but included in the Rasch analysis because Mahalanobis distances are calculated using raw mean scores, but the Rasch analysis does not use raw scores; Rasch measures are used instead.

Fourth, normality was checked with SPSS Descriptives. No items showed skewness values greater than 2.00. Out of the distributions of the 81 items, 56 items had skewness and/or kurtosis statistics that exceeded two times the absolute value of the standard error of skewness or kurtosis even after adjusting for outliers. However, with large samples, the visual appearance of the distribution is more important than the above statistics (Tabachnick & Fidell, 2007); therefore, histograms of all items were examined. It was found that items ANX3, ALSE5, ALSE6, and DLSE10 displayed a ceiling effect, and that item MI10 showed a floor effect. However, all items were retained at this point because the data were examined using a Rasch PCA of item residuals before the main analysis.

Fifth, linearity was checked using scatterplots for all variables; no problems were found.

Finally, multicollinearity and singularity were checked by examining the Pearson product-moment correlation matrix of all items. One pair of items from different parts of the questionnaire (items MI4 and DLSE6) correlated above .70, which is a potentially problematic level of collinearity (Tabachnick & Fidell, 2007); these items were checked when conducting the factor analysis.

The KMO statistic was .91, which is greater than the .70 criterion for factorability. Bartlett's Test ($X^2 = 16505.223$, $df = 2211$, $p < .001$) confirmed the factorability of the correlation matrix. A factor analysis using principal axis factoring extraction with varimax rotation was conducted for 81 items with a sample of 387 participants after data screening. A seven-factor solution was selected as the items were designed to measure seven constructs. Item loadings above .40 were used in determining factors (Field, 2005).

As hypothesized, the ten Self-Esteem items (SE1 to SE10) loaded together on Factor 1. Nine L2 Speaking Anxiety items (ANX1 to ANX9) loaded above .40 on Factor 2, while item ANX10 loaded at .29 on Factor 2. Because of the weak loading, item ANX10 was deleted from the analysis. Twelve WTC items (WCT1 to WTC12) loaded on Factor 3 above the .40 criterion.

Items SC1 to SC16 were hypothesized to load together and act as measures of L2 Speaking Self-Confidence. All of these items loaded on Factor 4 with loadings ranging from .52 (SC14) to .78 (SC6). Thus, these items loaded as hypothesized.

The Motivation items measuring ALSE, MI, and DLSE did not form an independent constructs. Items ALSE1 to ALSE13 were hypothesized to form a single factor, Attitude Toward Learning to Speak English. However, only items ALSE1 (*I enjoy speaking English*), ALSE2 (*I enjoy speaking English more than reading English*), ALSE3 (*I enjoy speaking English more than writing English*), ALSE8 (*I look forward to my English speaking classes*), and ALSE9 (*I enjoy English speaking classes more than other classes*) loaded on Factor 5.

Items MI1 to MI10 were hypothesized to load together and act as measures of L2 Speaking Motivational Intensity. However, only items MI2 (*I think I try to speak English more than other students*), MI6 (*I spend a long time studying English*), MI7 (*I study English more than most of my classmates*), MI8 (*I often think about how I can improve my English speaking skills*), MI9 (*I work hard to become an excellent speaker of English*), and MI10 (*I study English speaking on my own through radio or TV language program*) loaded on Factor 6. MI5 (*I make an effort not to make grammatical mistakes when I speak English*) loaded at .35, but it loaded most strongly on Factor 6, so it was retained. These items are related to motivational intensity. Other three MI items loaded on different factors; item MI1 on Factor 7, and items MI3 and MI4 loaded on Factor 3, which were deleted.

Items DLSE1 to DLSE10 were hypothesized to load together and act as measures of Desire to Learn to Speak English. Items DLSE3 (*I really want to learn to speak English better*), DLSE4 (*Learning to speak English is more important than learning to read English*), DLSE5 (*Learning to speak English is more important than learning to write English*), DLSE7 (*I plan to keep improving my English speaking skills even after graduating from college*), DLSE8 (*I believe that Japanese students should be taught to speak English at school*), DLSE10 (*I wish I could speak English perfectly*), ALSE4 (*I am very interested in learning to speak English*), ALSE6 (*I would enjoy talking with native English teachers*), ALSE11 (*I think that English is the most important subject in school*), ALSE12 (*Speaking English is important for engineers*), ALSE13 (*I consider speaking English to be*

one of the most important skills to learn in school), and MI1 (*I concentrate well when I speak English*) loaded on Factor 7.

Other motivation items did not load on the hypothesized motivation factors. Items ALSE5 and ALSE7 did not load strongly on any factor; thus, these items were deleted. Items DLSE2 and DLSE9 were complex; DLSE2 loaded on Factor 1 at .57 and Factor 3 at .41, and item DLSE9 loaded on Factor 1 at .53 and Factor 3 at .51. Therefore, items DLSE2 and DLSE9 were deleted. Five items hypothesized to load on motivation factors instead loaded on WTC (Factor 3); items ALSE10 (.55), MI3 (.46), MI4 (.49), DLSE1 (.44) and DLSE6 (.57). Because these items did not load on the intended constructs, they were deleted.

After deleting ten items (ANX10, ALSE5, ALSE7, ALSE10, MI3, MI4, DLSE1, DLSE2, DLSE6, and DLSE9), a factor analysis using principal axis factoring extraction with varimax rotation was conducted again. This time, items SC14 and SC15 formed a separate factor, so they were deleted. After deleting items SC14 and SC15, a factor analysis was conducted again, and two items, DLSE4 and DLSE5, formed a separate factor, so they were deleted. A factor analysis using principal axis factoring extraction with varimax rotation was conducted once more, and the results were satisfactory. The final results are shown in Table 2.

Table 2. Pattern Matrix of the Questionnaire Items

Item	Factor							h^2
	1	2	3	4	5	6	7	
SE3	.75	-.04	.07	.07	.05	-.03	.08	.58
SE8	.74	-.05	.08	.04	-.02	.03	.02	.56
SE5	.73	-.06	.11	.12	.07	.13	-.11	.60
SE4	.72	-.12	.11	.04	.06	.05	.08	.56
SE1	.66	-.13	.08	.13	.08	-.08	.11	.50
SE2	.65	-.04	-.00	.08	-.01	-.09	.19	.48
SE10	.63	-.05	.09	-.01	.00	.06	.01	.41
SE7	.62	-.09	.16	.06	.10	-.01	-.05	.43
SE6	.58	-.06	.04	.04	-.01	.09	-.24	.41
SE9	.53	.06	.00	.08	-.01	.03	.17	.32
ANX7	-.09	.78	-.16	-.18	-.11	.04	.03	.69
ANX9	-.09	.77	-.18	-.18	-.11	.01	.07	.68
ANX4	-.09	.67	-.07	-.14	-.05	-.06	-.01	.48
ANX8	-.01	.66	-.09	-.12	-.14	.10	.08	.50
ANX3	-.04	.60	-.14	-.16	.05	-.05	.11	.42
ANX6	-.12	.58	.10	-.08	.01	.13	.10	.40
ANX5	-.10	.57	.02	-.27	-.07	-.20	.10	.47
ANX1	-.07	.51	.12	-.07	-.07	.08	.18	.32
ANX2	-.01	.41	.02	-.10	-.04	-.15	-.04	.20
WTC6	.07	-.02	.80	.11	.07	.13	.11	.70
WTC4	.07	-.01	.73	.17	.14	.11	.18	.63
WTC5	.01	-.03	.71	.13	.06	.03	.29	.62
WTC3	.10	-.00	.70	.14	.08	.05	.32	.63
WTC8	.13	-.05	.70	.18	.12	.22	.05	.60
WTC9	.09	-.03	.70	.22	.01	.12	.20	.61
WTC7	.05	-.08	.68	.22	.10	.19	.09	.56
WTC10	.09	-.03	.68	.18	.03	.20	.13	.57
WTC2	.12	-.02	.66	.10	.04	-.05	.32	.56
WTC11	.16	-.07	.63	.17	.22	.26	.03	.57
WTC12	.22	-.10	.54	.16	.14	.08	.14	.41
WTC1	-.03	-.01	.52	.14	.10	.08	.21	.35
SC3	.03	-.17	.14	.79	.09	.16	.06	.71
SC6	.11	-.11	.12	.79	.03	.21	.04	.72
SC5	.00	-.12	.11	.76	.07	.15	.03	.63
SC10	.03	-.10	.15	.76	.05	.06	.13	.63
SC2	.00	-.23	.12	.73	.09	.22	.01	.66
SC1	.08	-.18	.04	.70	.07	.23	-.05	.59
SC8	.07	-.03	.13	.70	.15	.05	-.02	.53
SC4	.03	-.19	.11	.69	.10	.17	.00	.57
SC12	.11	-.10	.21	.69	.06	.04	.08	.55
SC13	.13	-.03	.17	.68	.03	-.05	.15	.53
SC7	.00	-.07	.15	.67	.01	.29	-.07	.57
SC9	.08	-.14	.16	.67	.11	.13	.08	.53
SC16	.13	-.18	.28	.57	.01	.13	.00	.47
SC11	.11	-.08	.11	.56	.04	-.07	.18	.38
ALSE9	.10	-.12	.20	.18	.64	.16	.26	.60
ALSE2	.10	-.19	.27	.18	.62	-.14	.11	.56

Item	Factor							h ²
	1	2	3	4	5	6	7	
ALSE3	.11	-.20	.19	.18	.56	-.12	.29	.53
ALSE8	.01	-.13	.31	.24	.54	.16	.21	.53
ALSE1	.20	-.14	.33	.28	.47	.03	.33	.57
MI6	-.07	.02	.16	.18	-.08	.70	.10	.58
MI9	-.01	-.03	.24	.24	.01	.70	.12	.62
MI7	.10	.04	.16	.24	-.07	.69	.02	.58
MI8	-.02	-.03	.28	.18	.13	.56	.27	.51
MI2	.01	-.14	.25	.35	.13	.52	.13	.51
MI10	.12	-.05	.18	.21	.10	.43	-.05	.29
MI5	.07	.24	-.06	.12	-.20	.35	.27	.31
ALSE13	.03	.09	.08	.00	.12	.08	.61	.40
DLSE3	.10	-.04	.33	.16	.14	.08	.60	.53
ALSE4	.08	-.08	.39	.14	.28	.02	.55	.57
ALSE6	.00	.03	.38	.03	.08	-.08	.54	.45
DLSE7	.08	-.10	.34	.17	.09	.25	.53	.51
MI1	.09	.12	.17	.13	.03	.21	.52	.39
DLSE8	.01	.10	.23	.08	.13	.06	.50	.34
DLSE10	.01	.12	.10	.00	-.05	-.05	.50	.28
ALSE12	.03	.10	.11	-.08	.06	-.02	.49	.27
ALSE11	-.02	.09	.12	.03	.06	.21	.43	.26
% of Variance	12.17	10.80	7.23	6.22	6.12	5.02	3.25	50.80

Note. N = 387. Boldface indicates factor loadings higher than .40.

Factor 1 obtained loadings greater than $\pm .40$ from ten items (SE1 to SE10) hypothesized to measure Self-Esteem; therefore, it was named Self-Esteem. The item with the strongest loading on this factor was item SE3 (*I feel useful most of the time*).

Factor 2 consisted of nine items (ANX1 to ANX9), designed to measure L2 Speaking Anxiety; thus, this factor was named L2 Speaking Anxiety. Item ANX7 (*I feel nervous speaking English*) loaded most strongly on this factor.

Factor 3 consisted of 12 items (WTC1 to WTC12), hypothesized to measure WTC; thus, this factor was named L2 Willingness to Communicate. Item WTC6 (*I*

would be willing to interview a teacher in English) had the strongest loading on this factor.

Factor 4 obtained loadings greater than $\pm .40$ from 14 items (SC1 to SC13 and SC16), which were designed to measure L2 Speaking Self-Confidence; thus, this factor was named L2 Speaking Self-Confidence. The two items with the strongest loadings were items SC3 (*I can hold a 5-minute conversation in English with pair partner*) and SC6 (*I can express my opinion about common topics*).

Factor 5 consisted of five ALSE items, ALSE1, ALSE2, ALSE3, ALSE8, and ALSE9; thus, this factor was named Attitude Toward Learning to Speak English. Item ALSE 9 (*I enjoy English speaking classes more than other classes*) had the strongest loading on this factor.

Factor 6 obtained loadings greater than $\pm .40$ from six items, MI2, MI6, MI7, MI8, MI9, and MI10; thus, this factor was named L2 Speaking Motivational Intensity. In addition to these six items, item MI5 (*I make an effort not to make grammatical mistakes when I speak English*) loaded on this factor at .35, which is below the .40 criterion. Despite the low loading, this item was included in this factor because this item is an original Motivational Intensity item and it loaded on this factor more strongly than on the other factors. Item MI6 (*I spend a long time studying English*) had the strongest loading on this factor.

Factor 7 consisted of ten items; five ALSE items (ALSE4, ALSE6, ALSE11, ALSE12, and ALSE13), four DLSE items (DLSE3, DLSE7, DLSE8, and DLSE10) and one MI item (MI1). This factor was named Desire to Learn to Speak

English/Attitude Toward Learning to Speak English (DLSE/ALSE). Item ALSE13 (*I consider speaking English to be one of the most important skills to learn in school*) loaded on this factor most strongly.

Rasch Analysis

As a second step to examine the construct validity of the questionnaire, the Rasch measurement model was employed for each affective variable. The rating scale criteria are as follows (Linacre, 2007):

1. At least 10 responses are made for each category.
2. Outfit MNSQ should be below 2.00.
3. No threshold should be disordered.

The category separation criteria are shown in Table 3 (Wolfe & Smith, 2007). The category separation is determined following the criteria. However, person and item reliabilities and separations are prioritized when deciding on the rating scale.

Table 3. *Category Separation Criteria (Wolfe & Smith, 2007)*

Category	Minimum separation (logits)
3	1.40
4	1.10
5	.81
6	.59
7	.41
8	.25
9	.15

Self-Esteem

The six-point rating scale was examined using the criteria set (See Table 4). At least 10 responses were made for each category, Outfit MNSQ was below 2.00, and no thresholds were disordered. All categories were separated by at least .59 logits, which is the criterion for six rating-scale categories, and the shape of the probability curves was peaked for each category (Figure 9). Thus, the six-point rating scale was working acceptably well.

Table 4. *Six-Point Rating Scale Functioning for the Self-Esteem Items*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	283 (7)	1.35	1.44	None	(-3.84)
2 Disagree	517 (13)	.84	.82	-2.48	-2.22
3 Slightly disagree	1250 (32)	.84	.84	-1.75	-.80
4 Slightly agree	1208 (31)	.83	.84	-.01	.76
5 Agree	462 (12)	.99	.99	1.66	2.22
6 Strongly agree	179 (5)	1.34	1.35	2.58	(3.91)

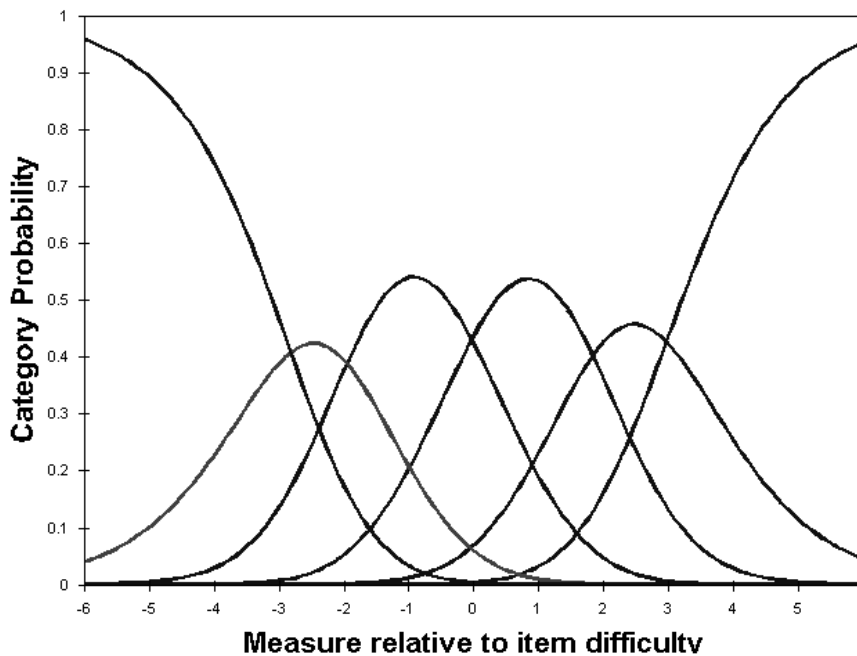


Figure 9. Category probability curves for the six-point rating scale for Self-Esteem.

As shown in Table 5, Item SE 7 (*I view myself positively rather than negatively*) misfit the Rasch model. The Infit MNSQ was 1.76 and the Outfit MNSQ was 1.76. The expression, “positively rather than negatively” might have confused some participants. Therefore, item SE7 was deleted from the SE construct.

Table 5. Rasch Item Statistics for the Self-Esteem Items

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
SE6	1.25	.06	1.19	2.6	1.19	2.5	.63
SE5	.76	.06	.90	-1.4	.90	-1.4	.76
SE10	.44	.06	.94	-.8	.95	-.7	.69
SE8	.23	.06	.75	-3.7	.75	-3.8	.75
SE3	-.12	.06	.66	-5.3	.66	-5.3	.75
SE7	-.15	.06	1.76	8.6	1.76	8.5	.69
SE4	-.17	.06	.81	-2.8	.81	-2.8	.75
SE9	-.41	.06	1.20	2.7	1.31	4.0	.60
SE2	-.89	.06	.93	-1.0	.94	-.8	.67
SE1	-.93	.06	.79	-3.2	.81	-2.8	.70

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .85.

The other nine items were entered into the Rasch analysis and they all fit the Rasch model acceptably well (Table 6) with Infit MNSQ indices between .69 and 1.31. Pt-measure correlations were also high and positive in each case. Thus, these nine items were used for the Self-Esteem construct.

Table 6. Rasch Item Statistics for the Self-Esteem Items Excluding Item SE7

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
SE6	1.38	.07	1.31	4.0	1.33	3.9	.63
SE5	.83	.07	1.02	.2	1.01	.2	.76
SE10	.49	.07	1.06	.9	1.07	.9	.69
SE8	.24	.07	.85	-2.1	.85	-2.1	.75
SE3	-.16	.07	.69	-4.8	.69	-4.7	.77
SE4	-.21	.07	.90	-1.4	.90	-1.4	.76
SE9	-.48	.07	1.29	3.7	1.39	4.8	.62
SE2	-1.02	.07	.98	-.3	.98	-.2	.70
SE1	-1.06	.07	.84	-2.3	.86	-1.9	.72

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .86.

The item-person map for the Self-Esteem items (Figure 10) shows the logit scale on the far left. Persons are placed along the logit scale according to their ability estimates, and items are placed according to their endorsement difficulty levels. Two of the easiest to endorse items, items SE1 and SE9, concerned viewing oneself as having equal abilities with others. Two items in the middle of the scale, items SE3 and SE4, were more related to the person's positive value in a social context. The most difficult to endorse items, items SE5 and SE6, concerned having a strongly positive image of oneself; these statements might not have been in accord with traditional Japanese notions of humility.

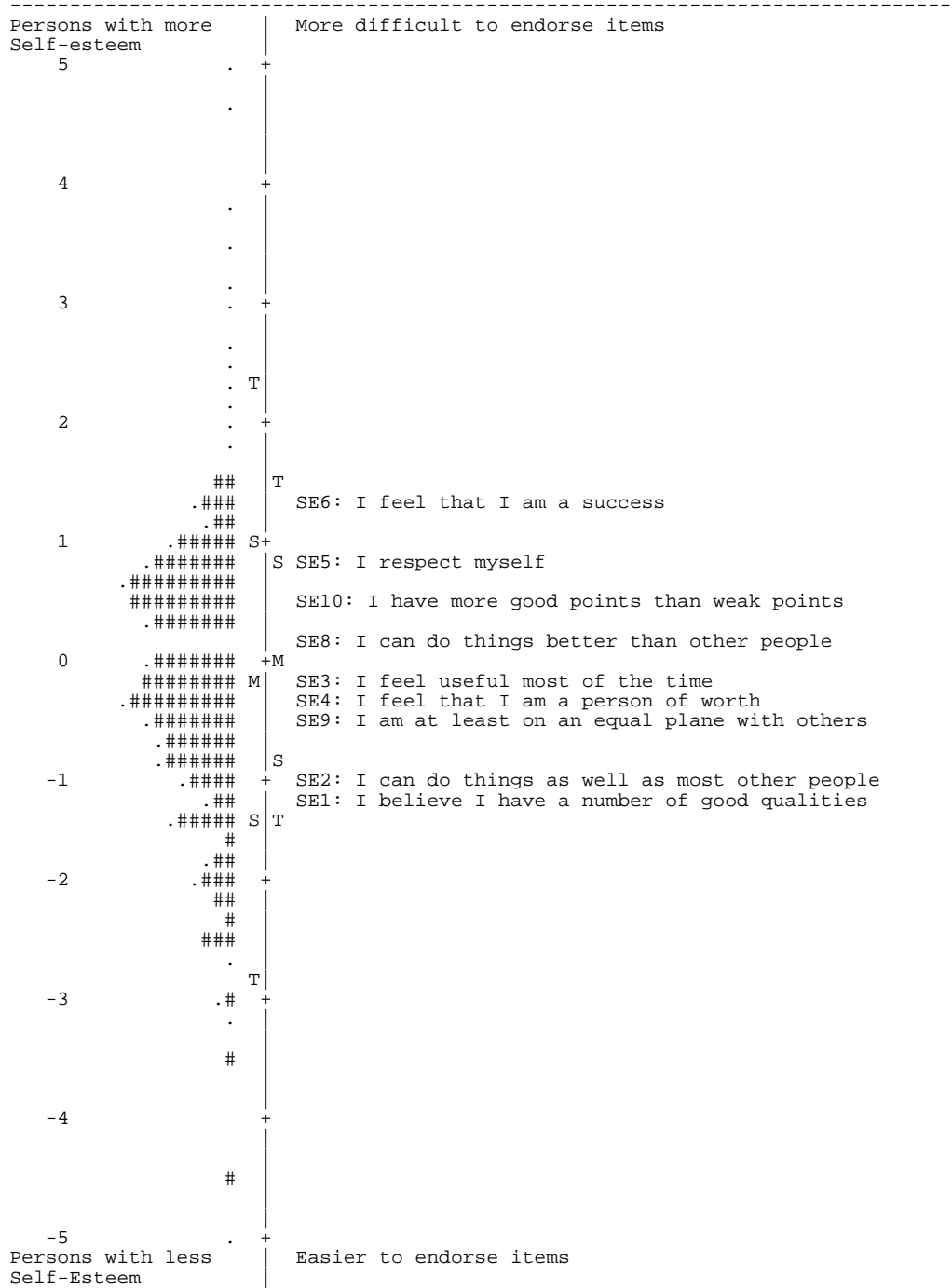


Figure 10. Item-person map for Self-Esteem. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

The mean person ability estimate was $-.20$, which was slightly lower than the mean item difficulty estimate, which is set by convention at 0.00 . This indicates that the items were slightly difficult for the participants to endorse overall. The Rasch person reliability (separation) estimate was $.86$ (2.47), and the Rasch item reliability (separation) estimate was $.99$ (11.06).

The nine SE items were next examined through the Rasch PCA of item residuals using the criteria for unidimensionality (Table 1). The Rasch model accounted for 56.6% of the variance (eigenvalue = 11.8), which exceeded the 50% criterion suggested by Linacre. The unexplained variance explained by the first residual was 8.9% (eigenvalue = 1.9), which met the criterion. These findings indicated that the items formed a fundamentally unidimensional construct. As shown in Table 7, Items SE2 and SE3 had high positive loadings above $.40$, and items SE5 and SE6 had high negative loadings above $-.40$. The items with positive loadings were related to “doing things well,” while the items with negative loadings were concerned with “a person’s worth.”

Table 7. Rasch Principal Component Analysis for the Self-Esteem Items

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
SE2	.76	-1.02	.98	.98
SE3	.62	-.16	.69	.69
SE8	.31	.24	.85	.85
SE1	.16	-1.06	.84	.86
SE6	-.60	1.38	1.31	1.33
SE5	-.58	.83	1.02	1.01
SE10	-.24	.49	1.06	1.07
SE4	-.18	-.21	.90	.90
SE9	-.01	-.48	1.29	1.39

L2 Speaking Anxiety

The six-point rating scale for L2 Speaking Anxiety was examined using the criteria (See Table 8). The six-point rating scale did not work well, as the thresholds for categories 2 and 3 were not greater than .59 logits, so they were not separated sufficiently.

Table 8. *Six-Point Rating Scale Functioning for L2 Speaking Anxiety*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	229 (6)	1.56	1.61	None	(-2.62)
2 Disagree	320 (8)	.94	1.01	-1.09	-1.33
3 Slightly disagree	730 (19)	.85	.85	-1.06	-.48
4 Slightly agree	964 (25)	.87	.85	-.07	.33
5 Agree	941 (24)	.86	.86	.68	1.33
6 Strongly agree	714 (18)	1.04	1.05	1.53	(2.88)

Therefore, categories 1 and 2 were combined to make a five-point rating scale (Table 9), in which categories 1 and 2 were separated by .68 logits, which is less than the .81 criterion for a five-point scale.

Table 9. *Five-Point Rating Scale Functioning for L2 Speaking Anxiety*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	549 (14)	1.29	1.33	None	(-2.55)
2 Slightly disagree	730 (19)	.86	.91	-1.17	-1.05
3 Slightly agree	964 (25)	.87	.94	-.49	-.03
4 Agree	941 (24)	.90	.93	.36	1.03
5 Strongly agree	714 (18)	.98	1.05	1.29	(2.62)

Next, categories 1 and 2 were combined to make the four-point rating scale. The results are shown in Table 10. Categories 2 and 3 were separated by only .66 logits, which is smaller than 1.10, the criterion for the four-point rating scale.

Table 10. *Four-Point Rating Scale Functioning for L2 Speaking Anxiety*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	1279 (33)	1.13	1.18	None	(-2.20)
2 Slightly agree	964 (25)	.84	1.01	-.82	-.65
3 Agree	941 (24)	.95	1.04	-.16	.61
4 Strongly agree	714 (18)	.94	1.08	.97	(2.27)

A three-point rating scale was created by combining categories 2 and 3. As shown in Table 11, the resulting scale separated well. However, the results of the PCA showed that the Rasch model accounted for 44.3% of the variance, which failed to meet the 50% criterion suggested by Linacre. The unidimensionality of the construct is the top priority of the analysis, so I decided not to utilize three-point scale.

Table 11. *Three-Point Rating Scale Functioning for L2 Speaking Anxiety*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	1279 (33)	1.07	1.07	None	(-2.75)
2 Agree	1905 (49)	.94	.94	-1.63	.00
3 Strongly agree	714 (18)	.93	.93	1.63	(2.75)

Because categories 2 and 3 were separated more widely in the five-point rating scale than the four-point rating scale, the five-point rating scale was selected for use with the L2 Speaking Anxiety items. The probability curve for the five-point scale is shown in Figure 11. The shape of the probability curves was peaked for each category.

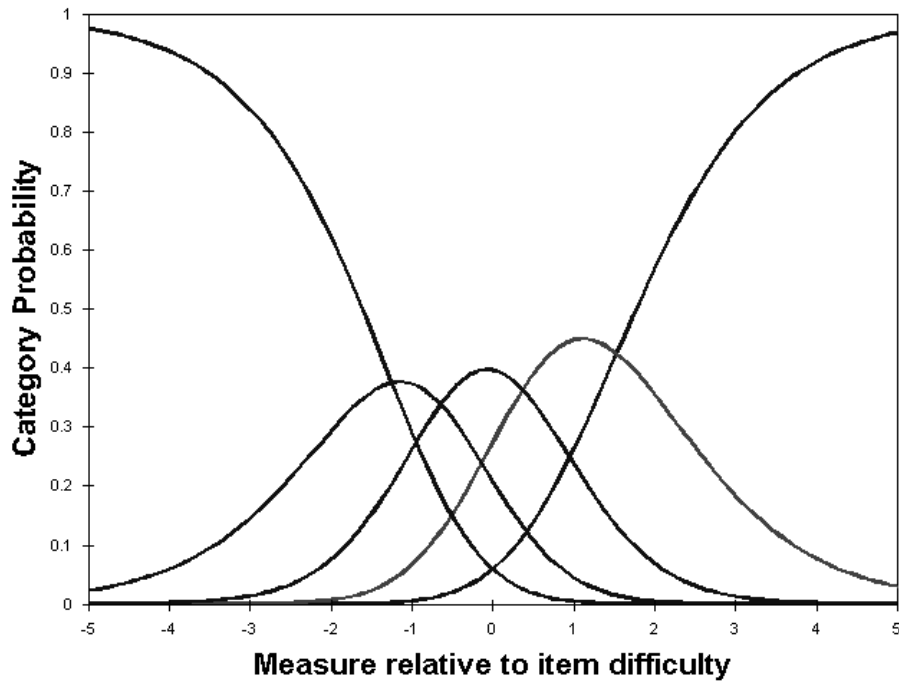


Figure 11. Category probability curves for the five-point rating scale for L2 Speaking Anxiety.

All the L2 Speaking Anxiety items except item ANX10 (*I feel self-conscious when I speak English*) met the .50-1.50 Infit and Outfit MNSQ fit criterion (See Table12). Item ANX10 displayed an Infit MNSQ statistic of 1.50, and an Outfit MNSQ statistic of 1.75. In the factor analysis item ANX10 loaded at .29 on this construct, suggesting that it is a weak measure of L2 Speaking Anxiety. Probably the word “self-consciousness” was difficult for the participants to understand. Therefore, I decided to delete item ANX10.

Table 12. *Rasch Item Statistics for the L2 Speaking Anxiety Items*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ANX10	1.42	.06	1.50	5.8	1.75	7.1	.43
ANX6	.41	.05	1.08	1.1	1.05	.7	.65
ANX1	.31	.05	1.21	2.9	1.20	2.7	.59
ANX7	.11	.05	.64	-6.1	.62	-6.2	.76
ANX4	-.13	.06	.97	-.4	.96	-.5	.67
ANX9	-.21	.06	.66	-5.8	.64	-5.8	.75
ANX2	-.22	.06	1.26	3.6	1.48	5.8	.52
ANX8	-.30	.06	.83	-2.6	.80	-2.9	.69
ANX3	-.67	.06	1.14	1.9	1.14	1.8	.63
ANX5	-.73	.06	.86	-2.1	.81	-2.5	.65

Note. N= 390. Rasch item reliability = .99. Rasch person reliability = .81.

After item ANX10 was deleted, the remaining nine items were entered into analysis. Item ANX2 (*I feel that other students speak English better than me*) showed an Outfit MNSQ statistic of 1.56 (Table 13). Because the participants were placed in the English classes that matched their L2 proficiency levels, most classmates had similar L2 abilities. That is why this item did not work well for the participants.

Table 13. *Rasch Item Statistics for the L2 Speaking Anxiety Items Excluding Item ANX10*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ANX6	.60	.06	1.13	1.9	1.12	1.6	.66
ANX1	.50	.06	1.28	3.8	1.29	3.8	.60
ANX7	.28	.06	.67	-5.5	.65	-5.6	.76
ANX4	.03	.06	1.01	.2	1.02	.3	.68
ANX2	-.06	.06	1.33	4.3	1.56	6.6	.54
ANX9	-.06	.06	.69	-5.0	.68	-5.0	.75
ANX8	-.15	.06	.90	-1.5	.89	-1.6	.68
ANX3	-.54	.06	1.21	2.8	1.24	2.9	.62
ANX5	-.61	.06	.87	-1.9	.82	-2.4	.66

Note. N= 390. Rasch item reliability = .98. Rasch person reliability = .81.

After item ANX2 was deleted, the remaining eight items were analyzed. All the items met the criterion and person reliability improved to .82 (Table 14).

Therefore, these eight items were used for the L2 Speaking Anxiety construct.

Table 14. *Rasch Item Statistics for the L2 Speaking Anxiety Items Excluding Items ANX10 and ANX2*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ANX6	.63	.06	1.16	2.2	1.13	1.8	.68
ANX1	.52	.06	1.39	5.1	1.47	5.6	.60
ANX7	.29	.06	.63	-6.2	.61	-6.3	.79
ANX4	.02	.06	1.07	.9	1.10	1.3	.68
ANX9	-.07	.06	.68	-5.2	.66	-5.2	.76
ANX8	-.16	.06	.91	-1.3	.89	-1.5	.70
ANX3	-.58	.06	1.23	3.0	1.25	3.0	.64
ANX5	-.65	.06	1.00	.0	1.00	.1	.65

Note. N= 390. Rasch item reliability = .98. Rasch person reliability = .82.

The item-person map for the L2 Speaking Anxiety construct is shown in Figure 12. The easiest to endorse item was ANX5, which indicates that most of the participants feel a general lack of confidence in their speaking ability. Two items in the middle of the item hierarchy were ANX4, ANX8, and ANX9; therefore, many participants feel nervousness speaking English. The most difficult to endorse items were ANX1 and ANX6, indicating that relatively few participants report the fear of negative evaluation by others. Therefore, most participants lack confidence in their speaking ability, so many feel nervous speaking English. However, only those with greater speaking anxiety are concerned with being evaluated negatively by other students and teachers.

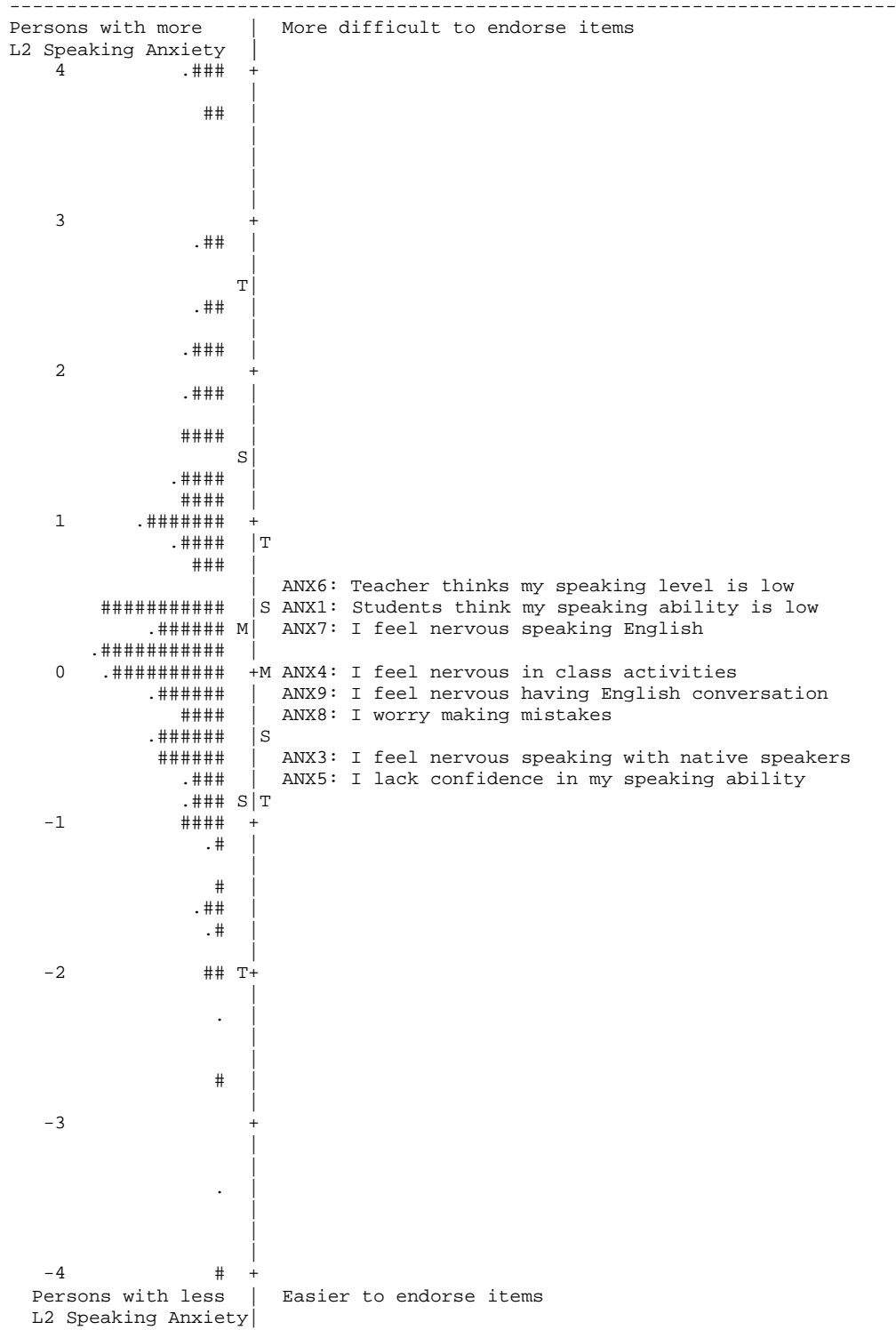


Figure 12. Item-person map for L2 Speaking Anxiety. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

The mean person ability estimate was .36, which was slightly higher than the mean item difficulty estimate. Therefore, the items were slightly easy for the participants to endorse overall. The Rasch person reliability (separation) estimate was .82 (2.13), and the Rasch item reliability (separation) estimate was .98 (7.00).

The Rasch model accounted for 51.5% of the variance (eigenvalue = 8.5), which exceeded the 50% criterion. The unexplained variance explained by the first residual was 14.4% (eigenvalue = 2.4). Given the amount of variance accounted for by the Rasch model and the small eigenvalue of the first contrast, the items measuring L2 Speaking Anxiety appeared to form a fundamentally unidimensional construct.

Table 15 shows the Rasch PCA for all L2 Speaking Anxiety items. Items ANX1 and ANX6 had high positive loadings, and items ANX4, ANX7, and ANX9 loaded negatively between -.37 and -.73. The positively loading items were related to “worry about other people’s evaluation,” while the negatively loading items were related to “feeling nervous speaking English,”

Table 15. Rasch Principal Component Analysis for the L2 Speaking Anxiety Items

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
ANX1	.78	.52	1.39	1.47
ANX6	.66	.63	1.16	1.13
ANX5	.28	-.65	1.00	1.00
ANX9	-.73	-.07	.68	.66
ANX7	-.67	.29	.63	.61
ANX4	-.37	.02	1.07	1.10
ANX3	-.34	-.58	1.23	1.25
ANX8	-.11	-.16	.91	.89

L2 Willingness to Communicate

The six-point rating scale for L2 WTC was examined using the criteria (See Table 16). At least 10 responses were made for each category, Outfit MNSQ was below 2.00, no thresholds were disordered, each threshold was separated by at least .59 logits, and the shape of the probability curve was peaked for each category (Figure 13).

Table 16. *Six-Point Rating Scale Functioning for L2WTC*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	465 (10)	1.13	1.10	None	(-3.72)
2 Disagree	792 (17)	.87	.86	-2.41	-2.00
3 Slightly disagree	1441 (31)	.83	.84	-1.38	-.59
4 Slightly agree	1136 (24)	.94	.99	.18	.73
5 Agree	531 (11)	1.01	1.02	1.43	1.96
6 Strongly agree	315 (7)	1.27	1.30	2.18	(3.54)

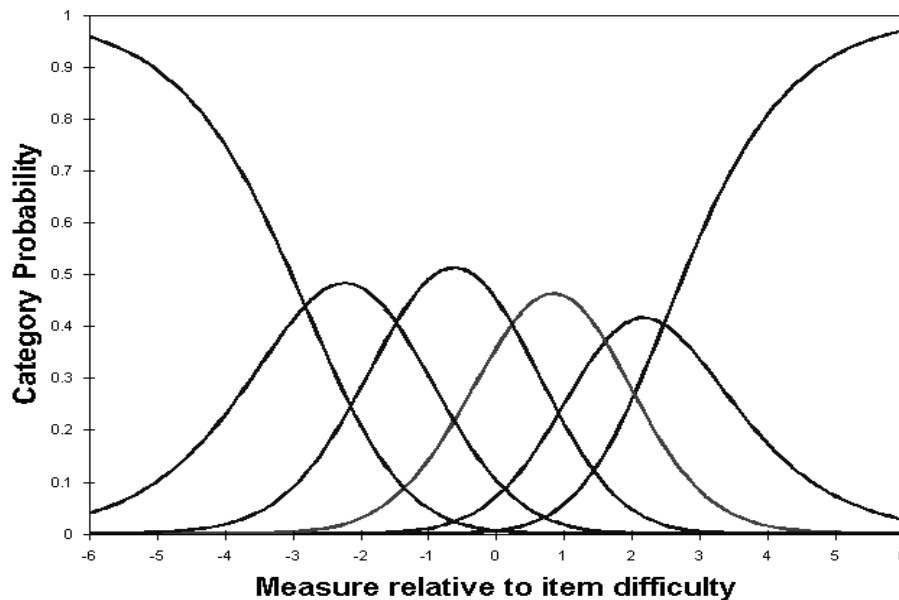


Figure 13. Category probability curves for the six-point rating scale for L2 WTC.

As shown in Table 17, items WTC12 (*I would be willing to guide a group of three Canadian students around Tokyo*) misfit the Rasch model according to the .50-1.50 MNSQ criterion. Guiding a tour requires more than speaking English, such as planning where to go, so this is probably why this item misfit the L2 WTC construct.

Table 17. *Rasch Item Statistics for the L2 WTC Items*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
WTC11	.94	.06	1.03	.4	1.02	.3	.70
WTC8	.55	.06	.98	-.2	.96	-.5	.74
WTC10	.34	.06	.88	-1.7	.88	-1.6	.74
WTC12	.27	.06	1.53	6.4	1.59	6.9	.64
WTC4	.25	.06	.78	-3.2	.76	-3.6	.77
WTC6	.23	.06	.66	-5.3	.65	-5.4	.80
WTC7	.23	.06	1.00	.0	.98	-.3	.74
WTC9	-.19	.06	.79	-3.2	.81	-2.8	.77
WTC5	-.46	.06	.96	-.5	.94	-.8	.75
WTC3	-.50	.06	.91	-1.2	.90	-1.4	.76
WTC1	-.60	.06	1.34	4.4	1.47	5.7	.63
WTC2	-1.08	.06	1.04	.6	1.04	.5	.73

Note. N=390. Rasch item reliability = .99. Rasch person reliability = .90.

Item WTC12 was removed and the remaining WTC are shown in Table 18. Now item WTC1 (*I will be willing to answer a question from my teacher in English class*) misfit the model; Item WTC1 had an Outfit MNSQ statistic of 1.52, which was above the 1.50 criterion. Answering questions requires the learners to tell the answers, so this item might be different from other items in which learners speak English more freely. Thus, item WTC1 was removed and the remaining items were analyzed.

Table 18. *Rasch Item Statistics for the L2 WTC Items Excluding Item WTC12*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
WTC11	1.02	.06	1.13	1.8	1.13	1.7	.70
WTC8	.62	.06	1.07	1.0	1.04	.6	.74
WTC10	.39	.06	.92	-1.0	.93	-1.0	.75
WTC4	.30	.06	.80	-3.0	.77	-3.4	.78
WTC6	.27	.06	.67	-5.1	.66	-5.3	.81
WTC7	.27	.06	1.04	.6	1.01	.2	.75
WTC9	-.17	.06	.80	-2.9	.83	-2.4	.78
WTC5	-.46	.06	1.00	.0	.98	-.2	.76
WTC3	-.50	.06	.96	-.6	.94	-.8	.77
WTC1	-.61	.06	1.39	4.9	1.52	6.2	.64
WTC2	-1.13	.06	1.14	1.9	1.13	1.7	.73

Note. N=390. Rasch item reliability = .99. Rasch person reliability = .90.

Table 19 shows that the remaining items met the Infit and Outfit MNSQ fit criterion. Therefore, these 10 items were used for the L2 WTC construct.

Table 19. *Rasch Item Statistics for the L2 WTC Items Excluding Items WTC12 and WTC1*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
WTC11	1.02	.07	1.17	2.2	1.18	2.3	.74
WTC8	.59	.07	1.10	1.4	1.07	.9	.75
WTC10	.35	.06	.95	-.7	.95	-.7	.76
WTC4	.25	.06	.84	-2.3	.81	-2.7	.76
WTC6	.22	.06	.70	-4.5	.70	-4.6	.76
WTC7	.22	.06	1.06	.9	1.04	.5	.76
WTC9	-.25	.06	.83	-2.4	.86	-2.0	.77
WTC5	-.55	.06	1.05	.7	1.04	.6	.77
WTC3	-.60	.06	1.04	.6	1.03	.4	.77
WTC2	-1.25	.06	1.20	2.6	1.22	2.7	.77

Note. N=390. Rasch item reliability = .99. Rasch person reliability = .90.

The item-person map for L2 WTC is shown in Figure 14. The easiest to endorse item, item WTC2, was related to willingness to speak English with an international student. Thus, casually talking with someone approximately the same age seems easy for the participants. Other relatively easy to endorse items, items

WTC3 and WTC5, concerned a willingness to speak English with a teacher.

Difficult to endorse items concerned more formal situations (items WTC8 and WTC10) and speaking in front of relatively large groups of people (items WTC8 and WTC11).

The mean person ability estimate was $-.32$, which was slightly lower than the mean item difficulty estimate. This indicates that the items were slightly difficult for the participants to endorse overall. The Rasch person reliability (separation) estimate was $.90$ (3.01), and the Rasch item reliability (separation) estimate was $.99$ (9.40).

The Rasch model accounted for 60.5% of the variance (eigenvalue = 15.3), and the unexplained variance explained by the first residual contrast was 10.3% (eigenvalue = 2.6). Because the Rasch model accounted for more than 50% of the variance and because of the small eigenvalue in the first residual contrast, it was concluded that the items formed a fundamentally unidimensional construct.

Table 20 shows the Rasch PCA for the L2 WTC items. Items WTC2, WTC3, WTC4, and WTC5 had high positive loadings above $.40$, and items WTC7, WTC8, WTC9, WTC10, and WTC11 loaded negatively above $-.39$. The items with positive loadings were related to willingness to speak English with native English speakers, such as a teacher, while the items with negative loadings were related to willingness to speak English with classmates.

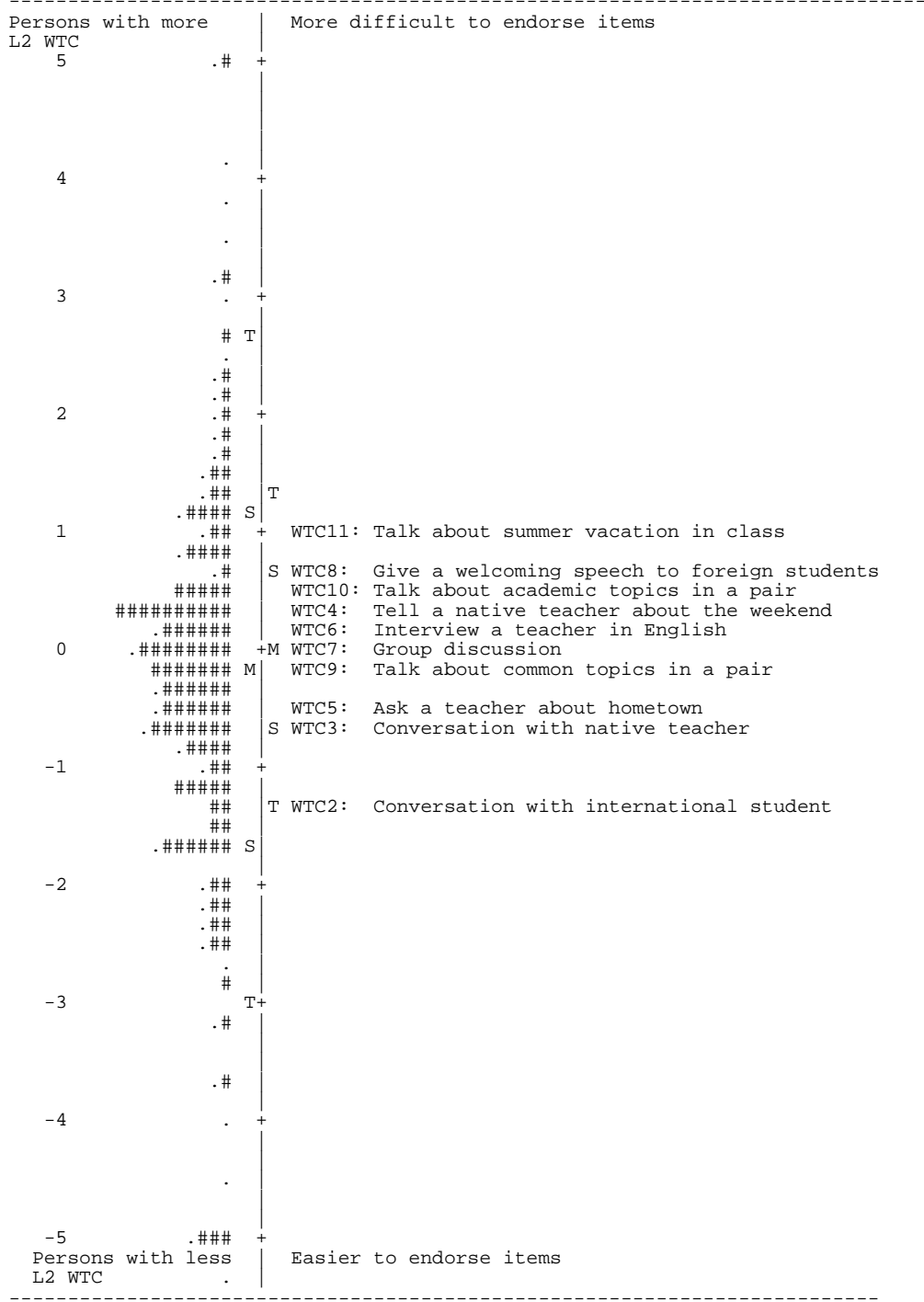


Figure 14. Item-person map for L2 WTC. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

Table 20. *Rasch Principal Component Analysis for the L2 WTC Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
WTC3	.68	-.60	1.04	1.03
WTC5	.63	-.55	1.05	1.04
WTC4	.48	.25	.84	.81
WTC2	.43	-1.25	1.20	1.22
WTC6	.26	.22	.70	.70
WTC10	-.64	.35	.95	.95
WTC7	-.51	.22	1.06	1.04
WTC8	-.49	.59	1.10	1.07
WTC9	-.47	-.25	.83	.86
WTC11	-.39	1.02	1.17	1.18

L2 Speaking Self-Confidence

The six-point rating scale for L2 Speaking Self-Confidence items was examined. As shown in Table 21, at least 10 responses were made for each category, no thresholds were disordered, and each threshold was separated by at least .59 logits. However, the Outfit MNSQ statistic for category 6 was above 2.00.

Table 21. *Six-Point Rating Scale Functioning for L2 Speaking Self-Confidence*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	1085 (17)	.90	.93	None	(-4.00)
2 Disagree	1434 (23)	.85	.79	-2.71	-2.25
3 Slightly disagree	1694 (27)	.82	.85	-1.53	-.85
4 Slightly agree	1469 (24)	.87	.89	-.31	.76
5 Agree	400 (6)	1.15	1.15	1.90	2.35
6 Strongly agree	158 (3)	2.34	2.26	2.66	(4.01)

Because the Outfit MNSQ statistics of category 6 was above 2.00, categories 5 and 6 were combined to make a five-point rating scale (Table 22), in which each threshold was separated by at least .81 logits. Therefore, the five-point scale was used for the L2 Speaking Self-Confidence construct. The probability curve for the

five-point scale is shown in Figure 15. The shape of the probability curves was peaked for each category.

Table 22. *Five-Point Rating Scale Functioning for L2 Speaking Self-Confidence*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	1085 (17)	.92	.96	None	(-3.45)
2 Disagree	1434 (23)	.88	.84	-2.18	-1.65
3 Slightly disagree	1694 (27)	.85	.94	-.90	-.17
4 Slightly agree	1469 (24)	.91	.92	.44	1.61
5 Agree	558 (9)	1.59	1.46	2.63	(3.81)

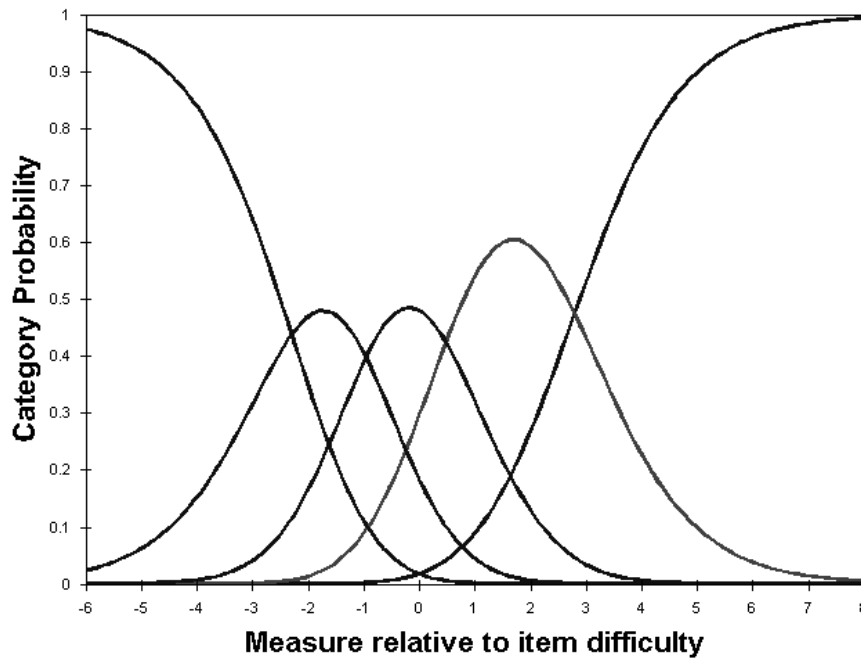


Figure 15. Category probability curves for the five-point rating scale for L2 Speaking Self-Confidence.

Table 23 shows that item SC11 (*I can tell the time to a foreigner in English*) misfit the Rasch model with an Infit MNSQ statistic of 1.61 and Outfit MNSQ

statistic of 1.55. Tellingtime is different from other items that require learners to communicate with others or produce longer sentences.

Table 23. Rasch Item Statistics for the L2 Speaking Self-Confidence Items

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
SC16	1.16	.07	1.04	.6	1.03	.4	.67
SC4	1.12	.07	.90	-1.4	.88	-1.4	.72
SC7	.96	.07	.94	-.8	.97	-.3	.69
SC1	.79	.07	.91	-1.3	1.07	.9	.73
SC2	.67	.07	.88	-1.8	.89	-1.4	.75
SC8	.56	.07	1.05	.8	1.05	.6	.71
SC12	.33	.07	.95	-.6	.93	-.9	.74
SC3	.19	.07	.76	-3.7	.74	-3.9	.80
SC9	.17	.07	1.03	.5	1.05	.7	.73
SC6	.09	.07	.68	-5.1	.67	-5.1	.80
SC13	.04	.07	1.09	1.3	1.06	.8	.72
SC10	.03	.07	.83	-2.5	.83	-2.4	.78
SC5	-.32	.07	.86	-2.1	.82	-2.6	.78
SC11	-1.38	.07	1.61	7.2	1.55	6.3	.65
SC15	-2.18	.08	1.21	2.7	1.16	1.9	.62
SC14	-2.23	.08	1.19	2.4	1.12	1.4	.63

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .92.

Table 24 shows that after removing SC11, all of the L2 Speaking Self-Confidence items met the .50-1.50 Infit and Outfit MNSQ fit criterion. Therefore, these 15 items were used to measure the L2 Speaking Self-Confidence construct.

Table 24. *Rasch Item Statistics for the L2 Speaking Self-Confidence Items Excluding Item SC11*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
SC16	1.12	.07	1.08	1.1	1.07	.8	.67
SC4	1.08	.07	.90	-1.4	.86	-1.6	.74
SC7	.91	.07	.96	-.5	.99	-.1	.70
SC1	.73	.07	.92	-1.2	1.06	.8	.74
SC2	.61	.07	.89	-1.7	.88	-1.6	.77
SC8	.49	.07	1.11	1.6	1.10	1.3	.71
SC12	.25	.07	1.02	.3	.99	-.1	.74
SC3	.10	.07	.78	-3.4	.75	-3.7	.81
SC9	.08	.07	1.07	1.0	1.08	1.1	.74
SC6	-.01	.07	.69	-5.0	.67	-5.0	.82
SC13	-.05	.07	1.20	2.7	1.17	2.3	.71
SC10	-.06	.07	.88	-1.7	.88	-1.6	.78
SC5	-.44	.07	.87	-1.9	.83	-2.4	.79
SC15	-2.38	.08	1.28	3.5	1.23	2.5	.61
SC14	-2.43	.08	1.27	3.4	1.20	2.2	.62

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .93.

The item-person map for L2 Speaking Self-Confidence is shown in Figure 16. The easiest to endorse items, items SC14 and SC15, concerned common greetings and a self-introduction in English. Many items in the middle of the scale, items SC3, SC5, SC6, SC9, SC10, and SC13, were related to easy classroom speaking activities. Relatively difficult to endorse items concerned longer stretches of speaking (items SC2 and SC4) or speaking about an academic topic (item SC7). The most difficult to endorse item, item SC16, concerned speaking English on the telephone in a real situation outside classroom. The mean person ability estimate was $-.51$, which was slightly lower than the mean item difficulty estimate. This indicates that the items were slightly difficult for the participants to endorse overall. The Rasch person reliability (separation) estimate was $.93$ (3.56), and the Rasch item reliability (separation) estimate was $.99$ (14.05).

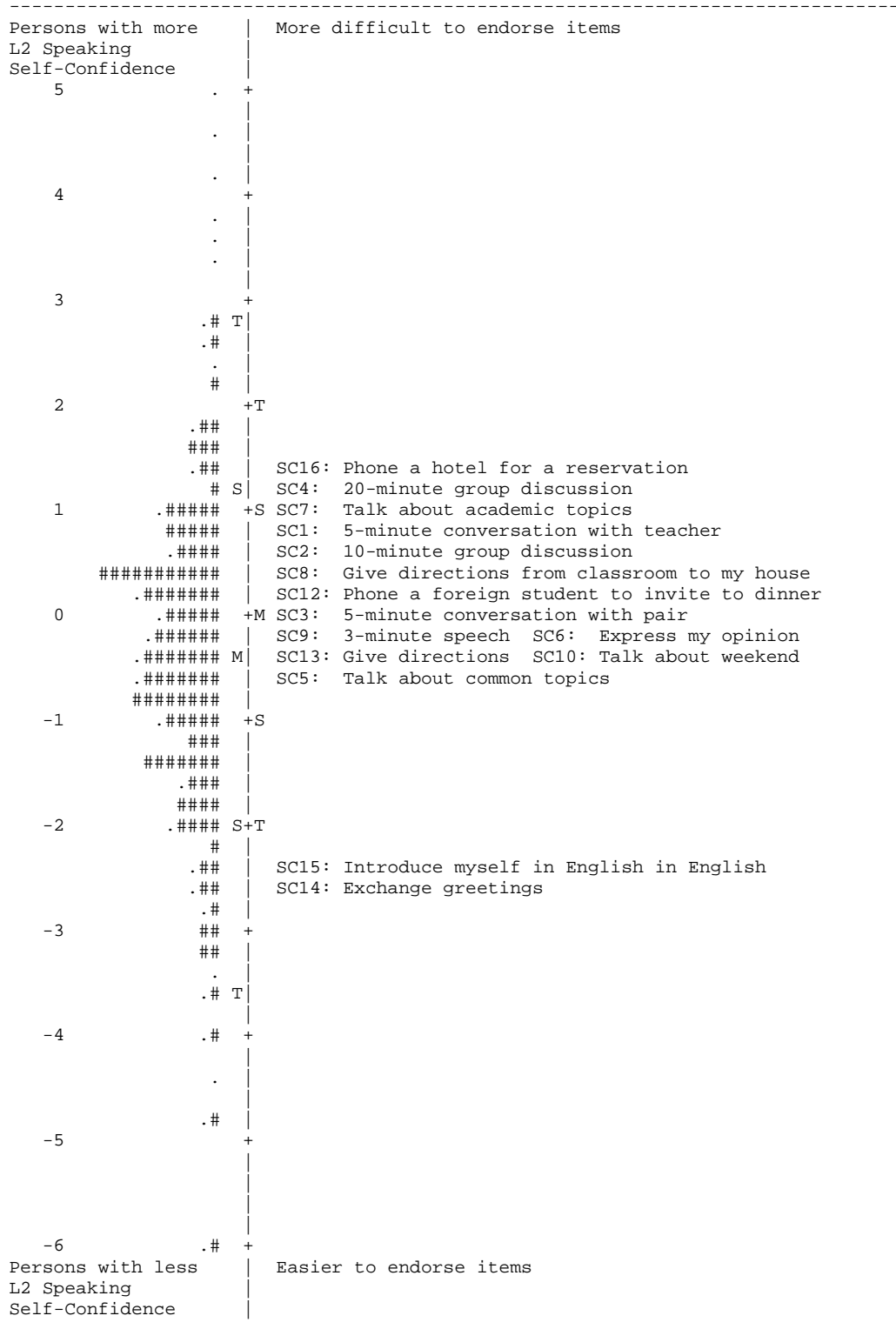


Figure 16. Item-person map for L2 Speaking Self-Confidence. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

Fifteen L2 Speaking Self-Confidence items were examined using a Rasch PCA of item residuals. The Rasch model accounted for 64.7% of the variance (eigenvalue = 27.5), which exceeded the 50% criterion. The unexplained variance explained by the first residual contrast was 7.0% (eigenvalue = 3.0). These findings indicated that the SC items formed a fundamentally unidimensional construct.

As shown in Table 25, items SC12, SC13, SC14, and SC15 had high positive loadings between .39 and .66, and items SC1, SC2, SC3, and SC4 had high negative loadings above -.40. The positively loading items were related to self-confidence in using English in particular situations, while the negatively loading items were concerned with self-confidence in engaging in English conversation with pair or group.

Table 25. *Rasch Principal Component Analysis for the L2 Speaking Self-Confidence Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
SC14	.66	-2.43	1.27	1.20
SC15	.65	-2.38	1.28	1.23
SC13	.54	-.05	1.20	1.17
SC12	.39	.25	1.02	.99
SC16	.30	1.12	1.08	1.07
SC9	.08	.08	1.07	1.08
SC8	.08	.49	1.11	1.10
SC10	.08	-.06	.88	.88
SC2	-.66	.61	.89	.88
SC4	-.64	1.08	.90	.86
SC3	-.53	.10	.78	.75
SC1	-.40	.73	.92	1.06
SC6	-.33	-.01	.69	.67
SC7	-.29	.91	.96	.99
SC5	-.28	-.44	.87	.83

L2 Speaking Motivation Items

In the SPSS Factor Analysis, the items measuring MI, ALSE, and DLSE did not form the hypothesized constructs; thus, all 33 items were input at the same time to allow them to interact with each other in the Rasch PCA of item residuals analysis (Table 26). The Rasch model explained 48.7% of the variance (eigenvalue = 31.3); the first contrast explained 6.9% of the variance (eigenvalue = 4.4).

Because the amount of the variance accounted for the Rasch model did not exceed 50% and the eigenvalue of the first contrast was very high, I concluded that the data were not fundamentally unidimensional. Table 26 shows that eight Motivational Intensity items had high positive loadings between .37 and .73; therefore, the ten items hypothesized to measure Motivational Intensity (MI1 to MI10) were analyzed first.

L2 Speaking Motivational Intensity

The six-point rating scale for L2 Speaking Motivational Intensity was examined, and the results are shown in Table 27. At least 10 responses were made for each category, the Outfit MNSQ statistic was below 2.00, no thresholds were disordered, and each threshold was separated by at least .59 logits. The category probability curve is shown in Figure 17. The shape of the probability curves was peaked for each category.

Table 26. *Rasch Principal Component Analysis for the MI, ALSE, and DLSE Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
MI9	.73	.92	.93	.91
MI7	.71	1.16	1.15	1.15
MI6	.70	.96	1.12	1.11
MI2	.56	.93	.78	.78
MI8	.54	.52	.93	.92
MI10	.49	1.58	1.31	1.26
MI4	.37	.72	.87	.86
MI5	.33	.08	1.58	1.69
MI3	.26	.33	.55	.56
DLSE6	.17	.58	.77	.76
DLSE1	.11	.12	1.04	1.05
DLSE7	.07	-.12	.78	.79
MI1	.01	-.20	.83	.83
ALSE11	.01	-.10	1.40	1.40
ALSE3	-.43	-.15	1.09	1.11
ALSE4	-.42	-.53	.74	.73
DLSE4	-.41	-.46	1.12	1.14
ALSE2	-.40	.04	1.24	1.26
DLSE5	-.40	-.47	1.15	1.15
ALSE7	-.38	-.21	1.09	1.11
ALSE6	-.38	-1.29	.99	.91
DLSE8	-.29	-.55	.96	.95
DLSE10	-.27	-1.41	1.71	1.61
DLSE3	-.27	-.53	.73	.72
ALSE1	-.26	-.09	.68	.68
ALSE5	-.26	-1.39	1.63	2.02
ALSE12	-.26	-.78	1.14	1.25
ALSE13	-.23	-.66	1.08	1.08
ALSE9	-.17	.20	.89	.90
ALSE10	-.11	.25	.61	.61
DLSE9	-.10	-.06	.71	.71
ALSE8	-.06	.33	.84	.86
DLSE2	-.05	.27	.77	.77

Table 27. *Six-Point Rating Scale Functioning for L2 Speaking Motivational Intensity*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	471 (12)	1.00	1.02	None	(-3.46)
2 Disagree	745 (19)	.83	.87	-2.09	-1.87
3 Slightly disagree	1256 (32)	.82	.82	-1.39	-.59
4 Slightly agree	903 (23)	.95	.97	.12	.64
5 Agree	376 (10)	.98	.98	1.31	1.84
6 Strongly agree	148 (4)	1.64	1.64	2.05	(3.41)

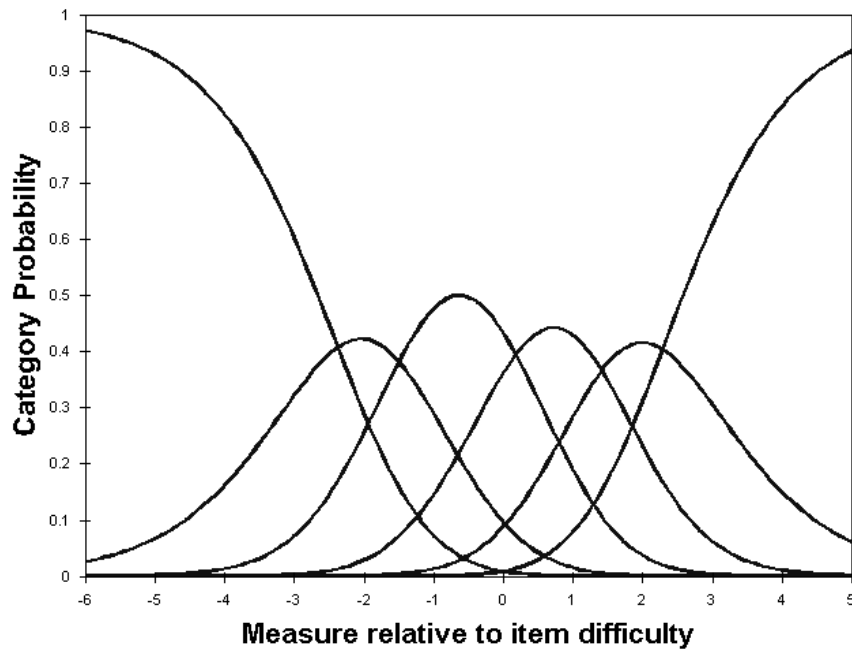


Figure 17. Category probability curves for the six-point rating scale for L2 Speaking Motivational Intensity.

Table 28 shows that item MI5 (*I make an effort not to make grammatical mistakes when I speak English*) misfit the Rasch model. Infit MNSQ was 1.62 and Outfit MNSQ was 1.69.

Table 28. Rasch Item Statistics for the L2 Speaking Motivational Intensity Items

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
MI10	1.16	.06	1.28	3.7	1.33	4.0	.53
MI7	.61	.06	.82	-2.8	.81	-2.8	.71
MI6	.34	.06	.86	-2.1	.85	-2.1	.69
MI2	.31	.06	.68	-5.1	.68	-5.0	.72
MI9	.29	.06	.68	-5.2	.67	-5.2	.76
MI4	.03	.06	1.11	1.5	1.10	1.4	.69
MI8	-.25	.06	.90	-1.5	.89	-1.6	.74
MI3	-.49	.06	.81	-2.9	.81	-2.8	.70
MI5	-.82	.06	1.62	7.5	1.69	8.2	.47
MI1	-1.19	.06	1.23	3.1	1.24	3.3	.51

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .84.

Item MI5 was also problematic in the factor analysis, as it loaded weakly (.35) on the Motivational Intensity construct. Thus, this item was deleted. An inspection of item MI5's content suggested that the negative wording "not to make grammatical mistakes" might have caused a problem. After deleting item MI5, all the items met the Rasch Infit and Outfit MNSQ fit criterion (Table 29).

Table 29. *Rasch Item Statistics for the L2 Speaking Motivational Intensity Items Excluding Item MI5*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt-measure correlation
MI10	1.19	.06	1.37	4.8	1.40	4.7	.55
MI7	.57	.06	.93	-.9	.91	-1.2	.71
MI6	.28	.06	.96	-.5	.95	-.6	.70
MI2	.24	.06	.71	-4.4	.70	-4.6	.75
MI9	.22	.06	.72	-4.3	.72	-4.3	.77
MI4	-.06	.06	1.13	1.7	1.11	1.5	.73
MI8	-.37	.06	.94	-.8	.93	-.9	.75
MI3	-.65	.06	.85	-2.2	.86	-2.1	.72
MI1	-1.41	.06	1.40	5.2	1.42	5.4	.50

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .84.

The item-person map for L2 Speaking Motivational Intensity is shown in Figure 18. The easiest to endorse item, item MI1, concerned the ability to concentrate while speaking English. Items in the middle of the scale concerned the effort invested in improving speaking English (items MI8 and MI9) or an attempt to speak frequently (items MI3 and MI4). More difficult to endorse items, items MI2 and MI7, concerned positive comparisons with other students. The most difficult to endorse item, MI10, concerned studying speaking English using TV or

radio programs. This approach to studying is probably used relatively little because of the advent of the Internet.

The mean person ability estimate was $-.62$, which was slightly lower than the mean item difficulty estimate. Therefore, the items were slightly difficult for the participants to endorse overall. The Rasch person reliability (separation) estimate was $.84(2.27)$, and the Rasch item reliability (separation) estimate was $.99(10.93)$.

The Rasch model accounted for 55.8% of the variance (eigenvalue = 11.4), which exceeded the 50% criterion. The unexplained variance explained by the first residual was 10.3% (eigenvalue = 2.1). The amount of variance accounted for by the Rasch model and the small eigenvalue for the first contrast indicated that the MI items formed a fundamentally unidimensional construct.

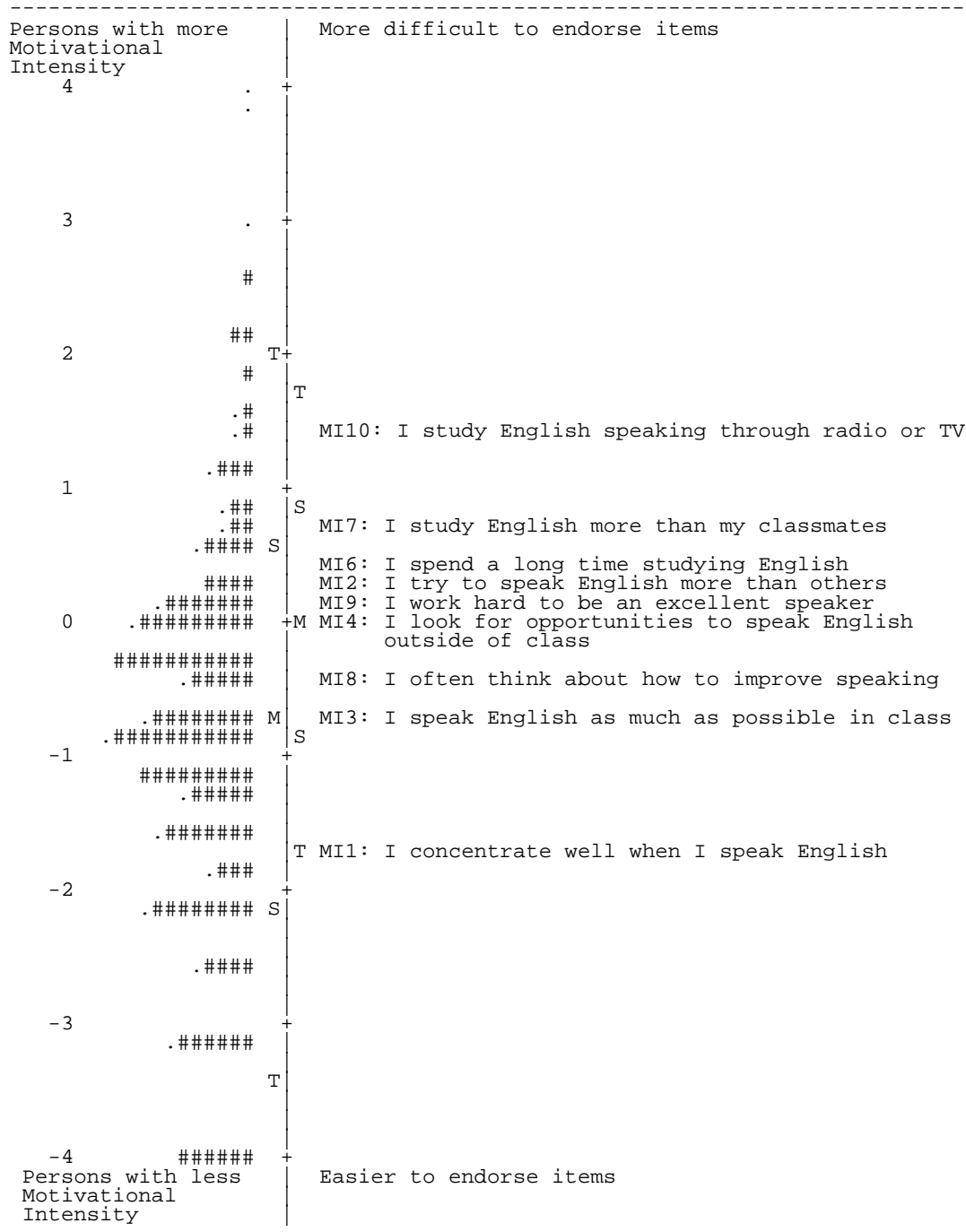


Figure 18. Item-person map for L2 Speaking Motivational Intensity. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

Table 30 shows the Rasch PCA of item residual results for the MI items. Items MI6, MI7, and MI9 loaded positively above .40, and items MI3 and MI4 loaded

negatively above -.40. The positively loading items concerned “studying English hard,” while the negatively loading items concerned “trying to speak English more often.”

Table 30. *Rasch Principal Component Analysis for the L2 Speaking Motivational Intensity Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
MI6	.76	.28	.96	.95
MI7	.65	.57	.93	.91
MI9	.49	.22	.72	.72
MI10	.09	1.19	1.37	1.40
MI4	-.58	-.06	1.13	1.11
MI3	-.57	-.65	.85	.86
MI1	-.35	-1.41	1.40	1.42
MI2	-.25	.24	.71	.70
MI8	-.05	-.37	.94	.93

Attitude Toward Learning to Speak English and Desire to Learn to Speak English

The remaining motivational ALSE and DLSE items were analyzed with the Rasch PCA of item residuals analysis (Table 31). Items ALSE1, ALSE2, ALSE8, ALSE9, ALSE10, and DLSE2 had high positive loadings between .38 and .65, while items ALSE 5, ALSE11, ALSE12, ALSE13, DLSE4, DLSE5, and DLSE10 had high negative loadings between -.37 and -.52. The negative loading ALSE items (items ALSE5, ALSE11, ALSE12, and ALSE13) were problematic items in the factor analysis; ALSE5 failed to load any factors above .40, and ALSE11, ALSE12, and ALSE13 loaded above .40 on the DLSE factor. Therefore, the dimensionality of the ALSE items was examined closely by inspecting those problematic items.

Table 31. *Rasch Principal Component Analysis for the ALSE and DLSE Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
ALSE10	.65	.58	.60	.60
ALSE1	.47	.23	.67	.73
ALSE8	.47	.67	.85	.90
DLSE2	.42	.61	.77	.76
ALSE9	.38	.54	.88	.90
DLSE6	.35	.92	.85	.85
ALSE2	.29	.36	1.18	1.18
ALSE4	.26	-.24	.68	.65
ALSE3	.24	.17	1.04	1.04
DLSE1	.23	.45	1.13	1.13
DLSE9	.18	.26	.72	.76
DLSE7	.15	.20	.87	.90
ALSE7	.13	.11	1.04	1.04
ALSE12	-.52	-.50	1.13	1.29
ALSE5	-.52	-1.16	1.65	2.11
DLSE5	-.46	-.17	1.12	1.13
ALSE13	-.44	-.37	1.10	1.09
DLSE10	-.43	-1.17	1.69	1.60
DLSE4	-.42	-.17	1.09	1.10
ALSE11	-.37	.22	1.50	1.69
DLSE8	-.29	-.26	.95	1.00
ALSE6	-.20	-1.04	.95	.85
DLSE3	-.02	-.23	.72	.71

Attitude Toward Learning to Speak English

The six-point rating scale for the ALSE items was examined (Table 32). At least 10 responses were made for each category, Outfit MNSQ was below the 2.00 criterion, and no thresholds were disordered. However, categories 2 and 3 were separated by only .01 logits and categories 5 and 6 were separated by only .45 logits, which were smaller than .59, the criterion for six rating-scale categories.

Table 32. *Six-Point Rating Scale Functioning for Attitude Toward Learning to Speak English*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	152 (3)	1.21	1.41	None	(-2.87)
2 Disagree	327 (6)	1.14	1.34	-1.33	-1.53
3 Slightly disagree	1084 (21)	.82	.83	-1.34	-.54
4 Slightly agree	1481 (29)	.90	.91	-.02	.47
5 Agree	1070 (21)	.81	.84	1.12	1.53
6 Strongly agree	954 (19)	1.10	1.08	1.57	(2.99)

First, categories 1 and 2 were combined creating a five-point rating scale (Table 33); categories 2 and 3 were separated by 1.02 logits, but categories 4 and 5 were separated by only .48 logits, which was smaller than .81, the criterion for five rating-scale categories.

Table 33. *Five-Point Rating Scale Functioning for Attitude Toward Learning to Speak English*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	479 (9)	1.22	1.40	None	(-2.79)
2 Slightly disagree	1084 (21)	.85	.87	-1.48	-1.12
3 Slightly agree	1481 (29)	.93	.94	-.46	.06
4 Agree	1070 (21)	.83	.85	.73	1.15
5 Strongly agree	954 (19)	1.06	1.06	1.21	(2.63)

Next, categories 1 and 2 were combined to make a four-point rating scale (Table 34). The results are shown in Table 30. The categories 3 and 4 were separated by .63, which is lower than 1.10, which is the criterion for the four-point category.

Table 34. Four-Point Rating Scale for Attitude Toward Learning to Speak English

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	1563(31)	1.06	1.21	None	(-2.20)
2 Slightly agree	1084 (29)	.98	1.01	-.88	-.59
3 Agree	1070 (21)	.88	.84	.13	.63
4 Strongly agree	954 (19)	1.05	1.05	.76	(2.14)

Categories 2 and 3 were combined to create the three-point rating scale. The results were shown in Table 35. Categories 2 and 3 were separated by 3.12, which is much greater than 1.40, the criterion for the three-point rating scale. Therefore, the three-point rating scale was used for the Attitude Toward Learning to Speak English construct. Figure 19 shows category probability curves for the three-point rating scale. The shape of the probability curves was peaked for each category.

Table 35. Three-Point Rating Scale for Attitude Toward Learning to Speak English

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	1564 (31)	1.03	1.09	None	(-2.69)
2 Slightly disagree	2551 (50)	.94	.97	-1.56	.00
3 Agree	954 (19)	1.00	1.00	1.56	2.69

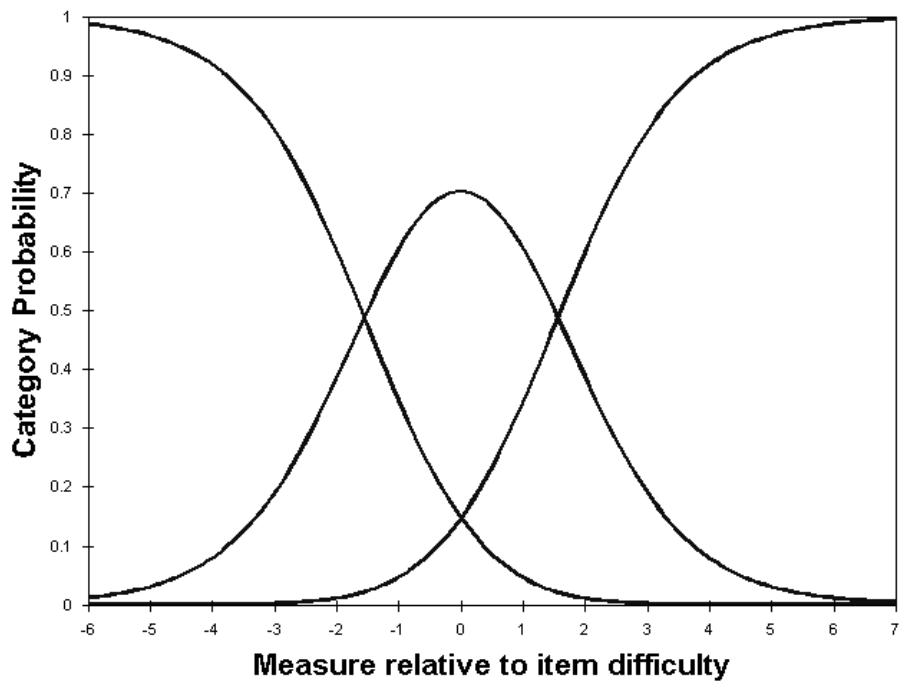


Figure 19. Category probability curves for the three-point rating scale for Attitude Toward Learning to Speak English.

As shown in Table 36, item ALSE11 (*I think that English is the most important subject in school*) misfit the Rasch model. Item ALSE11 was also problematic in the factor analysis because it loaded strongly with the ALSE and DLSE items; therefore, it was deleted.

Table 36. *Rasch Item Statistics for the Attitude Toward Learning to Speak English Items*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ALSE8	1.26	.10	.90	-1.6	.89	-1.3	.64
ALSE10	1.17	.10	.85	-2.4	.82	-2.3	.70
ALSE9	1.03	.10	.86	-2.3	.89	-1.4	.65
ALSE2	.68	.10	.99	-.1	1.07	.9	.61
ALSE1	.56	.10	.78	-3.7	.76	-3.5	.66
ALSE3	.31	.10	.89	-1.8	.87	-1.9	.65
ALSE11	.31	.10	1.51	6.8	1.62	7.4	.42
ALSE7	.19	.10	1.05	.8	1.04	.6	.60
ALSE4	-.45	.10	.75	-3.9	.73	-4.1	.70
ALSE13	-.71	.10	1.01	.1	1.00	.0	.54
ALSE12	-.87	.10	1.01	.2	1.13	1.7	.48
ALSE6	-1.67	.10	.96	-.5	.92	-.9	.58
ALSE5	-1.81	.10	1.33	4.6	1.55	5.4	.40

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .82.

After deleting item ALSE11, the remaining ALSE items were analyzed. The results are shown in Table 37. Item ALSE5 (*I admire Japanese students who can speak English well*), one of the problematic items in the Rasch PCA of item residuals analysis, misfit the Rasch model.

Table 37. *Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Item ALSE11*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ALSE8	1.37	.11	.93	-1.1	.91	-1.0	.64
ALSE10	1.26	.10	.86	-2.3	.82	-2.2	.71
ALSE9	1.12	.10	.88	-1.8	.91	-1.1	.66
ALSE2	.75	.10	.98	-.2	1.06	.8	.64
ALSE1	.62	.10	.78	-3.7	.77	-3.4	.68
ALSE3	.36	.10	.89	-1.7	.86	-2.0	.68
ALSE7	.23	.10	1.07	1.0	1.05	.6	.62
ALSE4	-.45	.10	.78	-3.4	.76	-3.5	.70
ALSE13	-.72	.10	1.14	2.0	1.14	1.8	.51
ALSE12	-.90	.10	1.13	1.8	1.34	4.0	.45
ALSE6	-1.75	.10	1.01	.2	.98	-.1	.58
ALSE5	-1.90	.10	1.45	6.0	1.76	6.5	.39

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .82.

Item ALSE5 was deleted because it did not load strongly on any factor in the factor analysis and it misfit the Rasch model. The notion of “admiring other students” might have differed conceptually from attitude toward learning to speak English. After deleting item ALSE5, the remaining ALSE items were analyzed. The results are shown in Table 38. Another problematic item, item ALSE12 (*Speaking English is important for engineers*), displayed poor fit, and in the factor analysis, that item loaded strongly with both the ALSE and DLSE items. For these reasons, it was deleted. An inspection of item ALSE12 suggested that although the participants were majoring in science and engineering, not all of them will become engineers in the future; thus, this item might have been difficult for some of them to respond to.

Table 38. *Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5 and ALSE11*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ALSE8	1.26	.11	.93	-1.0	.93	-.7	.66
ALSE10	1.16	.11	.86	-2.2	.80	-2.3	.72
ALSE9	1.01	.11	.89	-1.7	.90	-1.1	.67
ALSE2	.61	.10	1.00	.0	1.09	1.1	.66
ALSE1	.48	.10	.79	-3.4	.78	-3.0	.70
ALSE3	.20	.10	.88	-1.8	.85	-2.1	.70
ALSE7	.07	.10	1.10	1.4	1.08	1.1	.63
ALSE4	-.66	.10	.82	-2.7	.80	-2.8	.71
ALSE13	-.95	.10	1.26	3.4	1.30	3.5	.49
ALSE12	-1.14	.10	1.23	3.1	1.57	6.0	.44
ALSE6	-2.04	.10	1.13	1.9	1.14	1.4	.56

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .83.

After deleting item ALSE12, a Rasch analysis was conducted with the ten remaining ALSE items. As shown in Table 39, item ALSE13 (*I consider speaking English to be one of the most important skills to learn in school*), another problematic item, misfit the Rasch model. In the factor analysis, this item loaded strongly with ALSE and DLSE items. Because the participants were majoring in science and engineering, it might have been difficult for them to consider speaking English as the most important skill.

Table 39. Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5, ALSE11, and ALSE12

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ALSE8	1.20	.11	.94	-.9	.97	-.2	.67
ALSE10	1.09	.11	.87	-2.0	.80	-2.3	.73
ALSE9	.93	.11	.90	-1.5	.91	-1.0	.69
ALSE2	.52	.11	.98	-.2	1.04	.5	.68
ALSE1	.38	.11	.79	-3.4	.78	-2.9	.71
ALSE3	.09	.11	.89	-1.6	.86	-1.9	.71
ALSE7	-.05	.10	1.11	1.6	1.09	1.1	.65
ALSE4	-.81	.10	.83	-2.6	.80	-2.7	.72
ALSE13	-1.11	.10	1.42	5.4	1.66	6.7	.45
ALSE6	-2.24	.11	1.18	2.5	1.18	1.6	.56

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .83.

After deleting items ALSE5, ALSE11, ALSE12, and ALSE13, all problematic items in the Rasch PCA analysis, the remaining ALSE items met the .50-1.50 Infit and Outfit MNSQ fit criterion. The results are shown in Table 40.

Table 40. *Rasch Item Statistics for the Attitude Toward Learning to Speak English Items Excluding Items ALSE5, ALSE11, ALSE12, and ALSE13*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
ALSE8	1.16	.12	.98	-.2	1.10	.9	.68
ALSE10	1.04	.11	.90	-1.5	.84	-1.7	.74
ALSE9	.87	.11	.96	-.6	.95	-.5	.69
ALSE2	.42	.11	1.00	.0	1.04	.5	.70
ALSE1	.28	.11	.82	-2.7	.81	-2.4	.72
ALSE3	-.03	.11	.92	-1.2	.87	-1.7	.72
ALSE7	-.19	.11	1.16	2.2	1.15	1.9	.66
ALSE4	-1.00	.11	.90	-1.5	.86	-1.8	.72
ALSE6	-2.55	.11	1.30	4.0	1.47	3.5	.56

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .83.

The item-person map for Attitude Toward Learning to Speak English is shown in Figure 20. The easiest to endorse item, item ALSE6, was related to positive attitudes toward speaking English with native speakers of English. Many of the items in the middle of the scale, items ALSE7, ALSE3, and ALSE2, concerned enjoying speaking English more than listening, reading, and writing English. The more difficult to endorse items, items ALSE9 and ALSE10, concerned willingness to have more opportunities to speak English. The most difficult to endorse item, item ALSE8, concerned enjoying English speaking class.

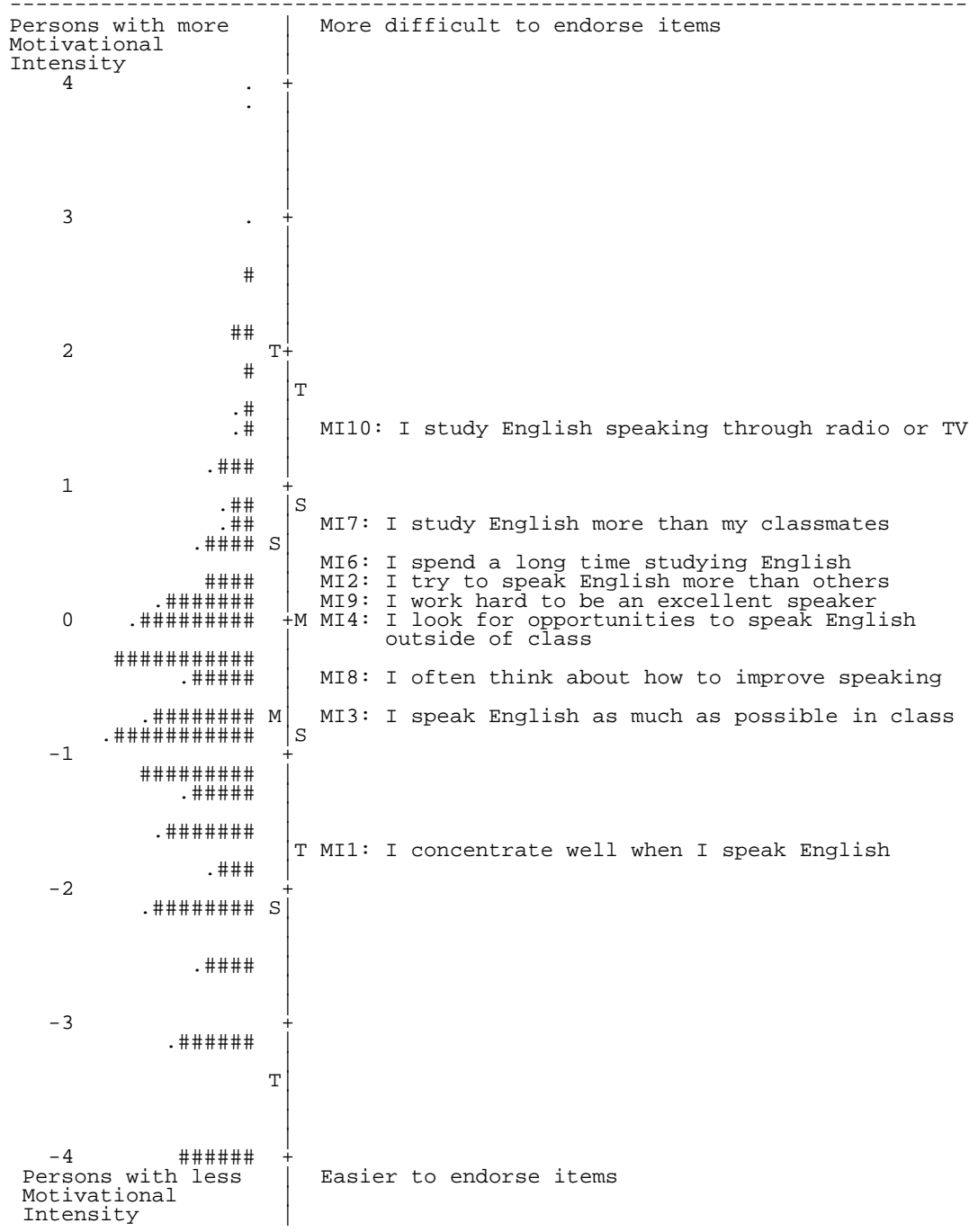


Figure 18. Item-person map for L2 Speaking Motivational Intensity. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

The mean person ability estimate was $-.94$, which was slightly lower than the mean item difficulty estimate. Therefore, the items were slightly difficult for the participants to endorse overall. The Rasch person reliability (separation) estimate was $.83$ (2.19), and the Rasch item reliability (separation) estimate was $.99$ (9.67).

The Rasch PCA of item residuals for the nine ALSE items were examined. The Rasch model accounted for 52.1% of the variance (eigenvalue = 9.8), which exceeded the 50% criterion. The unexplained variance explained by the first residual was 10.0% (eigenvalue = 1.9). These findings indicated that the items measuring ALSE formed a fundamentally unidimensional construct.

Table 41 shows that items ALSE2 and ALSE3 had high positive loadings above $.70$, and items ALSE4 and ALSE10 had high negative loadings above $-.50$. The items with positive loadings were related to “enjoying speaking English more than reading or writing,” while the items with negative loadings were related to “interest in learning to speak English.”

Table 41. *Rasch Principal Component Analysis for the Attitude Toward Learning to Speak English Items*

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
ALSE2	.75	.42	1.00	1.04
ALSE3	.71	-.03	.92	.87
ALSE7	.16	-.19	1.16	1.15
ALSE9	.08	.87	.96	.95
ALSE10	-.56	1.04	.90	.84
ALSE4	-.51	-1.00	.90	.86
ALSE6	-.37	-2.55	1.30	1.47
ALSE8	-.20	1.16	.98	1.10
ALSE1	-.06	.28	.82	.81

Desire to Learn to Speak English

The six-point rating scale for the Desire to Learn to Speak English items was examined using the Rasch rating scale model. The results are shown in Table 42.

At least 10 responses were made for each category, the Outfit MNSQ statistics were below 2.00, no thresholds were disordered, and each threshold was separated by at least .59 logits except for categories 2 and 3, which were separated by .23 and categories 5 and 6, which were separated by .54.

Table 42. *Six-Point Rating Scale Functioning for Desire to Learn to Speak English*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Strongly disagree	147 (4)	1.31	1.37	None	(-3.05)
2 Disagree	292 (7)	.97	1.10	-1.57	-1.64
3 Slightly disagree	848 (22)	.95	1.04	-1.34	-.57
4 Slightly agree	1149 (29)	.82	.87	-.04	.54
5 Agree	723 (19)	.79	.81	1.32	1.65
6 Strongly agree	741 (19)	1.11	1.10	1.63	(3.09)

Categories 1 and 2 were first combined given the relatively few responses in those categories and this created the five-point rating scale (Table 43). Categories 2 and 3 were separated 1.05 logits, but categories 4 and 5 were separated by only .36 logits, which is smaller than .81, the criterion for the five-point scale.

Table 43. *Five-Point Rating Scale Functioning for Desire to Learn to Speak English*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	439 (11)	1.16	1.25	None	(-2.90)
2 Slightly disagree	848 (22)	.99	1.11	-1.59	-1.19
3 Slightly agree	1149 (29)	.86	.90	-.54	.09
4 Agree	723 (19)	.81	.83	.89	1.23
5 Strongly agree	741 (19)	1.06	1.07	1.25	(2.69)

Next, categories 4 and 5 were combined to create a four-point rating scale. The results of that analysis are shown in Table 44. Each threshold was separated greater than 1.10 criterion for the four-point rating scale, and, as shown in Figure 21, the shape of the probability curves for each category was peaked.

Table 44. *Four-Point Rating Scale Functioning for Desire to Learn to Speak English*

	Count (%)	Infit MNSQ	Outfit MNSQ	Structure calibration	Category measure
1 Disagree	439(11)	1.11	1.42	None	(-2.50)
2 Slightly disagree	848(22)	.98	1.16	-1.23	-.73
3 Slightly agree	1149(29)	.88	.84	.03	.74
4 Agree	1464(38)	.98	.99	1.20	(2.49)

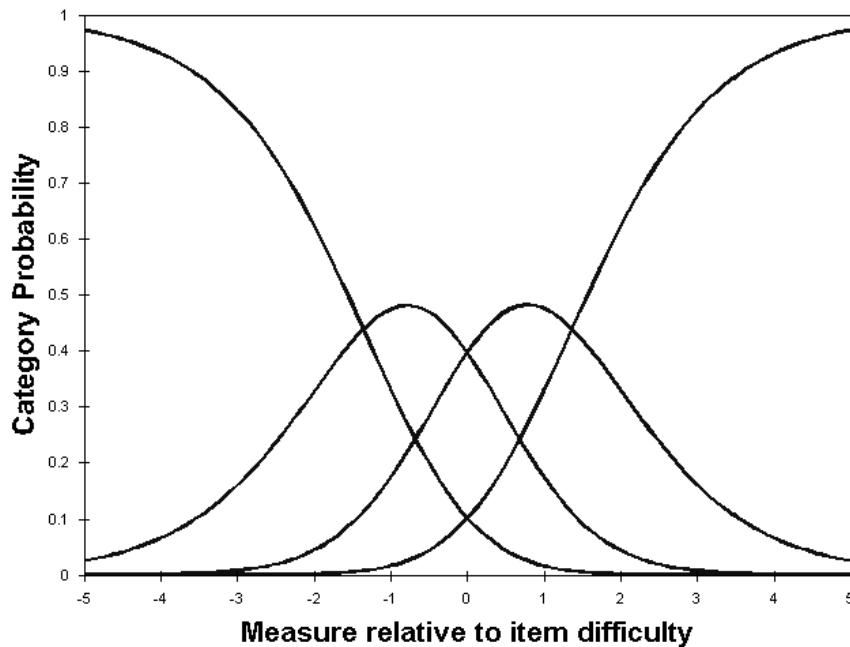


Figure 21. Category probability curves for the four-point rating scale for Desire to Learn to Speak English.

All of the Desire to Learn to Speak English items except item DLSE10 (*I wish I could speak English perfectly*) met the .50-1.50 Infit and Outfit MNSQ fit

criterion (Table 45). Item DLSE10 probably misfit because the word *perfectly* might have caused confusion. In the factor analysis, item DLSE10 loaded strongly with the ALSE items. Because of problems identified in both the Rasch analysis and factor analysis, item DLSE10 was deleted.

Table 45. *Rasch Item Statistics for the Desire to Learn to Speak English Items*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
DLSE6	1.37	.07	.88	-1.8	.92	-1.1	.72
DLSE2	.83	.07	.86	-2.2	.93	-.9	.73
DLSE1	.57	.07	1.11	1.5	1.12	1.6	.66
DLSE9	.27	.07	.75	-4.0	.74	-3.8	.73
DLSE7	.20	.07	.92	-1.2	1.01	.2	.66
DLSE5	-.31	.07	1.12	1.6	1.15	1.7	.58
DLSE4	-.34	.07	1.08	1.1	1.22	2.4	.57
DLSE3	-.44	.07	.73	-4.0	.69	-3.8	.69
DLSE8	-.52	.08	1.00	.0	1.03	.4	.61
DLSE10	-1.62	.10	1.88	7.6	2.01	5.5	.36

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .80.

As shown in Table 46, after deleting item DLSE10, the remaining DLSE items met the fit criterion. These nine items were used for the DLSE construct.

Table 46. *Rasch Item Statistics for the Desire to Learn to Speak English Items Excluding Item DLSE10*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
DLSE6	1.28	.07	.91	-1.3	.92	-1.0	.73
DLSE2	.70	.07	.88	-1.9	.91	-1.2	.74
DLSE1	.42	.07	1.16	2.3	1.17	2.2	.67
DLSE9	.09	.07	.80	-3.0	.79	-2.9	.74
DLSE7	.02	.07	.99	-.1	1.07	.9	.66
DLSE5	-.53	.08	1.19	2.5	1.25	2.7	.59
DLSE4	-.56	.08	1.15	2.0	1.30	3.1	.58
DLSE3	-.67	.08	.78	-3.2	.75	-3.0	.70
DLSE8	-.76	.08	1.10	1.3	1.13	1.3	.61

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .81.

Before continuing the analysis, three ALSE items (items ALSE5, ALSE12 and ALSE13) that had negative loadings above $-.40$ in the Rasch PCA analysis (See Table 31) were examined. Because these three items misfit the Rasch model, I examined whether they formed a unidimensional construct with the DLSE items. Table 47 shows that item ALSE5 misfit the Rasch model. In the factor analysis, item ALSE5 did not load strongly on any factor.

Table 47. *Rasch Item Statistics for the Desire to Learn to Speak English Items Including Items ALSE5, ALSE12, and ALSE13*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
DLSE6	1.39	.07	.88	-1.8	.92	-1.0	.69
DLSE2	.90	.07	.80	-3.1	.80	-3.0	.72
DLSE1	.66	.07	1.07	1.1	1.09	1.2	.63
DLSE9	.39	.07	.75	-4.0	.75	-3.7	.71
DLSE7	.33	.07	.87	-2.1	.93	-.9	.64
DLSE4	-.17	.07	1.02	.3	1.13	1.5	.54
DLSE5	-.14	.07	1.04	.6	1.09	1.1	.55
DLSE3	-.26	.07	.70	-4.6	.65	-4.5	.67
DLSE8	-.34	.07	.94	-.9	.94	-.7	.59
ALSE13	-.48	.07	1.09	1.2	1.15	1.6	.50
ALSE12	-.69	.08	1.20	2.4	1.31	2.8	.43
ALSE5	-1.59	.10	2.08	8.3	2.74	8.1	.22

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .79.

After deleting item ALSE5, a Rasch analysis was conducted with the remaining items. As shown in Table 48, item ALSE12 mistfit the Rasch model. In the factor analysis, item ALSE12 loaded strongly with ALSE and DLSE items, so it was deleted.

Table 48. *Rasch Item Statistics for the Desire to Learn to Speak English Items Including Items ALSE12, and ALSE13*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
DLSE6	1.34	.07	.89	-1.6	.94	-.8	.71
DLSE2	.81	.07	.84	-2.4	.85	-2.1	.73
DLSE1	.55	.07	1.10	1.4	1.11	1.5	.65
DLSE9	.26	.07	.78	-3.5	.77	-3.3	.72
DLSE7	.19	.07	.91	-1.4	.97	-.4	.65
DLSE8	-.51	.07	1.00	.0	1.00	.0	.60
DLSE5	-.31	.07	1.12	1.6	1.19	2.2	.56
DLSE4	-.33	.07	1.09	1.2	1.23	2.5	.55
DLSE8	-.51	.07	1.00	.0	1.00	.0	.60
ALSE13	-.67	.08	1.20	2.6	1.27	2.6	.50
ALSE12	-.90	.08	1.31	3.7	1.53	4.5	.42

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .80.

The remaining items were examined. As shown in Table 49, item ALSE13 misfit the Rasch model. In the factor analysis, item ALSE13 loaded strongly with the ALSE and DLSE items. Therefore, although items ALSE5, ALE12, and ALSE13 had high negative loadings in the Rasch PCA of item residuals analysis (Table 31), these items showed poor fit to the Rasch model, so they were not included in the Desire to Learn to Speak English construct.

Table 49. *Rasch Item Statistics for the Desire to Learn to Speak English Items Including Item ALSE13*

Item	Measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Pt- measure correlation
DLSE6	1.30	.07	.90	-1.5	.93	-.9	.72
DLSE2	.75	.07	.84	-2.4	.88	-1.7	.74
DLSE1	.49	.07	1.12	1.7	1.12	1.5	.67
DLSE9	.18	.07	.78	-3.5	.76	-3.4	.74
DLSE7	.11	.07	.94	-.9	1.01	.2	.66
DLSE5	-.42	.07	1.16	2.2	1.23	2.5	.57
DLSE4	-.44	.07	1.12	1.7	1.26	2.8	.56
DLSE3	-.55	.08	.76	-3.6	.73	-3.3	.69
DLSE8	-.63	.08	1.04	.5	1.04	.4	.61
ALSE13	-.79	.08	1.34	4.1	1.51	4.6	.48

Note. $N = 390$. Rasch item reliability = .99. Rasch person reliability = .81.

The item-person map for Desire to Learn to Speak English is shown in Figure 22. The easiest to endorse item, item DLSE8, concerned the importance of teaching speaking English in Japanese schools. Other easy to endorse items, items DLSE4 and DLSE5, concerned the idea that speaking English is more important than other major skills, such as reading and writing. Three items in the upper part of the scale, items DLSE1, DLSE2, and DLSE7, concerned the desire to study speaking English in addition to required English speaking classes. The most difficult to endorse item, item DLSE6, concerned seeking opportunities to speak English.

The mean person ability estimate was .78, which was higher than the mean item difficulty estimate. This indicates that the items were easy for the participants to endorse overall. The Rasch person reliability (separation) estimate was .81 (2.05), and the Rasch item reliability (separation) estimate was .99 (8.54).

The Rasch model accounted for 51.2% of the variance (eigenvalue = 9.4), which is higher than the 50% criterion. The unexplained variance explained by the first residual was 12.8% (eigenvalue = 2.4). Given the amount of variance explained by the Rasch model and the small eigenvalue of the first contrast, the items measuring Desire to Learn to Speak English appeared to form a fundamentally unidimensional construct.

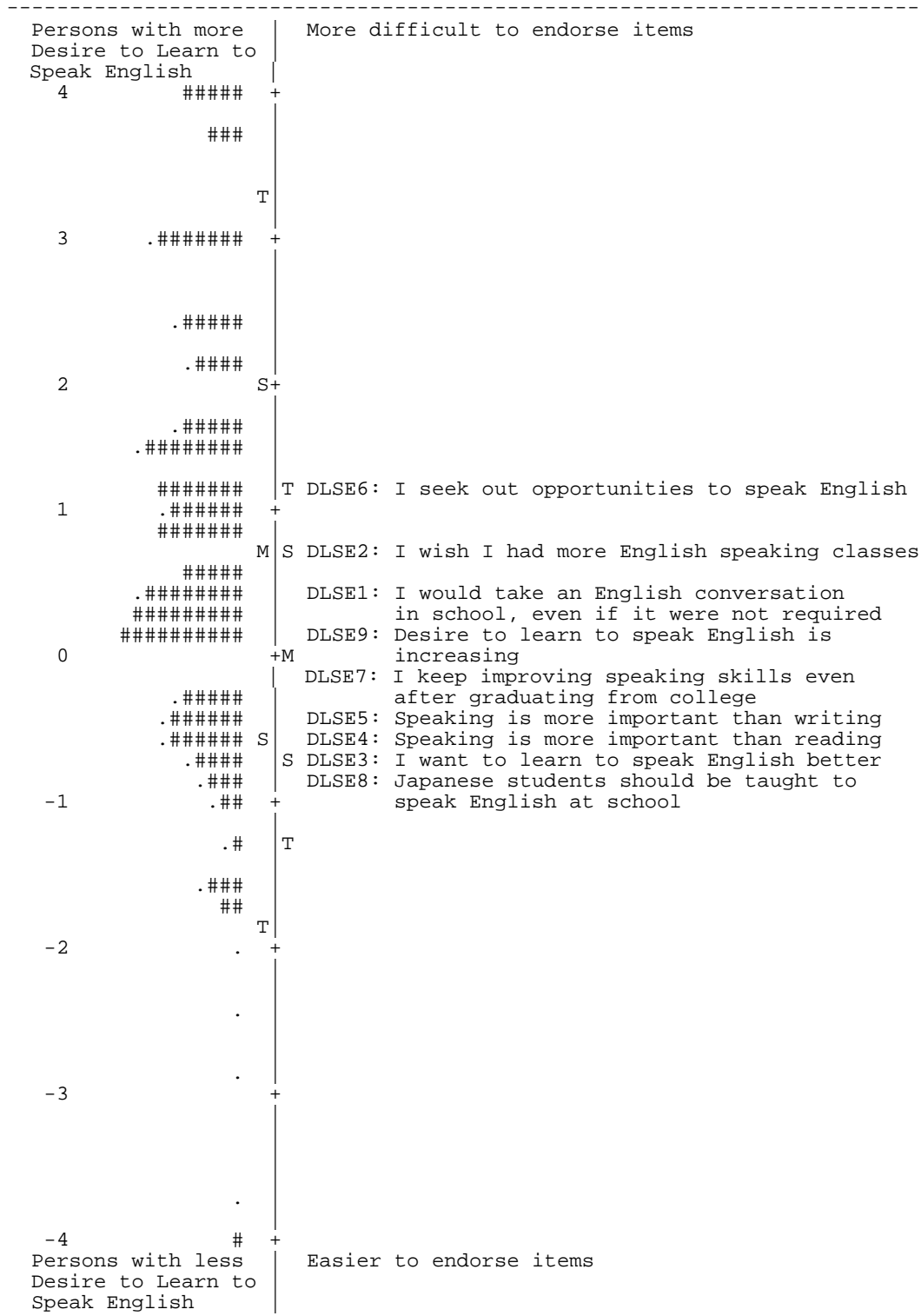


Figure 22. Item-person map for Desire to Learn to Speak English. Each '#' is 3 persons. Each '.' is 1 to 2 persons. M = Mean; S = 1 SD; T = 2 SD.

As shown in Table 50, items DLSE4 and DLSE5 loaded positively above .40, and items DLSE1, DLSE2, DLSE6, and DLSE7 had high negative loadings between -.38 and -.46. The positively loading items are related to the importance of speaking English, while the negatively loading items are concerned with desire to improve English-speaking skill.

Table 50. Rasch Principal Component Analysis for the Desire to Learn to Speak English Items

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
DLSE5	.87	-.53	1.19	1.25
DLSE4	.86	-.56	1.15	1.30
DLSE8	.24	-.76	1.10	1.13
DLSE2	-.46	.70	.88	.91
DLSE1	-.45	.42	1.16	1.17
DLSE6	-.40	1.28	.91	.92
DLSE7	-.38	.02	.99	1.07
DLSE9	-.31	.09	.80	.79
DLSE3	-.02	-.67	.78	.75

Comparison of the SPSS Factor Analysis and the Rasch Analysis Results

Table 51 shows a comparison of the Rasch PCA results and the SPSS Factor Analysis results. All seven constructs were identified by both analyses. All ten Self-Esteem items loaded above .40 in the factor analysis. Because item SE7 misfit the Rasch analysis, it was deleted.

Nine L2 Speaking Anxiety items loaded together above .40 in the factor analysis, while item ANX10 loaded weakly on this factor. Thus, item ANX10 was deleted from the L2 Speaking Anxiety construct.

Table 51.A Comparison of the Constructs Identified by Rasch Analysis and SPSS Factor Analysis

Factor/Construct	Factor Analysis Items loading onto factor	Rasch PCA Items loading onto factor
Self-Esteem	SE item1,2,3,4,5,6,7,8,9, and 10	SE items1,2,3,4,5,6,8,9, and 10
L2 Speaking Anxiety	ANX items1,2,3,4,5,6,7,8, and 9	ANX items1,3,4,5,6,7,8, and 9
L2 Willingness to Communicate	WTC items1,2,3,4,5,6,7,8,9,10,11, and 12	WTC items 2,3,4,5,6,7,8,9,10, and 11
L2 Speaking Self-Confidence	SC items1,2,3,4,5,6,7,8,9,10,11,12,13, and 16	SC items1,2,3,4,5,6,7,8,9,10,12,13,14,15, and 16
L2 Speaking Motivational Intensity	MI items2,5,6,7,8,9, and 10	MI items 1,2,3,4,6,7,8,9, and 10
Attitude Toward Learning to Speak English	ALSE items1,2,3,8, and 9	ALSE items1,2,3,4,6,7,8,9, and 10
Desire to Learn to Speak English	DLSE items3,7,8, and 10 ALSE items4,6,11,12,13 MI item1	DLSE items1,2,3,4,5,6,7,8, and 9

All twelve L2 WTC items loaded above .40 in the factor analysis. However, items WTC1 and WTC12 showed poor fit to the Rasch model. Examination of these items showed that answering teacher’s questions in class and guiding a tour might be different from the L2 WTC construct, so they were deleted.

In the factor analysis, two L2 Speaking Self-Confidence items (items SC14 and SC15) failed to load on the L2 Speaking Self-Confidence factor. However, in the Rasch analysis, items SC14 and SC15 showed good fit to the Rasch model. Because these two items are the easiest items in the construct, they were retained. On the other hand, item SC11 showed poor fit to the Rasch model. Telling the time

is probably different from the L2 Speaking Self-Confidence construct, so this item was deleted.

Only seven L2 Speaking Motivational Intensity items (items MI2, MI5, MI6, MI7, MI8, MI9, and MI10) loaded on the L2 Speaking Motivational Intensity construct; however, all the L2 Speaking Motivational Intensity items except item MI5 showed good fit to the Rasch model. Therefore, nine items are used for the L2 Speaking Motivational Intensity construct.

In the factor analysis, only five Attitude Toward Learning to Speak English items (items ALSE1, ALSE2, ALSE3, ALSE8, and ALSE9) loaded together above .40 on the same factor; on the other hand, the Rasch analysis showed that all the Attitude Toward Learning to Speak English items except item ALSE5 had good fit to the Rasch model. Therefore, all the items except item ALSE5 were retained for the Attitude Toward Learning to Speak English construct.

The Desire to Learn to Speak English construct was not well identified by the factor analysis because the items from three constructs (L2 Speaking Motivational Intensity, Attitude Toward Learning to Speak English, and Desire to Learn to Speak English) loaded above .40 on this factor. On the contrary, the Rasch analysis indicated that all the Desire to Learn to Speak English items except item DLSE10 formed a unidimensional construct. Thus, all the DLSE items except item DLSE10 are used for the Desire to Learn to Speak English construct.

Facets Analysis of the Speaking Data

In order to check the validity of speaking assessment rubric (Appendix C), I collected speaking data of six students who were not included in the main study. First, oral proficiency interviews were conducted individually and I asked each student if there were any questions that they found difficult to answer and if they could understand the cartoon. All six students said that no questions were difficult to answer and the cartoon was easy to understand. Then, they were shown the speaking assessment rubric and asked if they could understand each category. All six students said they could easily understand grammar, vocabulary, and pronunciation categories. At first, three of them were not familiar with the word fluency, but after I explained fluency by pointing out the keywords such as pauses and hesitation and they read the descriptions of the rubric, they understood what fluency means. Many Japanese students seem to understand grammar, vocabulary, and pronunciation categories, but some might not be familiar with the word fluency, but when fluency was explained with some keywords and they read the descriptions of the rubric, they could understand and distinguish the four categories very well. Thus, it was decided that during the interview when students do not understand the fluency category, the interviewer would provide an explanation to help them understand it.

Their speaking data were recorded during the interview, and I asked two raters who would participate as Rater 3 and Rater 4 in the main study to assess the six students based on the rubric. After they finished assessing the six students, I asked the raters about the speaking assessment rubric and both of them said that

they found no problem assessing the students using the rubric. Their scores were entered into a multi-faceted Rasch analysis using the Facets. The modeled facets were raters, students, and categories (grammar, fluency, vocabulary, and pronunciation).

Rating scale validation can be determined by checking the category statistics (Table 52). Because there were only six students, higher categories, 8 and 9, were not used, and categories 4 and 5 were not separated well. In the main study, the categories will be combined if the scale does not meet the criteria for a nine-point scale. The shape of probability curves is shown in Figure 23. Because category was underused, the probability curve for category 4 is not peaked well, and there are no probability curves for categories 8 and 9 because these categories were not used.

Table 52. Category Statistics for the Step Difficulties of the Rating Scale

Category score	Counts used (%)	Average measure	Expected measure	Outfit MNSQ	Measure	SE
1	4 (8%)	-2.80	-3.15	1.1	—	—
2	6 (13%)	-2.53	-2.26	1.3	-3.06	.65
3	8 (17%)	-1.60	-1.40	.7	-2.10	.53
4	4 (8%)	-.02	-.37	.1	-.22	.58
5	7 (15%)	.92	1.26	.7	-.18	.64
6	8 (17%)	3.69	3.10	1.9	2.10	.60
7	11 (23%)	4.25	4.43	1.0	3.46	.51
8	0 (0%)	.00	.00	.0	.00	.00
9	0 (0%)	.00	.00	.0	.00	.00

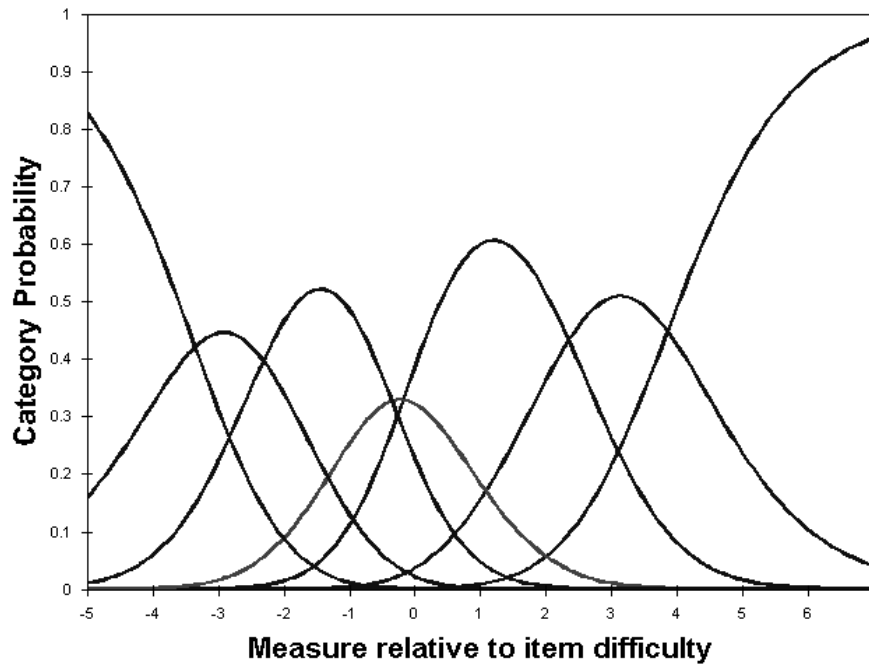


Figure 23. Rating scale for category probability

Figure 24 shows the Facets map for three facets, raters, students, and categories. The scale in the first column is the logit scale. The second column displays rater severity, which reveals that Rater 1 is more severe than Rater 2. The third column shows the students' estimated speaking ability. The most able speaker, Student 5, is at the top and the least able speaker, Student 3, at the bottom. The fourth column indicates category difficulty. Grammar was the most difficult category, while fluency and pronunciation were easier categories. The last column shows the scale used by the teacher raters.

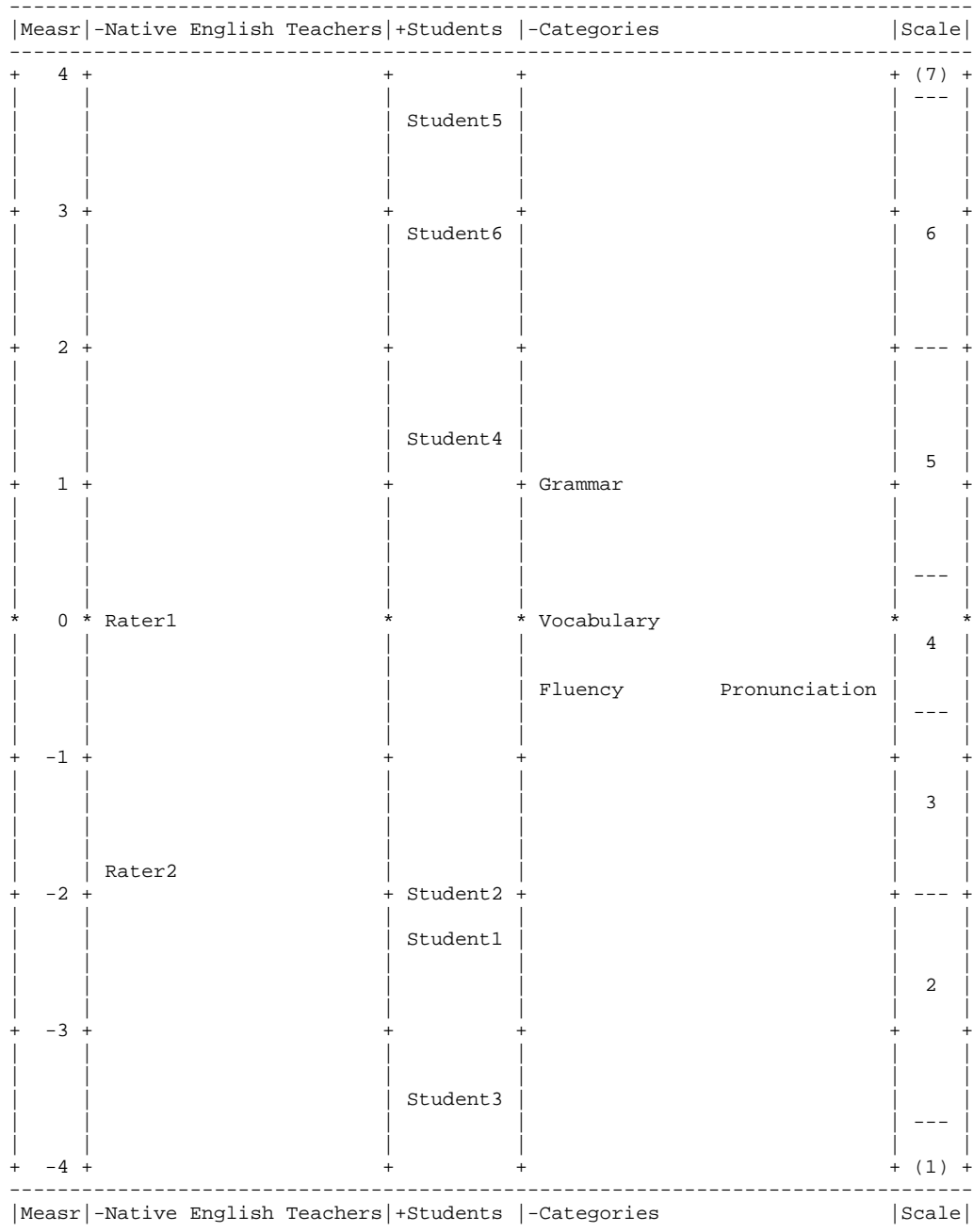


Figure 24. Facets map for two teacher-raters.

Table 53 provides a rater measurement report. The Infit and Outfit MNSQ statistics met the criterion of .5 to 1.5 (Linacre, 2007). The separation index was

2.99 and the reliability estimate was .90. The significant chi-square of 19.8, $df=1$, indicated that all raters were not equally severe.

Table 53. Calibration Report for Two Teacher-Raters

Rater	Observed		SE	Infit		Outfit	
	Score	Logit		MNSQ	ZSTD	MNSQ	ZSTD
1 (Rater3)	96	-.05	.27	.86	-.4	.88	-.3
2 (Rater4)	120	-1.76	.27	.70	-.9	1.07	.3

Note. Fixed (all same) chi-square: 19.8, $df = 5$, $p < .001$.

Table 54 shows the calibration report for students. The Infit MNSQ statistics met the criterion of .5 to 1.5, but the Outfit MNSQ statistics for student 5 was greater than the criteria. Student 5 was the most proficiency student, so his higher proficiency compared with others might have caused the misfit. The separation index was 5.30 and the reliability estimate was .97. The significant chi-square of 161.3, $df=5$, indicating that students differed in their speaking abilities.

Table 54. Calibration Report for Six Students

Student	Observed		SE	Infit		Outfit	
	Score	Logit		MNSQ	ZSTD	MNSQ	ZSTD
1	24	-2.28	.40	.72	-.4	.75	-.4
2	26	-1.97	.40	.60	-.7	.65	-.6
3	17	-3.52	.46	1.32	.7	1.35	.7
4	45	1.28	.47	.43	-1.2	.46	-1.2
5	53	3.63	.68	.90	.0	1.86	1.1
6	51	2.85	.57	.83	-.1	.80	-.1

Note. Fixed (all same) chi-square: 161.3, $df = 5$, $p < .001$.

Table 55 shows the calibration report for categories. Grammar was the most difficult category, while Pronunciation was the easiest. The Infit MNSQ statistics met the criterion of .5 to 1.5, but the Outfit MNSQ statistics for pronunciation was

greater than the criteria. The separation index was 1.26 and the reliability estimate was .61. The significant chi-square of 10.3, $df = 3$, indicating that categories differed in their difficulties. Therefore, the assessment rubric could separate the six students in terms of their oral proficiency.

Table 55. Calibration Report for Categories

Category	Observed		SE	Infit		Outfit	
	Score	Logit		MNSQ	ZSTD	MNSQ	ZSTD
Grammar	47	1.00	.38	.50	-1.3	.70	-.6
Vocabulary	54	.00	.38	.62	-.9	.69	-.6
Fluency	57	-.43	.38	.73	-.5	.69	-.5
Pronunciation	58	-.58	.39	1.27	.7	1.83	1.4

Note. Fixed (all same) chi-square: 10.3, $df = 3$, $p < .05$.

CHAPTER 5

RESULTS

In the preliminary analysis chapter, I reported how the participants' raw scores were converted to interval-level Rasch person measures, which were used for the statistical analyses reported in this chapter. The purpose of this chapter is to present the results for the four research questions.

Facets Analysis of the Speaking Data

In order to validate the speaking assessment, the data collected from 390 participants were analyzed. Five teacher raters assessed the participants based on the rubric in Appendix C. Rater 1 assessed Students 1 to 46, Rater 2 assessed Students 26 to 344, Rater 3 assessed Students 50 to 390, Rater 4 assessed Students 97 to 196, and Rater 5 assessed Students 1 to 25 and 97 to 390. The students also assessed their own oral performance. The raw scores were entered into a multi-faceted Rasch analysis using the Facets software package. The modeled facets were raters (five teacher raters and self-assessment done by students), students, and the four assessment categories of grammar, vocabulary, fluency, and pronunciation.

Student 17 received the lowest possible score from both teacher raters, making this participant an extreme outlier because the student's model standard error was 22.58, while the mean SE was 4.0, and Infit MNSQ was .30 and Outfit MNSQ was .32, which were below .50 criterion suggested by Linacre (2007). This

person's score was therefore deleted from the analysis, and data from 389 participants were entered into the Facets analysis.

The nine-point rating scale was examined using the criteria set by Linacre (2007) and Wolf and Smith (2007); At least 10 responses should be made for each category, Outfit MNSQ should be below 2.00, no thresholds should be disordered, and each category should be separated by at least .15 logits for a nine-point scale. As can be seen from Table 56, all criteria were met. However, categories eight and nine were underused, so it might be better to combine them. Yet if categories eight and nine were combined, the highest category of the KEPT, Level 5, would be lost. Therefore, because all criteria for the nine-point scale were satisfied, a nine-point rating scale was used in this study. Figure 25 shows the probability curve for a nine-point scale. The shapes of the probability curves were peaked for each category.

Table 56. *Category Statistics for the Step Difficulties of the Nine-Point Rating Scale*

Category score	Counts used (%)	Average measure	Expected measure	Outfit MNSQ	Measure	SE
1	127 (3%)	-3.83	-3.62	.7	—	—
2	267 (5%)	-3.06	-2.89	.7	-3.97	.10
3	597 (12%)	-2.26	-2.23	.9	-3.37	.07
4	1,312 (27%)	-1.31	-1.40	1.0	-2.62	.05
5	1,451 (30%)	-.40	-.44	1.0	-1.02	.04
6	734 (15%)	.47	.51	1.1	.72	.05
7	301 (6%)	1.40	1.51	1.2	1.89	.07
8	88 (2%)	2.39	2.66	1.2	3.30	.12
9	16 (0%)	4.02	3.98	1.0	5.08	.30

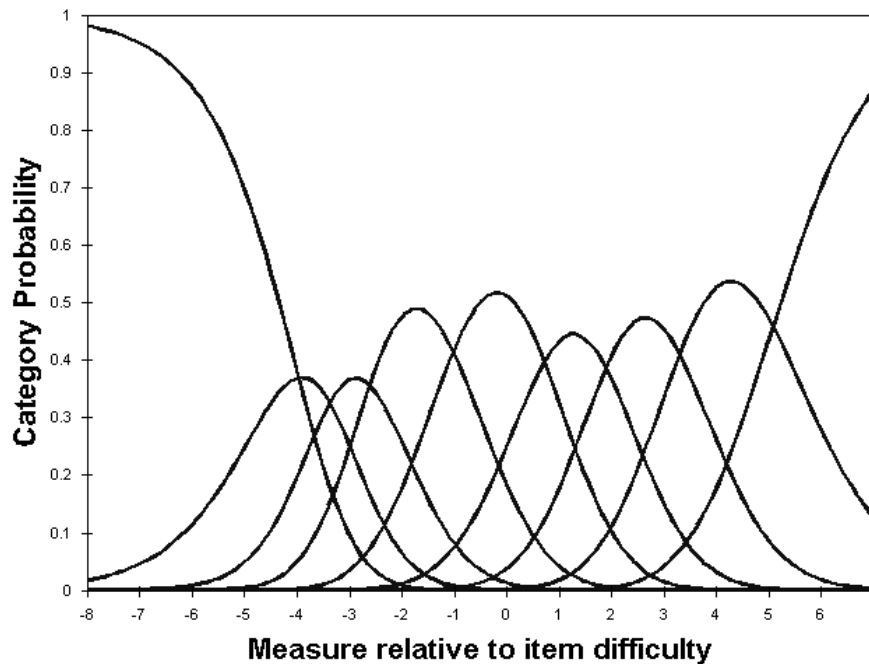


Figure 25. Rating scale category probability for the nine-point rating scale.

Research Question 1: The Validity of Self-Assessment

The first research question, “To what degree do Japanese students’ self-assessments of their L2 oral performance differ from teacher-assessments?,” was answered by examining the Facets results including the Facets map and bias interactions and by calculating the correlations between self-assessment measures and teacher-assessment measures.

Rater severity, students’ speaking ability, and item difficulty were checked with the Facets map (Figure 26). The teacher-ratings were separated from the self-ratings by weighting self-assessment scores at .001. This permitted bias interaction reports that compared self-ratings to the average of the teacher-ratings without the self-ratings unduly influencing the person measures. This approach also facilitated a comparison of how the teachers as a group and the self-raters constructed the

rating scale. Figure 26 shows the measures for rater severity, speaking ability, and item difficulty. The scale in the first column is the logit scale. The second column shows the students' estimated speaking ability. The most able speaker is at the top and the least able at the bottom. The third column displays rater severity. The most severe rater, Self-Rater (self-assessment), is at the top, while the least severe rater, Rater 2, is at the bottom. The fourth column indicates category difficulty. Grammar and vocabulary were the most difficult categories, while pronunciation was the easiest. The last two columns show the nine-point scales. Two nine-point rating scales were perceived in logit terms used by the teacher raters on the left and the students' self-assessment on the right, respectively. Comparing the two scales used by the teachers and students, we notice that the students are more likely to be severe when they use the lower categories (1 to 4) than the teacher raters. However, the students are more lenient than the teachers when they use the higher categories (7 to 9). That is, a student who performs at 3.2 logits is most likely to be rated as 7 by a teacher, but most likely to rate him or herself as 8. Conversely, a performance at -2.1 logits is most likely to be rated 4 by a teacher, but 3 by a student.

The person reliability was low at .45 (separation = .87). The person reliability indicates that the teacher-raters, when their ratings have been adjusted for severity, provide a reasonably reliable estimate of student speaking performances from the perspective of teachers. Because the person separation index was less than 1.0, the speaking test did not distinguish the sample into multiple levels. Table 56 shows that even with nine categories in this scale, more than half students fell into

categories 4 and 5 possibly because most of the participants had never been abroad and they studied English in Japanese secondary schools where they were not much taught oral communications, many participants' speaking skills were similar.

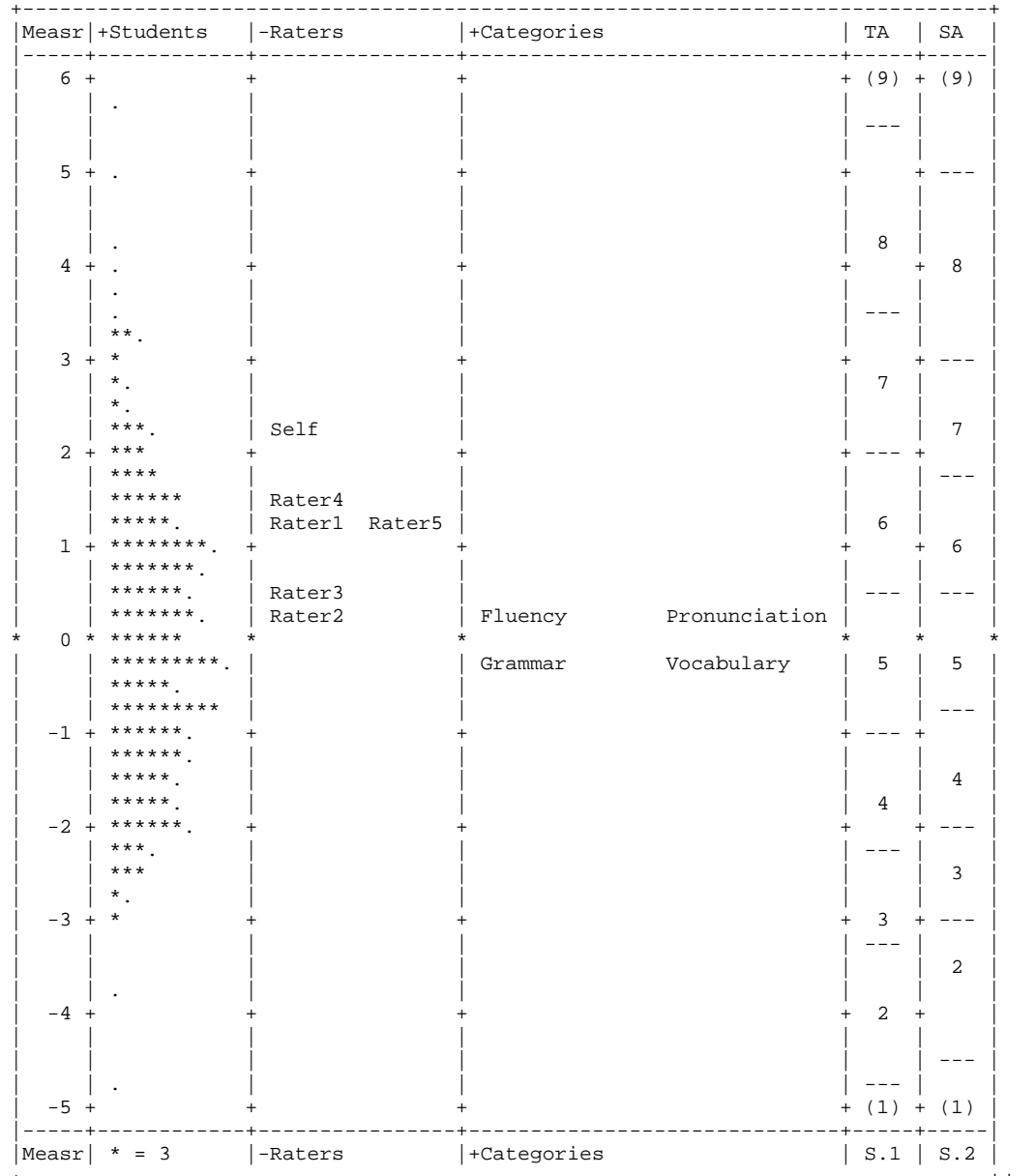


Figure 26. Facets map for teacher-raters and self-raters.

Table 57 provides a rater measurement report. The Infit and Outfit MNSQ statistics met the Infit MNSQ criterion of .5 to 1.5 except for Self-Rater. Compared with teacher raters, the self-raters did not fit the Rasch model. The separation index was 1.72 and the reliability estimate was .75. The significant chi-square of 847.4, $df = 5$, indicated that all raters were not equally severe. The rater severity of .7 (separation = 1.72) indicates that there was adequate variance in the speaking performances to reliably estimate the relative severity of teachers.

Table 57. Calibration Report for Teacher- and Self-Raters

Rater	Logit	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
1	1.27	.05	.89	-1.9	.93	-1.2
2	.16	.03	1.27	6.1	1.30	6.6
3	.57	.03	.64	-9.0	.65	-9.0
4	1.55	.06	1.04	.5	1.08	1.0
5	1.21	.03	1.05	1.2	1.06	1.5
Self	2.22	.09	3.20	1.6	3.98	1.8

Note. Fixed (all same) chi-square: 847.4, $df = 5$, $p < .001$.

Moreover, 95 out of the 100 most unexpected scores were self-assessment. Therefore, the students' self-assessment is neither reliable nor consistent compared with the teacher raters. Put another way, the self-assessments cannot be regarded as a valid measure of student performance from the perspective of the teacher raters.

Table 58 shows the category measurement report. Item reliability was .99 (separation = 8.58), indicating that the sample was large enough to precisely locate the items on the latent variable. Grammar and vocabulary were the most difficult categories, and pronunciation was the easiest.

Table 58. Calibration Report for Categories

Category	Logit	SE	Infit MNSQ	Infit ZSTD	Infit MNSQ	Infit ZSTD
Grammar	-.24	.03	.96	-.9	.99	-.1
Vocabulary	-.25	.03	.78	-5.8	.80	-5.2
Fluency	.15	.03	1.04	.9	1.05	1.0
Pronunciation	.34	.03	1.12	2.9	1.14	3.1

Note. Fixed (all same) chi-square: 223.0, $df = 3$, $p < .001$.

Figure 27 shows bias interaction of the average observations, indicating that the students awarded lower raw scores than the teacher raters in all categories. They gave the lowest raw scores for fluency.

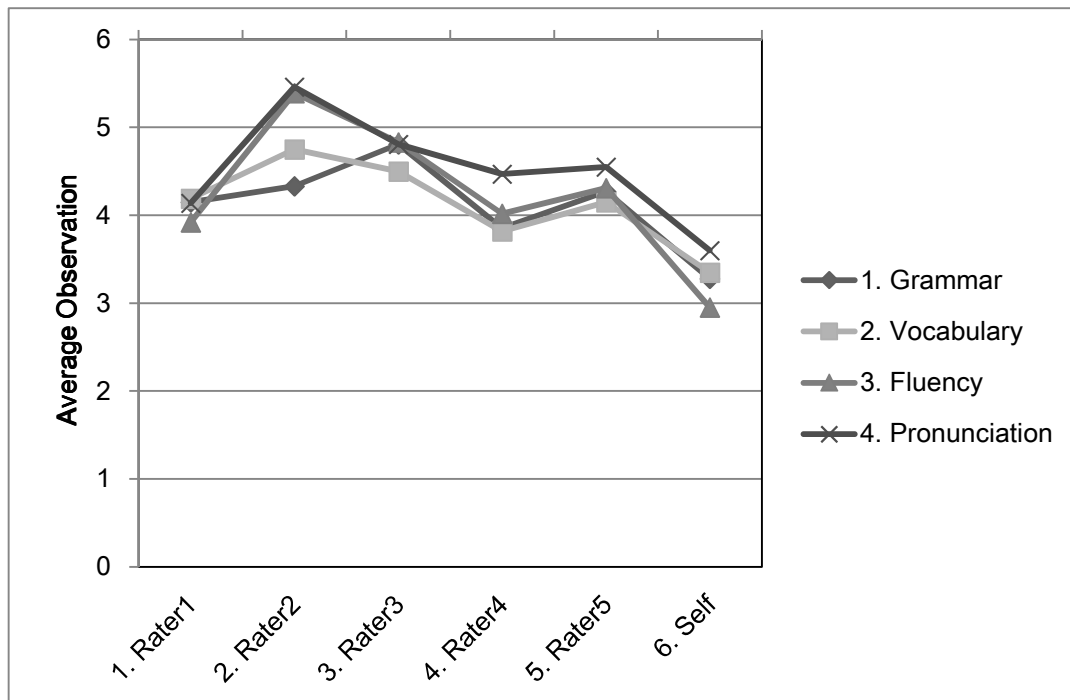


Figure 27. Bias interaction (Average observation).

Figure 28 displays bias interaction of raters and categories. This figure contrasts how the individual raters perceived the relative difficulty of the four

rating categories, relative to the overall performance measures. In contrast to the average difficulty measures of the teacher raters, the students perceived grammar and vocabulary as relatively easy categories, pronunciation as average, and fluency as relatively difficult. This perception was similar to Rater 1, but contrasts sharply with Rater 2, who ranked the categories in the opposite order. Overall, in contrast to the teachers, the students were relatively severe when rating fluency, but lenient toward vocabulary, grammar, and pronunciation.

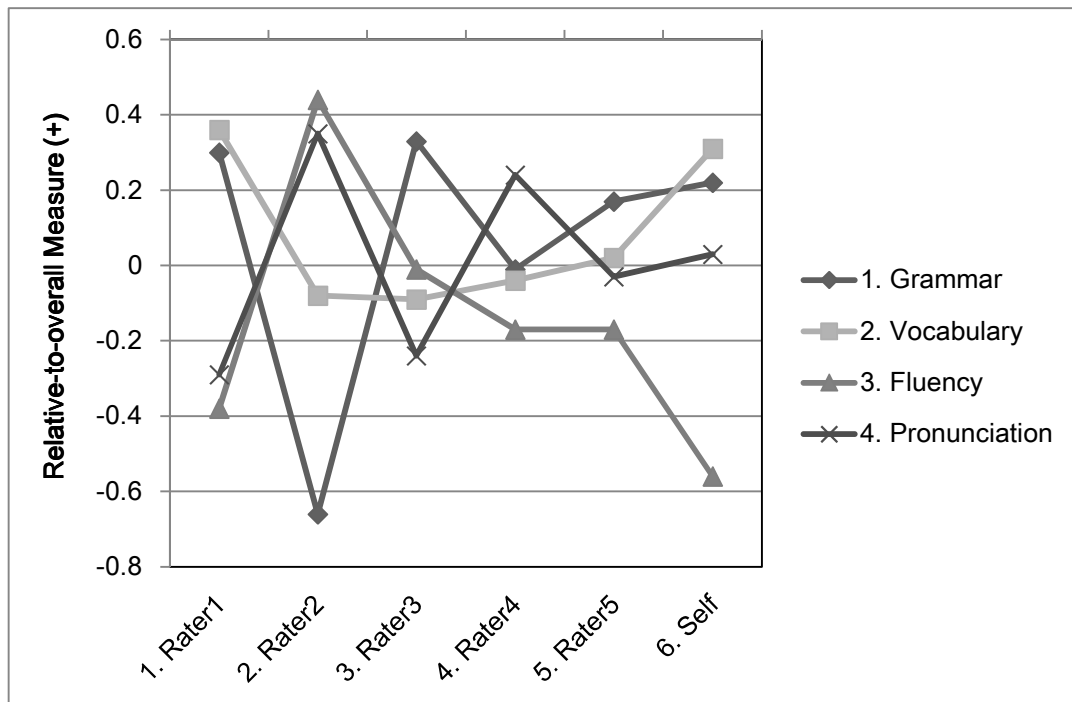


Figure 28. Bias interaction (Relative to overall measure).

Finally, using the Rasch person measures calculated by Facets, Pearson correlations were calculated between self-assessment and teacher-assessment. Table 59 shows the results of the Pearson correlations, indicating that all the

correlations were significant and had medium-sized relationships (.37 to .45). If interpreted as an indicator of inter-rater reliability, the correlation coefficients were quite low.

Table 59. *Pearson Correlations Among Self-Assessment and Teacher-Assessment Measures*

Self-assessment	Teacher-assessment				
	Total	Grammar	Vocabulary	Fluency	Pronunciation
Total	.42**				
Grammar	.40**	.41**			
Vocabulary	.39**	.38**	.37**		
Fluency	.47**	.44**	.43**	.45**	
Pronunciation	.43**	.43**	.42**	.42**	.38**

** $p < .01$

The correlations among five teacher raters were also examined. Because 50 students were assessed by all five raters, the correlations for teacher measures for these 50 students were calculated (Table 60). Although most correlations were higher than those between teacher and students, but only medium-sized correlations were found ($r = .49$ to $.80$, $p < .01$) and one correlation between Rater 2 and Rater 4 was not significant ($r = .28$, $p > .05$). Therefore, the low student/teacher correlations might be partly due to the low agreement among teachers.

Table 60. *Pearson Correlations Among Five Teacher-Raters*

Raters	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Rater 1	—				
Rater 2	.62**	—			
Rater 3	.80**	.49**	—		
Rater 4	.51**	.28	.55**	—	
Rater 5	.79**	.56**	.78**	.56**	—

** $p < .01$

Research Question 2: The Influence of L2 Affective Variables on Self-Assessment

The second research question is “Which affective variables, Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence, predict the students’ self-assessment of L2 oral performance?” In order to answer this question, first, Pearson correlations of affective variables with self-assessment, teacher-assessment, and bias size measures were compared. Second, a multiple regression analysis was conducted to investigate which affective variables predict teacher-assessment measures. Finally, a hypothetical structural model of bias size was tested.

Table 61 displays the Rasch mean logits, minimum scores, maximum scores, and the standard deviations for each affective variable. Many of the participants expressed a greater Desire to Learn to Speak English, while they had a negative Attitude Toward Learning to Speak English.

Table 61. Descriptive Statistics for Each Affective Variable (N=389)

	<i>M</i>	Minimum	Maximum	<i>SD</i>
SE	-.20	-6.47	6.64	1.45
ANX	.36	-4.63	4.89	1.44
WTC	-.32	-6.51	6.14	1.85
ALSE	-.94	-6.08	5.75	2.13
MI	-.02	-5.60	5.62	1.38
DLSE	.79	-4.92	4.97	1.59
SC	-.51	-6.93	4.99	1.72

Note. SE = Self-Esteem; ANX = L2 Speaking Anxiety; WTC = L2 Willingness to Communicate; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence.

Table 62 shows the Pearson correlations between each affective variable measures and the self-assessment measures. Table 62 also shows the Pearson correlations between each affective variable measures and the teacher-assessment measures. The results of the Pearson correlation coefficients show that self-assessment measures and teacher-assessment measures were correlated significantly with all affective variables except for Self-Esteem.

Table 62. *Pearson Correlations between Oral Performance Measures (SA and TA) and Affective Variables*

	SE	ANX	WTC	ALSE	MI	DLSE	SC
SA	.05	-.24**	.14**	.13*	.26**	.13*	.41**
SA Grammar	.10	-.20**	.09	.08	.21**	.09	.37**
SA Vocabulary	.08	-.20**	.14**	.16**	.22**	.15**	.36**
SA Fluency	.07	-.20**	.17**	.19**	.24**	.20**	.36**
SA Pronunciation	.11*	-.23**	.16**	.18**	.28**	.14**	.40**
TA	.05	-.16**	.19**	.26**	.31**	.31**	.39**
TA Grammar	.03	-.15**	.19**	.24**	.30**	.29**	.35**
TA Vocabulary	.05	-.17**	.19**	.25**	.31**	.31**	.39**
TA Fluency	.07	-.17**	.17**	.26**	.31**	.28**	.37**
TA Pronunciation	.04	-.13*	.17**	.25**	.24**	.27**	.31**

Note. SA = Self-Assessment; SE = Self-Esteem; ANX = L2 Speaking Anxiety; WTC = L2 Willingness to Communicate; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence; TA = Teacher-Assessment.

* $p < .05$. ** $p < .01$.

Table 63 shows the Pearson correlations between each affective variable measures and the bias size measures. The bias size measures for all categories were negatively correlated with Desire to Learn to Speak English, and the bias size measures for the pronunciation category were negatively correlated with L2 Speaking Anxiety. Bias size for pronunciation was also correlated with L2 Speaking Self-Confidence. On the other hand, four affective variables, Self-Esteem, L2 WTC, L2 Speaking Motivational Intensity, and Attitude Toward

Learning to Speak English were not significantly correlated with bias size measures.

Table 63. *Pearson Correlations between Oral Performance Measures (Bias) and Affective Variables*

	SE	ANX	WTC	ALSE	MI	DLSE	SC
Bias	.07	-.10	-.04	-.07	-.04	-.14**	.06
Bias Grammar	.08	-.08	-.05	-.06	-.05	-.15**	.04
Bias Vocabulary	.06	-.08	-.05	-.06	-.07	-.16**	.01
Bias Fluency	.04	-.08	-.03	-.08	-.05	-.13*	.05
Bias Pronunciation	.07	-.11*	-.02	-.07	.02	-.11*	.11*

Note. SE = Self-Esteem; ANX = L2 Speaking Anxiety; WTC = L2 Willingness to Communicate; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence; Bias = Bias Size of Self-Assessment.

* $p < .05$. ** $p < .01$.

A stepwise multiple regression was conducted using the affective variables as predictors and teacher-assessment measures as the dependent variable. Table 64 shows that students' speaking abilities were positively predicted by L2 Speaking Self-Confidence and Desire to Learn to Speak English. Thus, students with greater self-confidence and greater desire to speak English generally had higher English speaking proficiency.

Table 64. *Multiple Regression Predicting Teacher-Assessment from Affective Variables*

Model	R^2	Adjusted R^2	B	SEB	β	t
1. SC	.15	.15	.35	.04	.39	8.20**
2. SC	.18	.18	.29	.05	.31	6.33**
DLSE			.19	.05	.20	3.93**

Note. SC = L2 Speaking Self-Confidence; DLSE = Desire to Learn to Speak English. * $p < .05$. ** $p < .01$.

Stepwise multiple regression analyses were also conducted for each category, grammar, vocabulary, fluency, and pronunciation. The results are shown in Table 65. All the categories were predicted by L2 Speaking Self-Confidence and Desire to Learn to Speak English.

Table 65. Multiple Regression Predicting Teacher Assessment of Grammar, Vocabulary, Fluency, and Pronunciation with Affective Variables

Model	R^2	Adjusted R^2	B	SEB	β	t
Grammar						
1. SC	.12	.12	.48	.07	.35	7.35**
2. SC	.15	.15	.38	.07	.28	5.57**
DLSE			.28	.07	.19	3.80**
Vocabulary						
1. SC	.15	.15	.61	.07	.39	8.25**
2. SC	.18	.18	.49	.08	.32	6.37**
DLSE			.33	.08	.20	3.97**
Fluency						
1. SC	.14	.14	.50	.06	.37	7.83**
2. SC	.16	.16	.42	.07	.31	6.15**
DLSE			.25	.07	.17	3.37**
Pronunciation						
1. SC	.10	.10	.34	.05	.31	6.50**
2. SC	.13	.12	.27	.06	.25	4.83**
DLSE			.21	.06	.18	3.60**

Note. SC = L2 Speaking Self-Confidence; DLSE = Desire to Learn to Speak English. * $p < .05$. ** $p < .01$.

Second, a hypothesized structural model of bias size was tested. The results of the hypothesized structural model for the bias size of students' self-assessment of speaking skills are shown in Figure 29.

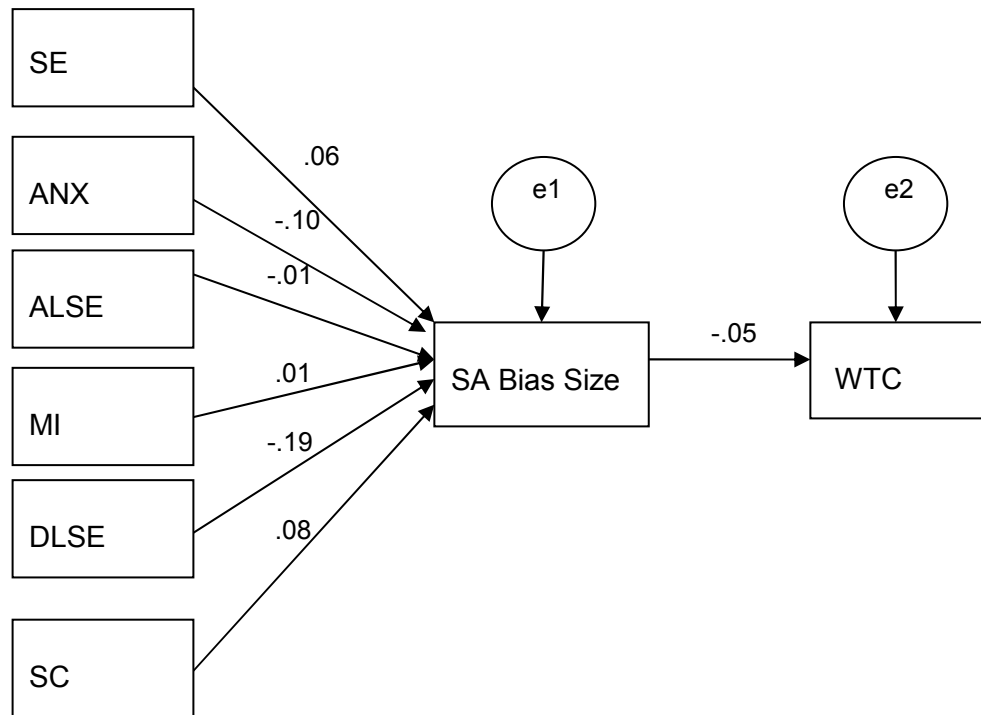


Figure 29. Path analysis results with bias size. SE = Self-Esteem; ANX = L2 Speaking Anxiety; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence; WTC = L2 Willingness to Communicate; SA Bias Size = Bias Size of Students' Self-Assessment of Speaking.

Table 66 shows that the χ^2 statistics: (= CMIN, $df = 21$, $N = 389$) was 954.493, $p = .00$, which rejected the model. Moreover, the goodness of fit index (GFI, .541) and the comparative fit index (CFI, .541) showed poor model fit, as both were below the .90 criterion. In addition, the root mean square error of approximation (RMSEA) of .338 was considerably greater than the critical value of .05.

Table 66. *Selected Fit Statistics with Bias Size (N = 389)*

	χ^2	<i>df</i>	χ^2/df	GFI	AGFI	CFI	PCFI	RMSEA	RMSEA 90%CI
Model	954.49	21	45.45	.541	.212	.013	.315	.338	[.320, .357]

$p < .001$

Additionally, Table 67 shows that only a negative path from DLSE to Bias Size was significant, implying that those with greater desire to learn to speak English tended to have a negative bias, that is to say, underestimation of their L2 speaking performance. However, other affective variables, ALSE, MI, SE, ANX, and SC, did not significantly predict Bias Size. Moreover, Bias Size did not predict WTC, implying that how the students perceive their own English speaking performances did not significantly predict the students' willingness to communicate in English, Therefore, the hypothesized structural model for the bias size was rejected.

Table 67. *Results of the Paths with Bias Size*

Outcome		Predictor	Unstandardized Coefficients (B)	SE	<i>t</i>	<i>p</i>
Bias Size	<---	ALSE	-.006	.024	-.261	.794
Bias Size	<---	DLSE	-.095	.024	-3.912	.000
Bias Size	<---	MI	.009	.031	.283	.777
Bias Size	<---	SE	.033	.029	1.151	.250
Bias Size	<---	ANX	-.065	.034	-1.930	.054
Bias Size	<---	SC	.039	.023	1.676	.094
WTC	<---	Bias Size	-.093	.104	-.900	.368

Next, stepwise multiple regression analyses were conducted using the affective variables as predictors and bias scores as dependent variables. The results are shown in Table 68. The R^2 was so low compared with the teacher-assessment

predicted by the affective variables; thus, the relationships between bias size and affective variables were not as strong as the relationships between speaking proficiency and affective variables. Table 68 shows that Desire to Learn to Speak English negatively predicted and L2 Speaking Self-Confidence positively predicted the bias size of students' self-assessment measures. Thus, students with greater Desire to Learn to Speak English tended to estimate their L2 speaking ability lower and those with greater L2 Speaking Self-Confidence tended to overestimate it. This result might have occurred because those students who had greater self-confidence in their speaking skill evaluated themselves higher, and those with greater desire to learn to speak English considered themselves as not good at speaking English and they thought they needed to improve it, so they evaluated their speaking skills lower.

Table 68. Multiple Regression Predicting Bias Size from Affective Variables

Model	R^2	Adjusted R^2	B	SEB	β	t
1. DLSE	.02	.02	-.07	.03	-.14	-2.84**
2. DLSE	.03	.03	-.09	.03	-.19	-3.52**
SC			.06	.02	.13	2.35*

Note. DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence.

* $p < .05$. ** $p < .01$.

Stepwise multiple regression analyses were also conducted for each category (i.e., grammar, vocabulary, fluency, and pronunciation). The results are shown in Table 69. The R^2 for each category was so low that bias sizes for grammar, vocabulary, fluency, and pronunciation did not have strong relationships with affective variables. The bias size for the grammar category was predicted

negatively by Desire to Learn to Speak English and positively by Self-Esteem. The bias size of the vocabulary category was negatively predicted by Desire to Learn to Speak English and L2 Speaking Anxiety. The bias size of the fluency category was predicted negatively by Desire to Learn to Speak English and positively by L2 Speaking Self-Confidence. The bias size of the pronunciation category was negatively predicted by Desire to Learn to Speak English and L2 Speaking Anxiety, and positively by L2 Speaking Self-Confidence.

Table 69. *Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables*

Model	R^2	Adjusted R^2	B	SEB	β	t
Grammar						
1. DLSE	.02	.02	-.09	.03	-.15	-2.93**
2. DLSE	.03	.03	-.11	.03	-.17	-3.25**
SE			.07	.04	.11	2.10*
Vocabulary						
1. DLSE	.03	.02	-.13	.04	-.16	-3.18**
2. DLSE	.04	.03	-.14	.04	-.17	-3.41**
ANX			-.09	.05	-.10	-2.01*
Fluency						
1. DLSE	.02	.01	-.10	.04	-.13	-2.57*
2. DLSE	.03	.02	-.13	.04	-.17	-3.16**
SC			.08	.04	.11	2.07*
Pronunciation						
1. ANX	.01	.01	-.07	.03	-.11	-2.26*
2. ANX	.03	.02	-.08	.03	-.13	-2.56*
DLSE			-.07	.03	-.13	-2.51*
3. ANX	.04	.04	-.05	.03	-.08	-1.55
DLSE			-.10	.03	-.17	-3.24**
SC			.08	.03	.14	2.50*
4. DLSE	.04	.03	-.10	.03	-.18	-3.26**
SC			.09	.03	.17	3.23**

Note. DLSE = Desire to Learn to Speak English; SE = Self-Esteem; ANX = L2 Speaking Anxiety; SC = L2 Speaking Self-Confidence.

* $p < .05$. ** $p < .01$.

Research Question 3: Characteristics of Those Who Conducted Accurate Self-Assessment

The third research question is “What are the characteristics of the students who could assess their own L2 oral performance accurately?” This question was answered by selecting the participants whose ratings were similar to the teachers’ ratings and their speaking measures and affective variable measures were compared with those of other participants whose self-assessments were less accurate. One hundred six students were selected for the accurate group, as their bias scores ranged from -0.25 to 0.25.

First, speaking proficiency of accurate group students and inaccurate group students were compared in Table 70. Speaking proficiency measures of the inaccurate group students were slightly higher than those of the accurate group students, but no significant differences were found between the two groups.

Table 70. Comparison of Accurate and Inaccurate Proficiency Groups for Speaking Proficiency

Speaking category	Group				<i>t</i> (df =387)
	Accurate (N=106)		Inaccurate (N=283)		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Total	-.13	1.45	.06	1.63	-1.05
Grammar	-.19	2.66	.07	2.19	-1.00
Vocabulary	-.22	2.65	.08	2.71	-.98
Fluency	-.22	2.05	.08	2.42	-1.12
Pronunciation	-.25	1.60	.09	1.93	-1.61

The comparison of affective variable measures between the two groups is shown in Table 71. No significant differences were found between the two groups.

Table 71. *Comparison of Accurate and Inaccurate Proficiency Groups for Affective Variables*

Affective variable	Group				<i>t</i> (<i>df</i> =387)
	Accurate (<i>N</i> =106)		Inaccurate (<i>N</i> =283)		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
SE	-.18	1.24	-.21	1.52	.18
ANX	.32	1.06	.38	1.56	-.40
WTC	-.15	1.63	-.38	1.93	1.11
ALSE	-1.03	1.86	-.90	2.23	1.14
MI	.12	1.33	-.06	1.39	-.51
DLSE	.81	1.50	.78	1.62	.17
SC	-.24	1.44	-.61	1.80	1.89

Note. SE = Self-Esteem; ANX = L2 Speaking Anxiety; WTC = L2 Willingness to Communicate; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence.

Therefore, there were no significant differences in speaking measures and affective variable measures between the accurate and inaccurate group students. There were not any distinctive features for the accurate group, so it is possible that some agreements between student raters and teacher raters occurred by chance alone.

Research Question 4: The Influence of L2 Affective Variables on Self-Assessment of High and Low Proficiency Students

The fourth research question asked, “To what degrees do the seven affective variables affect the self-assessment of high and low proficiency students differently?” This question was answered by dividing the students into high and low proficiency groups and for each group multiple regressions were conducted using affective variables (i.e., Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking

Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence) as predictors and bias size measures as dependent variable. The high proficiency group consisted of the top 100 students whose Rasch person measures ranged from 1.05 to 5.66. The low proficiency group consisted of the lowest 100 students whose Rasch person measures for L2 speaking proficiency ranged from -1.13 to -4.85. The oral interview scripts for each low, intermediate, and high proficiency student are shown in Appendix O.

Table 72 shows the descriptive statistics of the seven affective variables for the high and low proficiency groups and the *t*-test results for the two groups. The affective variable measures significantly differed between two proficiency groups except for Self-Esteem.

Table 72. Comparison of High and Low Proficiency Groups for Affective Variables

	Proficiency				<i>t</i> (<i>df</i> =198)
	High (N=100)		Low (N=100)		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
SE	-.07	1.36	-.40	1.62	1.57
ANX	-.16	1.57	.50	1.31	-2.52*
WTC	.12	1.94	-.78	2.01	3.20**
ALSE	-.28	2.10	-1.68	2.04	5.55**
MI	.50	1.38	-.60	1.40	4.78**
DLSE	1.46	1.63	.18	1.55	5.67**
SC	.38	1.50	-1.25	1.60	7.46**

Note. SE = Self-Esteem; ANX = L2 Speaking Anxiety; WTC = L2 Willingness to Communicate; ALSE = Attitude Toward Learning to Speak English; MI = L2 Speaking Motivational Intensity; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence.

p* < .05. *p* < .01.

High Proficiency Group

Stepwise multiple regressions were conducted using affective variables as predictors and the bias size of self-assessment Rasch measures as dependent variables. The results are shown in Table 73, which indicates that Self-Esteem and L2 Speaking Self-Confidence positively predicted the high proficiency students' bias size measures, while Desire to Learn to Speak English negatively predicted them.

Table 73. *Multiple Regression Predicting Bias Size of Self-Assessment from Affective Variables with High Proficiency Group*

Model	R^2	Adjusted R^2	B	SEB	β	t
1. SE	.07	.06	.14	.05	.27	2.75**
2. SE	.14	.12	.18	.05	.36	3.59**
DLSE			-.12	.04	-.28	-2.82**
3. SE	.19	.17	.14	.05	.27	2.67**
DLSE			-.15	.04	-.34	-3.39**
SC			.12	.05	.25	2.44*

Note. SE = Self-Esteem; DLSE = Desire to Learn to Speak English; SC = L2 Speaking Self-Confidence.

* $p < .05$. ** $p < .01$.

Stepwise multiple regression analyses were also conducted for each analytical category. The results are shown in Tables 74. The bias size of the Rasch grammar measures was negatively predicted by Desire to Learn to Speak English, and positively predicted by Self-Esteem and L2 Speaking Self-Confidence. The bias size of the Rasch vocabulary measures was negatively predicted by Desire to Learn to Speak English, and positively predicted by Self-Esteem and L2 Speaking Self-Confidence. The bias size of the Rasch fluency measures was negatively predicted by L2 Speaking Anxiety and L2 Willingness to Communicate. The bias size of the

Rasch pronunciation measures was positively predicted by Self-Esteem, and negatively by Desire to Learn to Speak English and L2 Speaking Anxiety.

Table 74. Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables with High Proficiency Group

Model	R^2	Adjusted R^2	B	SEB	β	t
Grammar						
1. DLSE	.06	.05	-.13	.05	-.24	-2.43*
2. DLSE	.16	.13	-.19	.05	-.34	-3.44**
SE			.21	.07	.32	3.26**
3. DLSE	.19	.17	-.22	.06	-.39	-3.93**
SE			.16	.07	.25	2.41*
SC			.14	.06	.23	2.21*
Vocabulary						
1. SE	.06	.05	.22	.09	.24	2.44*
2. SE	.11	.09	.29	.09	.32	3.13**
DLSE			-.18	.08	-.24	-2.40*
3. SE	.15	.12	.22	.09	.25	2.33*
DLSE			-.22	.08	-.29	-2.85**
SC			.18	.09	.22	2.06*
Fluency						
1. ANX	.05	.04	-.14	.07	-.21	-2.14*
2. ANX	.09	.07	-.19	.07	-.28	-2.69**
WTC			-.12	.06	-.21	-2.06*
Pronunciation						
1. SE	.11	.10	.23	.07	.33	3.51**
2. SE	.17	.15	.29	.07	.42	4.26**
DLSE			-.15	.06	-.26	-2.64*
3. SE	.24	.22	.23	.07	.33	3.35**
DLSE			-.18	.06	-.31	-3.20**
ANX			-.17	.06	-.28	-2.93**

Note. DLSE = Desire to Learn to Speak English; SE = Self-Esteem; SC = L2 Speaking Self-Confidence; ANX = L2 Speaking Anxiety; WTC = L2 Speaking Willingness to Communicate.

* $p < .05$. ** $p < .01$.

Low Proficiency Group

Stepwise multiple regressions were conducted using affective variables (i.e., Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to

Learn to Speak English, and L2 Speaking Self-Confidence) as predictors and bias size measures as dependent variable. The results are shown in Table 75. L2 Speaking Self-Confidence positively predicted the bias size of the low proficiency students' self-assessment measures.

Table 75. Multiple Regression Predicting Bias Size of Self-Assessment from Affective Variables with Low Proficiency Group

Model	R^2	Adjusted R^2	B	SEB	β	t
1. SC	.16	.15	.17	.04	.40	4.31**

Note. SC = L2 Speaking Self-Confidence.
* $p < .05$. ** $p < .01$.

A stepwise multiple regression was also conducted for each category (i.e., grammar, vocabulary, fluency, and pronunciation). The results are shown in Table 76. Bias size measures of grammar, vocabulary, fluency, and pronunciation were all positively predicted by L2 Speaking Self-Confidence.

Table 76. Multiple Regression Predicting Bias Size of Grammar, Vocabulary, Fluency, and Pronunciation from Affective Variables with Low Proficiency Group

Model	R^2	Adjusted R^2	B	SEB	β	t
Grammar						
1. SC	.15	.15	.21	.05	.39	4.22**
Vocabulary						
1. SC	.18	.17	.27	.06	.42	4.58**
Fluency						
1. SC	.14	.13	.24	.06	.38	4.02**
Pronunciation						
1. SC	.13	.12	.16	.04	.36	3.81**

Note. SC = L2 Speaking Self-Confidence.
* $p < .05$. ** $p < .01$.

Summary

There are several findings from the results of the analysis. First, a multi-faceted Rasch analysis indicated that the participants tended to rate their own L2 speaking performance more severely than the teacher raters and that the students' self-assessments were neither reliable nor consistent.

Second, multiple regression analysis revealed that the participants with greater Desire to Learn to Speak English tended to underestimate their own speaking performance and those with greater L2 Speaking Self-Confidence tended to overestimate it.

Third, 106 participants who evaluated their own speaking similar to teachers were examined, but there were no significant differences in their affective variable measures and oral proficiency measures from those who self-assessed inaccurately. Therefore, there is a possibility that their self-assessment agreed with teacher-assessment by chance alone.

Finally, 100 high proficiency students with greater Self-Esteem and L2 Speaking Self-Confidence tended to overestimate their own speaking, and those with greater Desire to Learn to Speak English tended to underestimate it. On the other hand, 100 low proficiency students with greater L2 Speaking Self-Confidence tended to overestimate their own L2 speaking performance.

CHAPTER 6

DISCUSSION

In this chapter, I reexamine and interpret the findings of the four research questions that were answered in Chapter 5, and describe the theoretical and practical implications of the study.

Research Question 1: The Validity of Self-Assessment

The first research question asked, “To what degrees do Japanese students’ self-assessments of their L2 oral performance differ from teacher-assessments?” In order to answer this research question, 389 students’ self-assessments were compared with teacher-assessments conducted by five English teachers.

First, rater severity was examined and it was shown that student raters (2.22 logits) were more severe than teacher raters (.16 to 1.27 logits). This result was not consistent with so-called above-average effect, in which on average people tend to overestimate their own abilities and put themselves above average (Dunning et al., 1989). This above-average effect can be applied to L2 learners, and some previous findings showed that students tend to overrate their performance. For example, Jafapur (1991) investigated the self-assessed L2 comprehension and L2 speaking scores of 37 Iranian students and found that mean self-assessment scores were significantly higher than teacher-assessed scores. Jassen-van Dieten (1989) also found that L2 learners from 73 countries assessed their own L2 four skills: listening,

speaking, reading, and writing, and the proportion of overestimation was six times the size of proportion of underestimation. In Anderson (1982), L2 learners from the Middle East tended to rate their own L2 four skills higher than teachers.

A few researchers, however, have reported results that were not consistent with the above-average effect, but were similar to the present study, finding that students assessed themselves more severely than teachers. For example, Anderson (1982) reported that the students from the Far East rated their English abilities lower than the teachers. Matsuno (2007) examined Japanese university students' L2 writing skills and compared self-assessments of L2 essays with teacher-assessments. The Facets map indicated that the students' self-assessed Rasch ability estimates were lower than those provided by teachers. Matsuno (2007) pointed out that the result might have been due to the Japanese traditional virtue of modesty that caused Japanese students to underestimate their abilities. Likewise, Chen (2008) found that before receiving self-assessment training, 28 Chinese students' self-assessed L2 speaking scores were significantly lower than teacher-assessed scores, and Chen argues that this result probably derived from Chinese students' tendency to underestimate their own abilities.

Some researchers reviewed in the Literature Review section compared the self-esteem of Western people such as Americans and East Asian people such as Japanese, and discovered that East Asian people tend to have lower self-esteem that leads to self-criticism, while Western people tend to have higher self-esteem that leads to self-enhancement (Brown, 2006; Markus & Kitayama, 1991). This is

because East Asian people are considered to possess modesty bias which is often necessary in order to maintain harmonious relationships with others, while Western people value individualism and tend to evaluate themselves higher than others in order to stand out (Farh et al., 1991). Moreover, according to Heine et al. (1999), most Japanese believe that self-criticism is a good thing because it serves as the basis for future improvement and achievement. Therefore, considering how Japanese people view and evaluate themselves, the result of this study that the Japanese students' self-assessments were lower than the teacher-assessments seems reasonable.

Next, the analysis of the category difficulty estimates showed that Grammar and Vocabulary were perceived by the teacher raters as more difficult categories than Fluency and Pronunciation. Therefore, the finding of this study shows that for the participants in this study Grammar and Vocabulary was more difficult than Fluency and Pronunciation. This result is in accordance with some studies such as Adams (1980) and McNamara (1990), but different from others such as Higgs and Clifford (1982) who reported that Grammar and Fluency distinguished lower students, while Vocabulary and Pronunciation distinguished more proficiency students, Iwashita et al. (2008) who found grammatical accuracy did not contribute to the overall scores of L2 speaking proficiency, and Sato (2010) who found that Fluency and Vocabulary were more difficult categories than Grammar and Pronunciation. The results on the degrees of category difficulties seemed to vary according to many factors such as context, participants, assessment rubric. When

the result of this study was compared with Bonk and Ockey (2003) and Ockey (2009) who examined Japanese university students using the KEPT, their results were similar to the present study; Grammar and Vocabulary were more difficult than Fluency and Pronunciation. That is to say, Grammar and Vocabulary contributed to distinguish the participants' L2 oral proficiency more than Fluency and Pronunciation. Possible explanation for this result might be that because the KEPT was developed to assess oral proficiency of Japanese university students, the categories were created to meet Japanese university students' speaking levels. For example, Level 5 of the KEPT, the highest level, requires "near native-like fluency and pronunciation," while the highest level of other speaking assessment rubrics such as ACTFL and IELTS demands "native-like." Although Level 4 of the KEPT is the second highest level, it allows for some unnatural pauses, some errors in rhythm, foreign accent, and occasional mispronunciations. Therefore, the Pronunciation and Fluency categories were not so difficult for Japanese university students compared with other speaking assessment rubrics that demand more native-like fluency and pronunciation.

Similar to Fluency and Pronunciation, Grammar and Vocabulary in the KEPT also require "near-native like" proficiency for the highest level, but these categories might be more difficult for Japanese students. To reach Level 5, the highest level, speakers' utterances need to have high levels of discourse structures and wide range of vocabulary, and to reach Level 4, the utterances need to contain full range of grammatical structures with some errors that do not impede the meanings, and lexis

sufficient for the task while some mistakes are allowed. Therefore, the researchers who investigated Japanese university students by using the KEPT, including the present study, reported the similar results; Grammar and Vocabulary were more difficult categories than Fluency and Pronunciation.

Additionally, because many Japanese university students had little experience speaking English, fluency and pronunciation did not contribute greatly to differentiate the participants' proficiency levels. On the other hand, grammar and vocabulary are usually tested for entrance examinations, so high proficiency students in Japan are often better at grammar and vocabulary than low proficiency students. This might be the reason why these two categories, Grammar and Vocabulary, contributed more to differentiate the participants' proficiency levels.

While the teacher-raters assessed the Grammar and Vocabulary categories as more difficult than Fluency and Pronunciation for the students of this study, the students' perception of category difficulties somewhat differed from the teacher-raters. Figures 27 and 28 indicate the results of bias interaction. To the best of my knowledge, no previous researchers have examined bias size or interactions when examining the validity of self-assessment of L2 speaking; in this study, the bias interaction was examined using Facets. It was found that the participants had negative bias regarding the relative difficulty of Fluency (-0.6 logits); this means that relative to the other rating categories, the participants generally perceived Fluency as a more difficult aspect of a speaking performance than the teachers did. That is, their self-assessments of Fluency tended to be lower than their self-

assessment of Grammar, Pronunciation, and Vocabulary. In contrast, the participants had a positive bias toward Grammar and Vocabulary (0.2 logits); this means that they regarded Grammar and Vocabulary as easier categories to excel at than Pronunciation and Fluency. In contrast, almost no bias was found for Pronunciation. There seems to be a gap between self-assessment and their actual performance. The participants regarded fluency was more difficult than grammar and vocabulary, but in fact their grammar and vocabulary skills were rated lower than fluency. The possible explanation for this gap might be that many Japanese university students spend longer time studying English grammar and vocabulary because these skills are tested for entrance examinations, so English classes at secondary schools mainly teach these skills rather than oral communication. Therefore, the participants probably had more confidence in grammar and vocabulary than fluency. However, there was a gap between actual proficiency and perception, and their grammar and vocabulary were rated lower than fluency. Nevertheless, it should be noted that the pattern of self-assessment bias was shared by some teacher-raters as well. Thus, although the students showed a notable bias toward the relative difficulty of the categories, teacher-raters varied as well.

Comparing the two scales used by the teachers and the students in the Facets map (Figure 26), it was found that the students were slightly more severe than the teacher-raters when they used the lower categories of 1 to 4. In contrast, the students were slightly more lenient than the teachers when they used the higher categories of 7 to 9. This finding suggests that the students with low L2 speaking

proficiency tended to underestimate their own L2 performance, while those with higher L2 speaking proficiency tended to overestimate their own speaking performance. Interestingly, this finding differs from the Dunning-Kruger Effect in which low proficiency people tend to overestimate their abilities because of their incompetence that prevents themselves from assessing accurately, while high proficiency people tend to underestimate their abilities because they consider their high proficiency is shared by others (Kruger & Dunning, 1999). Many studies found the consistent results with the Kruger-Dunning Effect by reporting that high achievers were more self-critical and low achievers were less self-critical (e.g., Patri 2002; Stefani, 1994; Sullivan & Hall 1997). However, Matsuno (2007), who investigated Japanese university students, reported that although those with higher proficiency tended to rate themselves severely, lower proficiency students did not display a common tendency toward overestimation or underestimation.

The findings of this study differed from other studies regarding the relationship between self-assessment and proficiency, possibly for two reasons. First, the participants in other studies were able to conduct self-assessment after the prior practice with peer-assessment. For example, Matsuno (2007) investigated both peer-assessment and self-assessment of L2 essays conducted at the same time. Through the peer-assessments, the students might have developed more realistic expectations of their essays in relation to their peers, as well as increased accuracy through practice assessing their classmates' L2 essays. Thus, Matsuno's participants were able to assess their essays more objectively than the students in this study,

who assessed their own L2 speaking performance immediately after completing the oral interview, and no comparison with peers was made. In Patri's (2002) study, the participants evaluated their oral presentation after they completed peer evaluations and discussions and in Sullivan and Hall (1997), the participants evaluated their own grades after analyzing their own work according to a guide. However, the participants in this study evaluated their own L2 performance immediately after they finished their oral performance. In comparison to other studies in which participants were given more time to think about their performance objectively, their self-evaluation might have been more influenced by their immediate subjective feelings. If the participants could not speak very well during the interview, their negative feelings about their performance might have caused them to evaluate themselves lower than necessary. In fact, after the interview, some students who could not speak English very well were hanging their heads or were saying things such as, “うわあ、俺の英語力、最悪! (*Uwa, oreno eigo ryoku, saiaku!*)”, which means “Oh, my English ability is so poor!” On the other hand, those who were able to speak some English might have felt more satisfied with their own performance and this feeling led them to evaluate themselves higher.

The second reason why Japanese participants did not show Dunning-Kruger Effect might be that the degree of Japanese students' self-esteem is different from that of western people; that is, Japanese tend to have lower self-esteem that leads to self-criticism instead of self-enhancement (Heine et al., 1999; Markus & Kitayama, 1991). The notion of the Dunning-Kruger Effect was developed by investigating

American students whose self-esteem is said to be higher than Japanese and who tend to evaluate themselves higher than others (Heine et al., 1999). Indeed, some researchers such as Brown (2006) showed that American participants' self-evaluations were higher than those of Japanese participants. Therefore, the Dunning-Kruger Effect may not apply to Japanese participants with lower self-esteem. In other words, Japanese participants with low proficiency do not necessarily show the Dunning-Kruger Effect and do not always overestimate their abilities as found in this study. Similarly, Matsuno's (2007) Japanese university students with low proficiency did not overestimate their L2 writing skills, either. However, there is little research investigating the differences in self-assessment of L2 by Japanese university students at different levels of proficiency, so further research is needed to confirm this finding.

Finally, when Pearson correlations were calculated between self-assessment and teacher-assessment, the correlations were significant and showed medium-sized positive relationships ($r = .37$ to $.47$). In the preliminary analysis, I tested six students with the speaking assessment rubric, and I asked them if they could understand the categories. They all understood the concept of the Grammar, Vocabulary, and Pronunciation categories, but half of them were not familiar with Fluency. After I explained fluency using the keywords such as pauses and hesitation and they read the descriptions of the rubric, they understood the concept of fluency well. Therefore, during the main study, before the participants were conducting the self-assessment, I briefly explained all the four categories and asked

them if they understood the concept of each category and after I made sure that they understood the concept, they conducted self-assessment. However, understanding the concept and conducting accurate self-assessment might be different. Even after the self-assessment training, the low proficient participants in Patri (2002) commented that they could not judge pronunciation and they could not identify grammar mistakes. Similarly, the participants of this study who were not used to speaking English could not evaluate themselves accurately. Additionally, this result might be partly due to the low agreements among five teachers, for their correlations were not very high (.49 to .80) and Rater 2 and Rater 3 were not significantly correlated (.28).

Compared with other studies, the correlation coefficients between teacher-assessment and students-assessment found in this study are larger than those reported by Trofimovich et al. (forthcoming) (.06 and .18), Pierce et al. (1993) (.12 to .19), and Jafarpur's (1991) senior students (.27). However, they are lower than those reported by MacIntyre et al. (1997) (.60 to .63), Patri (2002) (.48 to .50), Jafarpur (1991) (.53), and AlFalla's (2004) low self-esteem students (.85). If the correlations are interpreted as indicators of inter-rater reliability, the student self-assessments in this study cannot be considered reliable. However, it should be noted that except for two studies, Trofimovich (forthcoming) and AlFalla (2004), the researchers above did not include multiple raters but only utilized one teacher-rater, so the reliability of teacher-assessment of those studies are doubtful.

The self-assessment measures in this study are likely not reliable probably because the participants did not receive self-assessment training due to the purpose of the study, which was to examine self-assessment bias and determine whether bias size was influenced by affective variables. The following two studies reported the beneficial effects of self-assessment training for accurate self-assessment. First, Chen (2008) noted the importance of self-assessment training. In Chen's study, 28 Chinese students evaluated their own oral presentations and their self-assessments with and without training were compared to the teacher assessments. The results showed that at first the correlation between self-assessment and teacher-assessment was $.55$ ($p < .05$), but the correlation increased to $.79$ ($p < .05$) after the students received the self-assessment training. Kruger and Dunning (1999) also reported that after training in which the participants were given the peers' performance on the tests, high proficiency participants learned how poorly their peers performed and no longer underestimated their own abilities, while low proficiency participants realized their incompetency and conducted self-assessment more accurately than others who did not receive training. Kruger and Dunning (1999) maintained that through meditational analysis, they improved their metacognitive skills.

Unlike Chen (2008) and Kruger and Dunning (1999), Patri (2002) failed to show any beneficial effects of self-assessment training. Although 56 Chinese students received the two-hour assessment training on L2 oral presentation, the correlations between teacher-assessment and self-assessment were medium sized ($r = .50$). Even when the students in the experimental group were given peer

feedback, the correlation between self-assessment and teacher-assessment did not improve and was much lower ($r = .48$) than the correlation between peer-assessment and teacher-assessment ($r = .85$). Thus, although students received assessment training and peer feedback, the correlation between self-assessment and teacher-assessment was not very high.

Both Chen (2008) and Parti (2002) investigated Chinese university students, but the results of self-assessment training were different. When the comments of their participants were compared, a large difference was found between the two studies. In Chen (2008), during the first self-assessment, the students' comments about their own L2 presentation were mainly negative because they tended to ignore their merits but focused on their shortcomings. However, after the training, their negative comments reduced from 67% to 31%, and made more neutral and positive comments, which led to the increase in the correlations between self-assessment and teacher-assessment. In contrast, in Parti (2002) even though the participants received positive comments on their presentation from the peers such as "Your eye contact improved a lot" and "You are much better prepared this time," the participants did not believe the peers' positive feedback and commented, "They cannot identify some of my mistakes and say I'm good" and "Some of them are too subjective." Unlike the participants in Chen (2008), they could not recognize the good points of their own performance properly even after receiving the peer-feedback, so the correlations between self-assessment and teacher-assessment did not improve in the second self-assessment. Patri (2002) explained that the reasons

for inaccurate self-assessment are probably due to very low proficiency of the participants because they were from a remedial English class and had very little experience being autonomous learners. However, I assume that not only low proficiency but also low self-esteem is the reasons for inaccurate self-assessment. The participants were too critical toward their L2 presentation, and this self-criticism is said to be a tendency for East Asian people with low self-esteem (Heine et al., 1999). Probably due to their low proficiency and low self-esteem, self-assessment training did not work like Chen's (2008) participants who came to focus on good points rather than shortcomings of their own L2 presentation. Likewise, Japanese participants in the present study did not have much experience studying speaking English, so most of them were not competent English speakers and thus they could not evaluate their own speaking performance accurately. Moreover, because of their lower self-esteem, they became self-critical toward their L2 performance and rated their performance lower than teachers. Due to their low proficiency and lower self-esteem, self-assessment training might not work well unless they are taught to avoid being too self-critical and see their own skills more positively.

Research Question 2: The Influence of L2 Affective Variables on Self-Assessment

The second research question is, "Which seven affective variables, Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward

Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence, predict the students' self-assessment of L2 oral performance?"

The results of the Pearson correlations indicated that six affective variables correlated significantly with self-assessment measures and teacher-assessment measures. Among the six variables, L2 Speaking Self-Confidence correlated most strongly with both self-assessment measures ($r = .41, p < .01$) and teacher-assessment measures ($r = .39, p < .01$). Thus, those who had higher self-confidence in L2 speaking tended to assess their own L2 performance higher and tended to have higher L2 speaking proficiency. Only Self-Esteem did not correlate with self-assessment and teacher-assessment measures. Self-Esteem does not appear to have a strong relationship with students' L2 speaking proficiency nor with the self-assessment measures. The results of the correlations showed that with the exception of Self-Esteem, affective variables appear to have strong relationships with actual L2 proficiency and self-assessment measures. As reviewed in the Literature Review section, many researchers revealed significant relationships between L2 proficiency and affective factors, and the results of this study also confirmed the previous findings that high proficiency learners are more motivated than low proficiency learners. However, though some researchers reported that higher self-esteem is related to better language learning (Liu, 2012; Ortega, 2007; Shirahata et al. 1999), this relationship was not found in this study; There were no significant correlations between Self-Esteem and L2 speaking proficiency measures and also it was found

that high and low proficiency students possessed the same degree of Self-Esteem (Table 72).

Unlike self-assessment and teacher-assessment measures that were correlated with six affective variables, bias size measures, which represent the difference between the self-assessment measures and the teacher-assessment measures that are adjusted for teacher severity, were significantly correlated with only one variable, Desire to Learn to Speak English. Many previous studies have reported significant relationships between affective variables and L2 oral proficiency and between affective variables and self-assessment of L2 speaking, so the results of this study that both teacher-assessment measures and self-assessment measures were significantly correlated with affective variables are in accordance with the previous findings. On the other hand, so far no researchers have examined the relationship between the bias size measures of self-assessment of L2 speaking performance and affective variables, so the results of this study cannot be compared with other results. The possible explanation for the findings that the bias size measures did not correlate strongly with affective variables might be that the self-assessment conducted by the participants was inconsistent and unreliable. However, further research is needed to confirm the finding.

Although Desire to Learn to Speak English was positively correlated with self-assessment and teacher-assessment measures, it was negatively correlated with the bias size measures, implying that on average, those with a stronger Desire to Learn to Speak English tended to rate their performances slightly lower than the

teachers did, regardless of their overall ability relative to their peers. Therefore, it was found that different from self-assessment, bias size measures did not have much relationship with many affective variables. This result seems to indicate that the use of self-assessment measures is not appropriate if we want to investigate the bias size of self-assessment.

With regard to each category, higher Desire to Learn to Speak English was associated on average with higher teacher-assessment measures with all categories and self-assessment measures with the exception of grammar, where the correlation was not significant (Table 62). However, there was a negative correlation with the bias measures in all categories (Table 62). This implies that learners with higher Desire to Learn to Speak English tend to be better performers, but are cautious in their self-assessments; they tend to underrate themselves from the perspective of the teachers. Although the differences are small, the negative correlations were stronger between bias and Vocabulary (-.16) and bias and Grammar (-.15) than between bias and Fluency (-.13) and bias and Pronunciation (-.11). Perhaps students with a stronger Desire to Learn to Speak English view Grammar and Vocabulary as important aspects of a speaking performance that they need to improve in order to achieve their goals of being highly competent speakers. Their frustration might lead them to rate these categories more severely than teachers. Higher self-assessment and lower bias for Fluency and Pronunciation might be a result of their ongoing efforts to use English outside of class due to their higher Desire to Learn to Speak English. They might possess more confidence in these

categories, or they might view them as relatively less important than grammar and vocabulary in generating a strong speaking performance.

L2 Speaking Anxiety and L2 Speaking Self-Confidence had negative and positive correlations, respectively, with both teacher-assessment and self-assessment (Table 62). Their effects on bias were small, however. L2 Speaking Anxiety showed a correlation of $-.11$ with Pronunciation. This finding might be of interest to anxiety researchers as it would appear that anxious learners not only perform more poorly from a teacher's perspective, but they might further add to their anxiety by perceiving their pronunciation more severely than teachers. L2 Speaking Self-Confidence showed a correlation of $.11$ with Pronunciation, so L2 Speaking Self-Confidence produced a slight positive bias for Pronunciation.

Before testing the hypothesized structural modeling, stepwise multiple regression was conducted using the seven affective variables as predictors and teacher-assessment measures as a dependent variable. The results indicated that L2 Speaking Self-Confidence and Desire to Learn to Speak English predicted the total teacher-assessment measures as well as teacher-assessment measures in all four categories. Thus, students with greater L2 Speaking Self-Confidence and greater Desire to Learn to Speak English are likely to have better English speaking abilities. In many previous studies, L2 Speaking Anxiety had the strongest relationship with speaking skills (e.g., Hewitt & Stephenson, 2012; Phillips, 1992; Young, 1986), but in this study, L2 Speaking Self-Confidence was the strongest predictor. This result is consistent with Clément, Dörnyei, and Noels (1994), who

found the strongest relationship between self-confidence and speaking skills.

Desire to Learn to Speak English also predicted speaking proficiency and this result is associated with Gardner's (1985) socio-educational model, in which motivation leads to language learning outcomes (achievement).

Next, the hypothesized structural model for bias size shown in Figure 8 was tested; however, it showed very poor fit to the data, most likely due to the unreliability of the self-assessment. In Research Question 1, the results of the Rasch analysis indicated that unlike teacher raters, student raters misfit the model and were inconsistent in their self-assessment; 95 out of 100 of the most unexpected scores were the students' self-assessments. Thus, the self-assessment was too unreliable to use as an endogenous variable.

Following model testing, a multiple regression analysis was conducted using affective variables as predictors and bias size measures as dependent variable. The R^2 for the bias size measures was smaller than those for teacher-assessment, but the results revealed that Desire to Learn to Speak English negatively and L2 Speaking Self-Confidence positively predicted the bias size measures of students' self-assessment. Therefore, students who had a greater Desire to Learn to Speak English tended to underestimate their own speaking performance, while those with greater L2 Speaking Self-Confidence tended to overestimate their oral performance.

Interestingly, one variable, L2 Speaking Self-Confidence positively predicted both the teacher-assessment measures and bias size measures. Therefore, those with greater confidence in speaking English tended to be better L2 speakers and tended

to overestimate their own L2 performance. This result is in accordance with the Facets results shown in Figure 26, where those who used higher categories (7 to 9) estimated their own performance slightly higher than teacher-raters.

On the other hand, Desire to Learn to Speak English had a positive relationship with teacher-assessment measures and a negative relationship with bias size measures, implying that those with greater Desire to Learn to Speak English tended to be better English speakers, but they tended to underestimate their own L2 oral performance. In some previous studies, motivation was positively correlated with self-evaluation. For example, Noels, Clément, and Pelletier (1999) found that Amotivation was associated with lower self-evaluation, while Intrinsic Motivation led to higher self-evaluation. Masgoret and Gardner (2003) also reported that self-ratings of L2 achievement and Motivation displayed the largest correlation (.39). What should be noted here is that these researchers utilized self-assessment scores rather than bias size measures. In fact, Desire to Learn to Speak English was positively correlated with self-assessment measures in Table 62, but negatively correlated with the bias size measures in Table 63. This result seems to indicate that the bias size measures are different from self-assessment measures, so if we want to investigate the bias of self-assessment, we need to use the bias size measures, or we obtain different results. The possible explanation for the negative correlation between Desire to Learn to Speak English and the bias size measures might be that those students feel that they need to improve their speaking proficiency because

they think that they are not good at speaking English. So probably their low self-evaluation of L2 speaking skills might enhance their desire to be better L2 speakers.

With regard to each category, Desire to Learn to Speak English negatively predicted all four categories, suggesting that those who were willing to improve their speaking skills tended to underestimate all the categories, hoping that they would improve those skills. On the other hand, Self-Esteem positively predicted Grammar. I assume that grammar mistakes might be difficult to be recognized while speaking L2, especially novice speakers without much knowledge of grammar. Without much grammar knowledge, L2 speakers might not be able to distinguish accurate and inaccurate sentences, or simple and complex sentences. Indeed, Patri's (2002) participants commented that they could not identify their own grammar mistakes of their L2 presentation. Thus, after the L2 interview, those with higher self-esteem who had positive view of themselves evaluated their grammar abilities higher, while those with lower self-esteem who had negative view of themselves evaluated lower. Compared with grammar, lack of vocabulary knowledge might be easier to be recognized by the speakers. Because when the participants could not come up with the appropriate words, they often paused and hesitated, and thus they could not continue their speech smoothly (See Appendix O). Because L2 Speaking Anxiety was negatively correlated with vocabulary (Table 62), those with greater L2 Speaking Anxiety were likely not good at vocabulary, so they tended to have many pauses and silences during their speech.

This experience generated greater anxiety, which might have led them to evaluate their vocabulary skills lower than necessary.

L2 Speaking Self-Confidence was a predictor of the actual fluency, so those who spoke English relatively fluently tended to evaluate their own fluency higher. Pronunciation bias measures were predicted negatively by L2 Speaking Anxiety and positively by L2 Speaking Self-Confidence, so did actual pronunciation skill (Tables 62 and 63). Therefore, those who had confidence in their own L2 pronunciation tended to be better at pronunciation and they had a positive bias for self-assessment, while those who had greater anxiety in L2 pronunciation tended to pronounce poorly and they evaluated their pronunciation lower.

Research Question 3: Characteristics of Those Who Conducted Accurate Self-Assessment

The third research question was, “What are the characteristics of the students who could assess their own L2 oral performance accurately?” In order to answer this question, 106 students whose ratings were similar to the teacher ratings were compared with other students in terms of L2 proficiency and affective variables.

First, high proficiency people are said to be better able to evaluate themselves than low proficiency people (Kruger & Dunning, 1999). However, this tendency was not recognized in this study. The mean measure of L2 oral proficiency for accurate students was slightly lower than average (-.13), so the L2 speaking performance of the accurate group students was not necessarily good. Indeed, the

mean measure of inaccurate students was slightly higher than that of accurate group (.06), though no significant difference was found between them.

Second, the affective variable measures of accurate and inaccurate students were compared. In the L2 literature, some researchers showed the influence of affective variables on self-assessment accuracy. Anxiety was reported to have a negative bias on self-assessment (Gardner & MacIntyre, 1993; MacIntyre et al. 1997), the strongest path was shown from self-confidence to self-ratings of speaking in the structural equation model (Clément et al., 1994), and low self-esteem learners and integratively motivated learners had the highest correlations between their self-assessments and teacher-assessments of L2 presentation (AlFallay, 2004). Unlike the results of the previous studies, no significant differences in affective factors were found between the accurate and inaccurate students. Since no distinctive features for the accurate students were found, it is possible that their accurate self-assessments could have occurred by chance alone.

The participants of the present study were mostly those who have never studied abroad and have studied English at Japanese secondary schools where grammar and reading are focused rather than oral communication; therefore, many of them rarely had chance to speak English in their everyday life. Therefore, they might not possess the skill to evaluate their L2 oral performance. For example, Chinese students in the EFL context conducted self-assessment of L2 presentation and commented that they “could not judge pronunciation” and “could not identify mistakes in grammar” (Patri, 2002). Kruger and Dunning (1999) also mentioned

the difficulty of self-assessment because low proficiency people cannot evaluate their own skills accurately due to their incompetency. Thus, the participants in this study were not very skilled in speaking English, so they could not evaluate their speaking performance accurately. As Peirce et al. (1993) pointed out, many EFL learners have little access to L2 outside the classroom and thus do not possess a native speaker peer standard with which to compare their own L2 proficiency. Consequently, they often fail to assess their performances accurately.

Research Question 4: The Influence of L2 Affective Variables on Self-Assessment of High and Low Proficiency Students

The fourth research question asked, “To what degrees do the seven affective variables affect the self-assessment of high and low proficiency students differently?” The top 100 students were selected for the high proficiency group, while the bottom 100 students belonged to the low proficiency group. The affective variable measures and the bias size measures were compared between the two groups. It was found that high proficiency learners had greater L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence, and lower L2 Speaking Anxiety than low proficiency students, but no significant difference was found for Self-Esteem. This result is in accordance with many previous studies that reported motivated L2 learners have better proficiency and high proficiency learners are less anxious than low proficiency learners (e.g.,

Gardner & MacIntyre, 1993; Horwitz, 1999), but it is different from some researchers who found higher self-esteem is related with high proficiency of L2 skills (e.g., Liu, 2012).

The regression analysis showed that Self-Esteem and L2 Speaking Self-Confidence positively and Desire to Learn to Speak English negatively predicted the bias size measures of high proficiency students. Therefore, high proficiency students with higher Self-Esteem and greater L2 Speaking Self-Confidence tended to overestimate their L2 speaking, and those with greater Desire to Learn to Speak English were likely to underestimate their L2 oral performance. Desire to Learn to Speak English negatively influenced the self-assessments conducted by the high proficiency students. The underestimation related with Desire to Learn to Speak English seems positive because the feeling that they are poor at speaking English is related to the desire to improve it. Those motivated learners might pay a sufficient amount of attention toward their speech, so they can enhance their proficiency.

In Research Question 1, we discovered the tendency of the high proficiency students to overestimate their L2oral proficiency. Self-Esteem appears to be a factor in this overestimation. This result is the same as that reported by AlFallay (2004), who investigated 78 Arabic university students and discovered that the high self-esteem group tended to overestimate their L2 oral ability. However, Japanese people are often said to have lower self-esteem than people from other countries, especially westerners (R. A. Brown, 2006; Heine et al. 2001). Moreover, they traditionally value modesty (R. A. Brown, 2005), so they tend to be self-critical and

self-effacing. Thus, I predicted that the students' low self-esteem would probably lead to underestimations of their L2 oral performance. However, the results of this study indicated that Japanese students with high self-esteem and high proficiency tended to overestimate rather than underestimate their own L2 performance.

Although no significant difference in self-esteem was found between high and low proficiency groups, self-esteem positively affects bias measures ($\beta = .27$) for high proficiency students (Table 73), while self-esteem did not significantly influence bias measures for low proficiency students (Table 75). Thus, it appears that high self-esteem leads to overestimation only when high proficiency students rated themselves.

With regard to each category, those with a greater Desire to Learn to Speak English assessed the Grammar, Vocabulary, and Pronunciation categories lower because they think that they need to improve these skills in order to speak English more proficiently. High proficiency students tended to have greater Desire to Learn to Speak English than low proficiency students and high proficiency students' greater "desire to improve their speaking skills" might be engendering higher levels of expectation. In contrast, low proficiency students' self-assessment bias measures were not predicted by Desire to Learn to Speak English. Therefore, the higher proficiency students seem to be holding themselves to a higher standard than the lower proficiency students because of their greater desire to improve L2 speaking. Therefore, they tended to underestimate their L2 oral performance.

L2 Speaking Anxiety led to underestimations in Fluency and Pronunciation, but not Grammar and Vocabulary. Because they are high proficiency learners, they are relatively better at grammar and vocabulary than other students, because English education in Japan is often focused on the study of these skills. This might be why L2 Speaking Anxiety did not lead to underestimation of their knowledge of English grammar and vocabulary; rather, L2 Speaking Self-Confidence led them to overestimate their grammar and vocabulary ability.

Underestimation related with L2 Speaking Anxiety of high proficiency students seems negative because their anxiety toward speaking English made them think that their fluency and pronunciation skills are low, even though they are better at these skills than others. Many studies on L2 Anxiety have reported that high proficiency learners tend to have lower L2 Anxiety than low proficiency learners (e.g. Phillips, 1992; Young, 1986). This study also found that high proficiency students had significantly lower anxiety than low proficiency students. Even though high proficiency students could speak English better than other students and had lower anxiety than low proficiency students, their anxiety made them evaluate their fluency and pronunciation skills lower than expected. Thus, although Brown et al. (2001) found facilitating anxiety in Japanese university students, the anxiety felt by high proficiency students of this study seems debilitating, for even such a low level of speaking anxiety made them believe that their fluency and pronunciation are poorer than how it is perceived by teachers.

Self-Esteem led to overestimation of three categories, Grammar, Vocabulary, and Pronunciation. In particular, Self-Esteem had the strongest influence on Pronunciation. Apparently, feeling good about themselves led them to assess their pronunciation higher than teachers. This is probably because those who study English in the EFL context often learn vocabulary from reading and sometimes they do not know the correct pronunciation of English words, and consequently they fail to assess their own pronunciation accurately. Because those with higher self-esteem had positive views of themselves, their positive views might have led them to assess their own English pronunciation better than they actually were.

Next, the 100 students who had the lower Rasch person measures for L2 speaking proficiency than other students were categorized as low proficiency students and were examined. The stepwise multiple regression analysis revealed that bias sizes measures for the total proficiency as well as all the categories were predicted by L2 Speaking Self-Confidence only. When Japanese students evaluate their own L2 performance, those who can speak English better are likely to be influenced more by psychological factors than those who speak less well. In Research Question 1, it was discovered that the low proficiency students tended to underestimate their speaking abilities. During the interview, the low proficiency speakers had difficulty producing English sentences and accessing appropriate lexis, and they often had very long silences. For these reasons, they must have felt that their speaking performance was very poor and consequently evaluated themselves very low. Only those with higher self-confidence could overcome their

negative feelings about their own speaking performance and overestimate their L2 speaking skills.

Theoretical Implications

Three theoretical implications are suggested by the results of the study. First, most studies of self-assessment of L2 skills have been focused on investigating the validity of self-assessment by calculating correlations with teacher-assessment (Jassen-van Dieten, 1989; Le Blanc & Painchaud, 1985), while the influence of affective variables on self-assessment bias has not yet been fully researched. Some researchers who have investigated this issue only dealt with one or two variables; for example, anxiety in MacIntyre et al. (1997) and anxiety and WTC in Liu and Jackson (2008). This study filled this gap by investigating the influence of seven affective variables (e.g., Self-Esteem, L2 Speaking Anxiety, L2 Willingness to Communicate, Attitude Toward Learning to Speak English, L2 Speaking Motivational Intensity, Desire to Learn to Speak English, and L2 Speaking Self-Confidence) on self-assessment of L2 speaking performance. By using the many-facet Rasch measurement model, students' speaking performance, rater severity, and category difficulty were placed on the same scale; this feature provides informative comparisons among them. Additionally, the influence was examined closely by looking at subskills of speaking skills, grammar, vocabulary, fluency, and pronunciation.

Second, using Facets, I calculated bias measures of self-assessment. Past researchers investigating self-assessment have not utilized bias size Rasch measures, but have instead relied on self-assessment raw scores. However, it was discovered in this study that the use of bias size measures obtained different results from those that used self-assessment measures. For example, self-assessment measures were correlated with six affective variables, while bias measures were correlated with only two affective variables. Moreover, Desire to Learn to Speak English was positively correlated with self-assessment measures, but negatively correlated with Rasch bias size measures. Consequently, it is recommended that future researchers investigating self-assessment consider the use of bias size measures in place of raw self-assessment scores.

Finally, self-assessment plays an important role in L2 acquisition. In monitor hypothesis proposed by Krashen (1982), L2 learned rules are considered to serve as a monitor of utterances initiated by the acquired system of rules. However, the results of this study showed that participants who mainly studied grammar and vocabulary at secondary schools failed to evaluate these skills accurately. Such an inaccurate self-assessment may fail to serve as the monitor of their utterances and hinder acquisition. Moreover, in the noticing hypothesis Schmidt (1990) argues that noticing a gap is essential for L2 acquisition because noticing is necessary for input to become intake, but inaccurate self-assessment might prevent L2 learners from noticing their gaps. This is probably because Japanese university students have not had much experience of L2 output, so engaging in more L2 output may help them

to realize their own speaking performance better. For example, the interaction hypothesis proposed by Long (1981) suggests that L2 development is promoted by communication through face-to-face interaction; When L2 learners make some utterances and the interlocutors do not understand them, L2 learners try negotiation of meaning through which they are likely to notice which aspects they have problems with, such as grammar, vocabulary, and pronunciation. Additionally, Swain's (1985) output hypothesis states that when L2 learners make efforts to make their messages comprehensible, through negative feedback of their output, they notice a gap between their actual utterances and those of proficient speakers, which fosters L2 acquisition. Thus, although early SLA such as Krashen's input hypothesis put more emphasis on L2 input, L2 output seems important to improve self-assessment skills, for a lot of L2 output might help L2 learners notice a gap and enable them to evaluate their own L2 skills more accurately, which will enhance L2 acquisition.

Pedagogical Implications

Three pedagogical implications can be suggested by the results of the study. First, some researchers examined the participants' L2 speaking skills using only one teacher-rater (e.g., Chen, 2008; Jafarpur, 1991; MacIntyre et al., 1997; Peirce et al., 1993), but the accuracy of the teacher-assessment of these studies is doubtful. In fact, the five raters in this study displayed a great deal of diversity and had unique bias patterns. The correlations for 50 students were medium-sized ($r = .49$

to .80, $p < .01$), and Rater 2 and Rater 4 were not significantly correlated ($r = .28$, $p > .05$). This result suggests that multiple raters should be employed and Facets should be used, especially in high stakes speaking tests, because it can provide estimates of ability that are adjusted for rater bias.

Second, the results suggest that self-assessment conducted without prior training and evaluation is not sufficiently reliable or consistent to be used as an assessment tool. Therefore, teachers should not assume that student self-assessments of L2 oral performance are equivalent to actual speaking abilities. Moreover, as most students in this study assessed their L2 speaking more severely than the teacher raters, teachers might consider telling their students that their speaking skills are better than they think, which might encourage them to speak English more. In order to introduce self-assessment successfully in foreign language classrooms, teachers should conduct self-assessment training. For example, Chen (2008) conducted assessment training in which students created a set of criteria and scoring standards for assessing oral performance, they practiced assessing videotaped performances using the rubric they created, and finally they assessed their peers' oral presentations after which the students were engaged in group discussions to talk about their assessments. Teacher feedback was also provided on their assessments. This practice improved the correlation between self- and teacher-ratings. However, Patri's (2002) participants failed to conduct accurate self-assessment even after the self-assessment training probably because they failed to focus on their own positive aspects. Especially, East Asian people including

Japanese are said to have a tendency to be self-critical (Heine et al., 1999). Thus, if teachers are considering including self-assessment exercises in their speaking classes, they should include self-assessment training in which students should be trained to focus more on their positive points than shortcomings; otherwise, students' self-assessments are not likely to be reliable or consistent.

Finally, in addition to the explicit self-assessment training, speaking teachers should provide L2 learners with more opportunities for L2 output. As the interaction hypothesis (Long, 1981) and the output hypothesis (Swain, 1985) suggest, when L2 learners engage in L2 output, through negotiation of meaning and positive or negative feedback, they are able to notice a gap between their own productions and those of proficient speakers, which will improve the learners' abilities to evaluate their own speaking performance. Accurate self-assessment can be very important for L2 acquisition because noticing their own strengths and weaknesses will help learners improve their speaking skills. Many Japanese university students have mainly studied grammar and reading at secondary schools and rarely had much chance to speak English. Therefore, the results of this study that their self-assessments were unreliable and inconsistent seem reasonable. Therefore, providing them with not only the explicit self-assessment training, but also more opportunities to speak English will help them to evaluate their L2 oral performance more accurately.

CHAPTER 7

CONCLUSION

In this chapter, I summarize the findings of the study. Following the summary, I describe the limitations of the study, and make suggestions for future research.

Summary of the Findings

In this study, I investigated the validity of the students' self-assessments of L2 oral performance and the influences of affective variables on self-assessment bias.

First, it was found that the participants rated their own L2 speaking skills more severely than the teacher raters and that the students tended to be more severe than the teachers when they used lower categories (1 to 4), but more lenient when they used higher categories (7 to 9). Moreover, 95 out of the 100 most unexpected scores were student self-assessments. The Pearson correlations between self-assessment and teacher assessment showed medium relationships ($r = .37$ to $.45$, $p < .01$). Therefore, students' self-assessments were neither reliable nor consistent.

Second, Rasch bias measures for self-assessment were calculated and a hypothesized structural model was tested. The model did not fit the data, most likely due to the unreliability of the self-assessment measures. Following model testing, multiple regression analyses were conducted using affective variables as

predictors and bias size Rasch measures as the dependent variables. The results indicated that the students with greater Desire to Learn to Speak English tended to underestimate and those with greater L2 Speaking Self-Confidence tended to overestimate their own L2 speaking performance.

Third, 106 students whose ratings were similar to teachers were selected as an accurate group and their distinctive features were examined. Using the *t*-test, they were compared with other students in terms of L2 proficiency and affective variable measures. No significant differences were found between the two groups, so it was concluded that those whose self-assessments agreed with teachers could have resulted in some agreements that occurred by chance alone.

Finally, 100 high proficiency students were compared with 100 low proficiency students in terms of the influences of affective variables on their bias size measures. The results of a stepwise multiple regression analysis indicated that higher proficiency students with higher Self-Esteem and greater L2 Speaking Self-Confidence tended to overestimate and those with greater Desire to Learn to Speak English were likely to underestimate their L2 oral performance, and that lower proficiency students with greater L2 Speaking Self-Confidence tended to overestimate their L2 speaking proficiency.

Limitations of the Study

There are three limitations in this study. First, when I asked the participants to volunteer for this study, some students refused to participate in the speaking

interviews because they said that they could not speak English well and that they did not want their poor speaking performances to be recorded. These students were not necessarily poor at speaking English because some of them belonged to advanced English classes. They probably had a great deal of anxiety toward speaking English. I could not force them to participate in this study, so the data I gathered were from the students who agreed to participate in the interviews.

However, including those students who refused to have speaking interviews might have changed some of the results of this study, especially for L2 Speaking Anxiety.

The second limitation is that the individual oral interviews were conducted by me, a female Japanese teacher. Most of the students were familiar with me, so their attitudes might have been different had they been required to speak English to a stranger; for example, their anxiety might have been higher. In addition, when the students could not access appropriate English vocabulary, they sometimes used Japanese because they knew that I could understand them. Consequently, their spoken data would have been different had they been required to speak English to a native English speaker who does not understand Japanese.

Finally, in this study the hypothesized structural model for L2 speaking self-assessment bias failed to fit the data. Although the hypothesized model was constructed based on past studies, the model probably failed to fit, due to the unreliability and inconsistency of self-assessments conducted by students who did not receive self-assessment training. Moreover, in the third research question, I could not figure out the distinctive characteristics of the participants who could

evaluate themselves similar to teacher-ratings. Thus, I pointed out the possibility that their agreements might have occurred by chance alone. However, including other variables such as aptitude and personality or qualitative data might have helped to find the different answer to the third research question.

Suggestions for Future Study

First, this study included only analyses of quantitative data. Including a qualitative dimension might produce further results. When I conducted the interview tests, some participants showed great emotional disturbance because they felt that they could not speak English well enough, while others said they enjoyed speaking English so much that they wanted to be interviewed again. These differences in attitudes were seen regardless of their speaking proficiency. Therefore, in future research it might be informative to ask the participants to relate what they felt during the English interview immediately after the interview.

Second, this study failed to find out the answer to the third research question that investigated the distinctive characteristics of those who evaluated their L2 oral performance accurately. Their L2 oral proficiency and affective variable measures were compared with those of the students whose self-assessments were less accurate. However, no significant differences were found between them. In the future study, including other variables such as aptitude and personality might help to find the answer to this question. Moreover, conversational experience in EFL contexts might be a good predictor of accurate self-assessment skills. As a future

study, I could ask participants to report whether and to what degree they have used English with native and non-native speakers on a daily basis, and examine if the amount of L2 conversational experience makes an impact on their self-assessment patterns.

Finally, as the purpose of this study was to examine self-assessment bias, self-assessment training was not conducted. It was consequently discovered that the self-assessments conducted by the untrained students were unreliable and inconsistent. It would be useful to investigate what the results would be if the students received self-assessment training. In fact, there were contradictory findings in studies that conducted self-assessment training. Chen (2008) reported increased correlations between self- and teacher-assessments, while Patri (2002) failed to show beneficial effects of self-assessment training. As these studies calculated Pearson correlations between self- and teacher-assessed scores, more detailed information might be obtained if training were conducted, Rasch bias measures calculated, and differences between self- and teacher-assessments examined both before and after the training.

Final Conclusions

In this study the validity of self-assessment of L2 oral performance and the effects of affective variables on self-assessment bias were investigated. Although the results showed that self-assessment was unreliable, the act of self-assessment is important for language learners. By assessing their own language skills, learners

can recognize their own strengths and weaknesses and become more autonomous learners. However, as the results indicated that most Japanese students assessed their own L2 performance more severely than teacher raters, it is advisable that before introducing self-assessment activities in a language class, teachers should tell students their tendency to underestimate their speaking abilities, especially in fluency, in order to be able to better monitor their own language abilities.

There are still few studies investigating Japanese students' self-assessments of speaking skills. However, many benefits have been reported for self-assessment (e.g., Brown & Hudson, 1998; Kusnic & Finley, 1993; Oscarson, 1989; Todd, 2002) and the importance of English speaking skills have been recognized due to globalization. As Japan is an EFL context, Japanese L2 learners have few opportunities to speak English outside classrooms, unlike L2 learners in ESL contexts who can use their L2 in daily life and can monitor their own speaking skills every day. If Japanese L2 learners learn how to assess their own speaking proficiency, perhaps by becoming better informed about their biases, they can better gauge their own progress. Further research on the self-assessment of speaking skills will, I hope, help Japanese students improve their oral proficiency.

REFERENCES

- Abiko, T. (1997). *Jikohyooka de jyugyo ga kawaru. (Self-assessment changes classes)*. Tokyo: Meiji Tosho Shuppan.
- Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In J. Firth (Ed.), *Measuring spoken proficiency* (pp. 1-6). Washington, DC: Georgetown University Press.
- Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal*, 78, 155-168. doi:10.1111/j.1540-4781.1994.tb02026.x
- AlFallay, I. (2004). The role of selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System*, 32, 407-425. doi:10/1016/j.system.2004.04.006
- Anderson, P. L. (1982). Self-esteem in the foreign language: A preliminary investigation. *Foreign Language Annals*, 15, 109-114. doi:10.1111/j.1944-9720.1982.tb00234.x
- Andrés, V. (1999). Self-esteem in the classroom or the metamorphosis of butterflies. In J. Arnold (Ed.), *Affect in language learning* (pp. 87-102). Cambridge: Cambridge University Press.
- Ávila, J. (2007). Self-esteem and second language learning: The essential colour in the palette. In F. Rubio (Ed.), *Self-esteem and foreign language learning* (pp. 68-90). Cambridge: Cambridge Scholars.
- Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, 6, 14-29. doi:10.1177/026553228900600104
- Baker, S. C., & MacIntyre, P. D. (2000). The role of gender and immersion in communication and second language orientations. *Language Learning*, 50, 311-341. doi:10.1111/0023-8333.00119
- Baumeister, R. F., Campbell, D. C., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, 4, 1-44. doi:10.1111/1529-1006.01431
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89-110. doi: 10.1191/0265532203lt245oa
- Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, 34, 15-35. doi:10.1016/j.system.2005.08.004
- Brown, Jonathan Dean. (1998). *The self*. Boston, MA: McGraw Hill.
- Brown, James Dean. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653-675. doi:10.2307/3587999
- Brown, J. D., Robson, R., & Rosenkjar, P. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 361-398). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.
- Brown, R. A. (2005). The paradox of Japanese self-esteem. *Information & Communication Studies*, 32, 1-12. Retrieved from <http://www.bunkyo.ac.jp/faculty/lib/slib/kiyo/Inf/if32/if3201.pdf>
- Brown, R. A. (2006). The relationship between self-aggrandizement and self-esteem in Japanese and American university students. *Information & Communication Studies*, 35, 1-16. Retrieved from <http://www.bunkyo.ac.jp/faculty/lib/slib/kiyo/Inf/if35/if3501.pdf>
- Burgoon, J. K. (1976). The unwillingness-to-communicate scale: Development and validation. *Communicative Monographs*, 43, 60-69. doi:10.1080/03637757609375916
- Burroughs, N. F., Marie, V., & McCroskey, J. C. (2003). Relationships of self-perceived communication competence and communication apprehension with willingness to communicate: A comparison with first and second language in Micronesia. *Communication Research Reports*, 20, 230-239. doi:10.1080/08824090309388821

- Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal*, 90, 506-518. doi:10.1111/j.1540-4781.2006.00463.x
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47. doi: 10.1093/applin/I.1.1
- Chen, Y-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, 12, 235-262. doi:10.1177/1362168807086293
- Cheng, Y. S., Horwitz, E. K., & Schallert, D. L. (1999). Language anxiety: Differentiating writing and speaking components. *Language Learning*, 49, 417-446. doi:10.1111/0023-8333.00095
- Clément, R. (1980). Ethnicity, contact, and communicative competence in a second language. In H. Giles, W. Robinson, & P. Smith (Eds.), *Language: Social psychological perspectives* (pp.147-154). Oxford: Pergamon Press.
- Clément, R. (1986). Second language proficiency and acculturation: An investigation of the effects of language status and individual characteristics. *Journal of Language and Social Psychology*, 5, 271-290. doi:10.1177/0261927X8600500403
- Clément, R., Dörnyei, Z., & Noels, K. A. (1994). Motivation, self-confidence, and group cohesion in the foreign language classroom. *Language Learning*, 44, 417-448. doi:10.1111/j.1467-1770.1994.tb01113.x
- Clément, R., Gardner, R. C., & Smythe, P. C. (1977). Motivational variables in second language acquisition: A study of Francophones learning English. *Canadian Journal of Behavioural Science*, 9, 123-133. doi:10.1037/h0081614
- Clément, R., & Kruidenier, B. (1985). Aptitude, attitude, and motivation in second language proficiency: A test of Clément's model. *Journal of Language and Social Psychology*, 4, 21-37. doi:10.1177/0261927X8500400102
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum.
- Dörnyei, Z. (2001). *Teaching and researching motivation*. Harlow: Longman.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. London: Erlbaum.

- Dörnyei, Z. (1990). Conceptualizing motivation in foreign language learning. *Language Learning, 40*, 45-78. doi:10.1111/j.1467-1770.1990.tb00954.x
- Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090. doi.apa.org/journals/psp/57/6/1082
- Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.
- Farh, J., Dobbins, G. H., & Cheng, B. (1991). Cultural relativity in action: A comparison of self-ratings made by Chinese and U.S. workers. *Personnel Psychology, 44*, 129-147. doi:10.1111/j.1744-6570.1991.tb00693.x
- Field, A. (2005). *Discovering statistics using SPSS*. Thousand Oaks, CA: Sage.
- Folse, K. S. (2007). *The art of teaching speaking: Research and pedagogy for the ESL/EFL classroom*. Ann Arbor, MI: The University of Michigan Press.
- Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitude and motivation*. London: Edward Arnold.
- Gardner, R. C. (2001). Integrative motivation and second language acquisition. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 1-19). Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.
- Gardner, R. C., & Lambert, W. E. (1959). Motivational variables in second language acquisition. *Canadian Journal of Psychology, 13*, 266-272. doi:10.1037/h0083787
- Gardner, R. C., & Lambert, W. E. (1972). *Attitudes and motivation in second language learning*. Rowley, MA: Newbury House.
- Gardner, R. C., & MacIntyre, P. D. (1993). A student's contribution to second language learning. Part II: Affective variables. *Language Testing, 26*, 1-11. doi:http://dx.doi.org.libproxy.temple.edu/10.1017/S0261444800000045
- Gardner, R. C., Tremblay, P. F., & Masgoret, A. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal, 81*, 344-362. doi:10.1111/j.1540-4781.1997.tb05495.x

- Giles, H., & Byrne, J. L. (1982). An intergroup approach to second language acquisition. *Journal of Multilingual and Multicultural Development*, 3, 17-40. doi:10.1080/01434632.1982.9994069
- Ginther A. (2012). Assessment of speaking. *The Encyclopedia of Applied Linguistics*, 1-7. doi: 10.1002/9781405198431.wbeal0052
- Gregersen, T., & Horwitz, E. K. (2002). Language learning and perfectionism: Anxious and non-anxious language learners' reactions to their own oral performance. *The Modern Language Journal*, 86, 562-570. doi:10.1111/1540-4781.00161
- Halbach, A. (2000). Finding out about students' learning strategies by looking at their diaries: A case study. *System*, 28, 85-96. doi:10.1016/S0346-251X(99)00062-7
- Heine, S. J., Kitayama, S., & Lehman, D. R. (2001). Cultural differences in self-evaluation: Japanese readily accept negative self-relevant information. *Journal of Cross-Cultural Psychology*, 32, 434-443. doi:10.1177/0022022101032004004
- Heine, S.J., Markus, H. R., Lehman, D. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review*, 106, 766-794. doi:10.1037/0033-295X.106.4.766
- Hewitt, E., & Stephenson, J. (2012). Foreign language anxiety and oral exam performance: A replication of Phillips's MLJ study. *The Modern Language Journal*, 96, 170-189. doi:10.1111/j.1540-4781.2011.01174.x
- Higgs, T., & Clifford, R. (1982). The push towards communication. In T.V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57-79). Lincolnwood, IL: National Textbook Company.
- Horwitz, E. K. (1991). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. In E. K. Horwitz & D. J. Young (Eds.), *Language anxiety: From theory and research to classroom implications* (pp. 101-108). Upper Saddle River, NJ: Prentice Hall.
- Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, 70, 125-132. doi:10.1111/j.1540-4781.1986.tb05256.x

- Hughes, R. (2003). *Teaching and researching: speaking*. Pearson Education ESL.
- Imai, H., & Yoshida, T. (Ed.). (2007). *Hope: Chuukosei no tameno eigo speaking test (Hope: Speaking test for secondary school students)*. Tokyo: Kyooiku Shuppan.
- Inoue, M. (1997). "Ikiru chikara" no ikusei to jikohyooka no houhou. (*Training of strength to live and how to conduct self-assessment*). Tokyo: Meiji Tosho Shuppan.
- Irie, K. (2005). *Stability and flexibility of language learning motivation*. Unpublished Doctorial Dissertation, Temple University Japan, Tokyo, Japan.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29, 24-49. doi:10.1093/applin/amm017
- Jafarpur, A. (1991). Can native EFL learners estimate their own proficiency? *Evaluation and Research in Education*, 5(3), 145-157. doi:10.1080/09500799109533306
- Jassen-van Dieten, A-M. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing*, 6, 30-46. doi:10.1177/026553228900600105
- Kitayama, S., & Markus, H. R. (2000). The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. In E. Diener & S. M. Eunkook (Eds.), *Culture and subjective well-being* (pp. 113-161). Cambridge: The MIT Press.
- Kobayashi, C., & Brown, J. D. (2003). Self-esteem and self-enhancement in Japan and America. *Journal of Cross-Cultural Psychology*, 34, 567-580. doi:10.1177/0022022103256479
- Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research*, 4, 900-913. doi:10.4304/jltr.4.5.900-913
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.

- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessment. *Journal of Personality and Social Psychology, 77*, 1121-1134. doi:10.1037/0022-3514.77.6.1121
- Kudo, E., & Numazaki, M. (2003). Explicit and direct self-serving bias in Japan: Reexamination of self-serving bias for success and failure. *Journal of Cross-Cultural Psychology, 34*, 511-521. doi:10.1177/0022022103256475
- Kusnic, E., & Finley, M. L. (1993). Student self-evaluation: An introduction and rationale. In J. MacGregor (Ed.), *Student self-evaluation: Fostering reflective learning* (pp. 5-14). San Francisco, CA: Jossey-Bass.
- Le Blanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*, 673-87. doi:10.2307/3586670
- Lewkowicz, J. A., & Moon, J. (1985). Evaluation: A way of involving the learner. In J. C. Anderson (Ed.), *Lancaster practical papers in English language education, Volume 6: Evaluation* (pp. 45-80). Oxford: Pergamon.
- Linacre, J. M. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*, 486-512.
- Linacre, J. M. (2006). FACETS Rasch measurement computer program (Version 3.63) [Computer software]. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch model computer program*. Chicago, IL: MESA.
- Linacre, J. M. (2014). *A user's guide to FACETS: Rasch model computer program*. Retrieved from www.winsteps.com.
- Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*, 486-512.
- Liu, M. (2012). Predicting effects of personality traits, self-esteem, language class risk-taking and sociability on Chinese university EFL learners' performance in English. *Journal of Second Language Teaching and Research, 1*, 30-57. Retrieved from <http://pops.uclan.ac.uk/index.php/jsltr/article/view/9/2>
- Liu, M., & Jackson, J. (2008). An exploration of Chinese EFL learners' unwillingness to communicate and foreign language anxiety. *The Modern Language Journal, 92*, 71-86. doi:10.1111/j.1540-4781.2008.00687.x

- Long, M. H. (1981). Input, interaction, and second language acquisition. In H. Winitz (Ed.), *Native language and foreign language acquisition* (pp.259-278). Annual of the New York Academy of Science, 379.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- MacIntyre, P. D. (1994). Variables underlying willingness to communicate: A causal analysis. *Communication Research Reports*, 11, 135-142. doi: 10.1080/08824099409359951
- MacIntyre, P.D., Babin, P. A., & Clément, R. (1999). Willingness to communicate: Antecedents and consequences. *Communication Quarterly*, 47, 215-239. doi:10.1080/01463379909370135
- MacIntyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology*, 15, 3-26. doi:10.1177/0261927X960151001
- MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, 82, 545-562. doi:10.2307/330224
- MacIntyre, P.D., & Doucette, J. (2010). Willingness to communicate and action control. *System*, 38, 161-171. doi:10.1016/j.system.2009.12.013
- MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second language learning: Toward a theoretical understanding. *Language Learning*, 39, 251-275. doi:10.1111/j.1467-1770.1989.tb00423.x
- MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, 47, 265-287. doi:10.1111/0023-8333.81997008
- Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98, 224-253. doi: 10.1037//0033-295X
- Masgoret, A., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, 53, 123-163. doi:10.1111/1467-9922.00212

- Matsumoto, H., & Kitayama, S. (1998). False uniqueness effect in Japan and the United States: Effects of culture and domains. Unpublished manuscript. Kyoto University.
- Matsuno, S. (2007). *Self-, peer, and teacher-assessment in Japanese university EFL writing classrooms*. Unpublished Doctorial Dissertation, Temple University Japan, Tokyo, Japan.
- Matsuoka, R. (2006). *Japanese college students' willingness to communicate in English*. Unpublished Doctorial Dissertation, Temple University Japan, Tokyo, Japan.
- McCroskey, J. C., & Richmond, V. P. (1982). Communication apprehension and shyness: Conceptual and operation distinction. *The Central States Speech Journal*, 33, 458-468. Retrieved from www.jamescmccroskey.com/publications/103.pdf
- McCroskey, J. C., & Richmond, V. P. (1987). Willingness to communicate. In J. C. McCroskey & J. A. Daly (Eds.), *Personality and interpersonal communication* (pp. 129-156). Beverly Hills, CA: Sage.
- McCroskey, J. C., & Richmond, V. P. (1991). Willingness to communicate: A cognitive view. In M. Booth-Butterfield (Ed.), *Communication, cognition, and anxiety* (pp. 19-37). Newbury Park, CA: Sage.
- McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, 7, 52-75. doi: 10.1177/026553229000700105
- McNamara, T. F. (1996). *Measuring second language performance*. New York, NY: Longman.
- Moere, A. V. (2006). Validity evidence in a university group oral test. *Language Testing*, 23, 411-440. doi:10.1191/0265532206lt336oa
- Mortensen, D. C., Arnston, P. H., & Lusting, M. (1977). The measurement of verbal predispositions: Scale development and application. *Human Communication Research*, 3, 146-158. doi:10.1111/j.1468-2958.1977.tb00513.x
- Myers, J. L. (2001). Self-evaluation of the "stream of thought" in journal writing. *System*, 29, 481-488. doi:10.1016/S0346-251X(01)00037-9

- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4, 386-422.
- Noels, K. A. (2001). Learning Spanish as a second language: Learners' orientations and perceptions of their teachers' communication style. *Language Learning*, 51, 107-144. doi:10.1111/0023-8333.00149
- Noels, K. A., & Clément, R. (1996). Communicating across cultures: Social determinants and acculturative consequences. *Canadian Journal of Behavioral Science*, 28, 214-228. doi:10.1037/0008-400X.28.3.214
- Noels, K., Clément, R., & Pelletier, L.G. (1999). Perceptions of teachers' communicative style and students' intrinsic and extrinsic motivation. *The Modern Language Journal*, 83, 23-34. doi:10.1111/0026-7902.00003
- Obunsha.(2010). *Eiken jyun-1 kyu niji-shiken taisaku yosou mondai (The STEP test in practical English proficiency pre-1 level interview test questions)*. Tokyo: Obunsha.
- Ockey, G.J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, 26, 161-186. doi:10.1177/0265532208101005
- Ockey, G.J. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning*, 61, 968-989. doi:10.1111/j.1467-9922.2010.00625.x
- Ortega, A. (2007). Anxiety and self-esteem. In F. Rubio (Ed.), *Self-esteem and foreign language learning* (pp. 105-127). Cambridge: Cambridge Scholars.
- Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, 6, 1-13. doi:10.1177/026553228900600103
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C Clapham & D. Gorson (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment* (pp. 175-187). Dordrecht: Kluwer Academic.
- Oya, T., Manalo, E., & Greenwood, J. (2004). The influence of personality and anxiety on the oral performance of Japanese speakers of English. *Applied Cognitive Psychology*, 18, 841-855. doi:10.1002/acp.1063

- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19, 109-131. doi:10.1191/0265532202lt224oa
- Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, 14, 25-42. doi:10.1093/applin/14.1.25
- Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal*, 76, 14-26. doi:10.1111/j.1540-4781.1992.tb02573.x
- Price, M. L. (1991). The subjective experience of foreign language anxiety: Interviews with highly anxious students. In E. K. Horwitz & D. J. Young (Eds.), *Language anxiety: From theory and research to classroom implications* (pp. 101-108). Upper Saddle River, NJ: Prentice Hall.
- Richards, J. C., & Schmidt, R. (2002). *The Longman dictionary of language teaching and applied linguistics* (3rd ed.). London: Pearson Education.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing*, 15, 1-20. doi:10.1177/026553229801500101
- Saito, Y., & Samimy, K. K. (1996). Foreign language anxiety and language performance: A study of learner anxiety in beginning, intermediate, and advanced-level college students of Japanese. *Foreign Language Annals*, 29, 239-251. doi:10.1111/j.1944-9720.1996.tb02330.x
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25, 553-581. doi:10.1177/0265532208094276
- Sato, T. (2011). Japanese and native English-speaking teachers' perspectives on learners' oral English performance. *ARELE: Annual Review of English Language Education in Japan*, 22, 17-32. Retrieved from http://ci.nii.ac.jp/els/110009425245.pdf?id=ART0009903116&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1403638260&cp=
- Schmidt, R. (1990). The role of consciousness in language learning. *Applied Linguistics*, 11, 129-158. doi: 10.1093/applin/11.2.129

- Scovel, T. (1991). The effect of affect on foreign language learning: A review of the anxiety research. In E. K. Horwitz & D. J. Young (Eds.), *Language anxiety: From theory and research to classroom implications* (pp. 15-23). Upper Saddle River, NJ: Prentice Hall.
- Shirahata, T., Tomita, Y., Muranoi, H., & Wakabayashi, S. (1999). *A guide to English language teaching terminology*. Tokyo: Taisyukan Shoten.
- Shokraii, N. H. (1998, January). The self-esteem fraud: Why feel-good education does not lead to academic success. *USA Today Magazine*, 126, 66-68. Retrieved from Academic Search Premier
- Sick, J. R., & Nagasaka, J. P. (2000). A test of your willingness to communicate in English. Unpublished Questionnaire.
- Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19, 69-75. doi:10.1080/03075079412331382153
- Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment and Evaluation in Higher Education*, 22, 289-305. doi:10.1080/0260293970220303
- Swain, M. (1985). Communicative competence: Some roles of comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp.235-253). Rowley, MA: Newbury House.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.
- Thompson, G. (2009). Reevaluating the test specifications of an oral proficiency test. *The Journal of Kanda University of International Studies*, 21, 233-260. Retrieved from http://ci.nii.ac.jp/els/110007058175.pdf?id=ART0008988857&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1403644850&cp=
- Todd, R. W. (2002). Using self-assessment for evaluation. *English Teaching Forum*, 40, 16-19. Retrieved from <http://americanenglish.state.gov/resources/english-teaching-forum-2002-volume-40-number-1#child-110>
- Tremblay, P., & Gardner, R. C. (1995). Expanding the motivation construct in language learning. *The Modern Language Journal*, 79, 505-520. doi:10.1111/j.1540-4781.1995.tb05451.x

- Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (forthcoming). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition*.
- Tsui, A. B. M. (1996). Reticence and anxiety in second language learning. In K. M. Bailey & D. Nunan (Eds.), *Voices from the language classroom: Qualitative research in second language education* (pp. 145-167). Cambridge: Cambridge University Press.
- Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.
- Ushioda, E. (2001). Language learning at university: Exploring the role of motivational thinking. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 93-125). Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.
- Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, 33, 448-451.
doi:10.1037/h0027806
- Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II Validation activities. *Journal of Applied Measurement*, 8, 204-234.
- Woodrow, L. (2006). Anxiety and speaking English as a second language. *Regional Language Center Journal*, 37, 308-327. doi: 10.1177/0033688206071315
- Yamaguchi, S. (2006). *The cultural meaning of self-esteem and the cultural comparison of measurement*. Tokyo: Tokyo University.
- Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, 86, 54-66.
doi:10.1111/1540-4781.00136
- Yashima, T., Zenuk-Nishide, L., & Simizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning*, 54, 119-152.
doi:10.1111/j.1467-9922.2004.00250.x

- Yoshikawa, M. (2005). Structure of Japanese EFL learners' speaking ability: An empirical approach. *ARELE: Annual Review of English Language Education in Japan*, 16, 51-60. Retrieved from http://ci.nii.ac.jp/els/110008512286.pdf?id=ART0009707060&type=pdf&lang=jp&host=cinii&order_no=&ppv_type=0&lang_sw=&no=1403641115&cp=
- Young, D. J. (1986). The relationship between anxiety and foreign language proficiency ratings. *Foreign Language Annals*, 19, 439-445. doi:10.1111/j.1944-9720.1986.tb01032.x
- Zakahi, W. R., & McCroskey, J. C. (1989). Willingness to communicate: A potential confounding variable in communication research. *Communication Reports*, 2, 96-104. doi:10.1080/08934218909367489

APPENDICES

APPENDIX A CONSENT FORM

Participant's Name:

DATE:

Title: Effects of L2 affective factors on self-assessment of speaking

Investigator: Noriko Iwamoto

Purpose of Research: To examine how affective factors influence self-assessment of speaking

Study Procedures: Questionnaire data and interview data will be collected from the participant. The data are analyzed using SPSS, Winsteps, Facet, and Amos.

Participant's Understanding & Precautions:

I understand that circumstances may arise which might cause the investigators to terminate my participation in the study before its completion. These circumstances would include any situation in which the investigators believed that further participation would be dangerous to me. I understand that the results of this study may be published but my identity will not be disclosed. I agree to permit Temple University Japan to keep, publish, or dispose of the results of my study results.

I understand that the possible benefits of this study are that it will explain how affective factors influence self-assessment of speaking and that teachers will learn how their students think about their own speaking abilities.

I understand that I may refuse consent or withdraw from the research project at any time without penalty.

I understand that I will be compensated 500 yen for my participation in the study.

I have read and understood this consent form and I voluntarily agree to participate in this research project. I understand that I will be given a copy of the signed consent form.

Signature of the Subject
Date

Signature of Investigator
Date

APPENDIX B ORAL INTERVIEW FRAMEWORK

Part I

Participant will speak English with an interviewer for about 5 minutes.

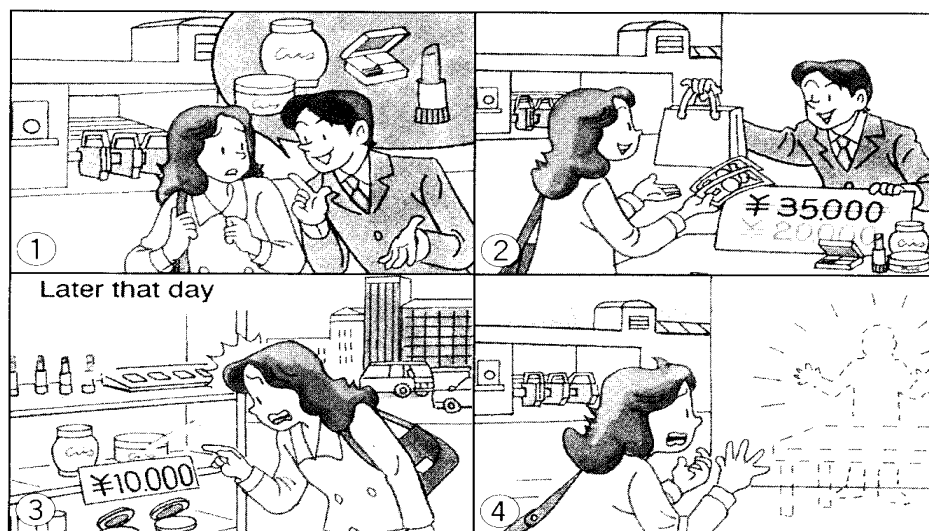
The interviewer asks the following questions.

Topics

1. Hobby
(What is your hobby?)
2. Weekend
(What did you do last weekend?)
3. University life
(How is your university life? Please tell me difference between high school life and university life.)
4. Future
(What do you want to do after you graduate from this university? Please tell me about your future plans.)
5. Opinion
(What do you think about learning English as a required subject at university?)

Part II (adapted from Obunsha, 2010)

Participant is given the following cartoon. The participant is given one minute to think about what to say. After one minute, the participant narrates the story for two minutes.



(実際の問題カードはカラーです)

資料提供：財日本英語検定協会

APPENDIX C
SPEAKING ASSESSMENT RUBRIC (ENGLISH VERSION)
(adapted from Bonk & Ockey, 2003)

	Grammar	Vocabulary	Fluency	Pronunciation
5	Uses high level of discourse structures with near native-like accuracy	Wide range of vocabulary with near native-like use, vocabulary is clearly appropriate to express opinion	Near native-like fluency, effortless, smooth, natural rhythm	Rarely mispronounces, able to speak with near native-like pronunciation
4.5	Shows ability to use full range of grammatical structures but makes some errors. Errors do not impede the meaning of the utterances	Lexis sufficient for task although not always precisely used	Speaks with confidence, but has some unnatural pauses, some errors in speech rhythm, rarely gropes for words	Pronunciation is clear, occasionally mispronounces some words, accent may sound foreign but does not interfere with meaning
4				
3.5	Relies mostly on simple (but generally accurate) sentences, has enough grammar to express meaning, complex sentences are used but often inaccurately	Lexis generally adequate for expressing opinion but often used inaccurately	Speech is hesitant, some unnatural rephrasing and groping for words	Pronunciation is not native like but can be understood, mispronounces unfamiliar words, may not have mastered some sounds
3				
2.5	Uses simple inaccurate sentences and fragmented phrases, doesn't have enough grammar to express opinions clearly	Lexis not adequate for task, cannot express opinion	Slow strained speech, constant groping for words and long unnatural pauses (except for routine phrases)	Frequently mispronounces, accent often impedes meaning, difficult to understand even with concentrated listening
2				
1.5	Only says a few words, cannot make a reasonable judgment of student's grammatical ability	Little lexis, inadequate for simple communication	Fragments of speech that are so halting that conversation is virtually impossible	Frequently mispronounces, heavy accent, may use Japanese katakana-like speech which is virtually not comprehensible
1				

APPENDIX D
SPEAKING ASSESSMENT RUBRIC (JAPANESE VERSION)
(adapted from Bonk & Ockey, 2003)

	文法	語彙	流暢さ	発音
5	ネイティブに近い正確さで、高度な文構造を用いている。	幅広い語彙をネイティブに近い使い方ができる。自分の考えを表現するのに適した語彙能力がある。	ネイティブに近い流量さ。楽にスムーズで自然なリズムで話す。	発音の誤りがほとんどない。ネイティブに近い発音で話すことができる。
4.5 4	幅広い文法を使える能力を示しているが、いくつか文法上の誤りがある。だが、言いたい事はきちんと伝わる。	今回の課題に十分な語彙を知っているが、常に正確に使用できていないわけではない。	自信を持って話しているが、いくつか不自然な休止や発話のリズムに間違いがある。しかし単語が浮かばずに考え込むことはめったにない。	発音が明確である。いくつかの単語の発音に誤りがある。日本語的ななまりがあるが、言いたい事はきちんと伝わる。
3.5 3	主に単純な文を使っている（大体正確である）。言いたい事を表現するのに十分な文法を知っている。複雑な文を使うと、しばしば間違える。	大体において、自分の考えを表現するのに適した語彙であるが、しばしば不正確な使い方をしている。	発話はためらいがちで、不自然な単語の言い換えや、単語を探して考え込むことが何度かある。	発音はネイティブのようではないが、理解できる。よく知らない単語の発音を間違える。いくつかの英語の音を習得していない。
2.5 2	単純で不正確な文や断片的なフレーズを使う。自分の考えを明確に表現する文法を十分に知らない。	今回の課題に対して十分な語彙を知らない。自分の考えを表現することができない。	ゆっくりで不自然な発話。常に単語を探して考え込み、長く不自然な休止がある（決まり文句を言うとき以外）	たびたび発音に誤りがある。なまりにより、しばしば言いたい事柄が伝わらない。集中して聞いていても、理解しにくい。
1.5 1	単語をいくつか言うのみ。正確に文法能力を把握することができない。	語彙に乏しい。単純なコミュニケーションも取れない。	発話の断片。とぎれとぎれで、会話は実質的に不可能。	ひんばんに発音を間違い、強いなまりがある。日本語のカタカナのような発音の仕方、ほとんど理解できない。

APPENDIX E
SELF-ESTEEM SCALE (ENGLISH VERSION)
(adapted from Rosenberg, 1965)

Indicate your level of agreement with each of the following statements by circling one number on the rating scale that best describe the way you feel about yourself. Use the following scale as your guide.

	1	2	3	4	5	6
	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. I believe that I have a number of good qualities.	1	2	3	4	5	6
2. I am able to do things as well as most other people.	1	2	3	4	5	6
3. I feel useful most of the time.	1	2	3	4	5	6
4. I feel that I am a person of worth.	1	2	3	4	5	6
5. I respect myself.	1	2	3	4	5	6
6. I feel that I am a success.	1	2	3	4	5	6
7. I view myself positively rather than negatively.	1	2	3	4	5	6
8. I am able to do things better than other people.	1	2	3	4	5	6
9. I feel that I am at least on an equal plane with others.	1	2	3	4	5	6
10. I have more good points than weak points.	1	2	3	4	5	6

APPENDIX F
SELF-ESTEEM SCALE (JAPANESE VERSION)
(adapted from Rosenberg, 1965)

次の各項目について、あなた自身にどの程度当てはまるか、尺度上の該当する項目に○をつけてください。

	1	2	3	4	5	6
	全くそう 思わない	そう 思わない	あまりそう 思わない	ややそう 思う	そう思う	強くそう 思う
1. 私にはいくつかいい所があると思う。	1	2	3	4	5	6
2. 私はたいいていの人ができる程度には物事ができる。	1	2	3	4	5	6
3. 私は自分がたいいていの場合において役に立つ人間だと思 う。	1	2	3	4	5	6
4. 私は自分が価値ある人間だと思う。	1	2	3	4	5	6
5. 私は自分のことを尊敬できる。	1	2	3	4	5	6
6. 私は自分が成功者だと思う。	1	2	3	4	5	6
7. 私は自分に対してネガティブというよりむしろポジテ ィブな見方をしている。	1	2	3	4	5	6
8. 私は人よりも物事を上手に出来ると思う。	1	2	3	4	5	6
9. 私は他人と少なくとも同レベルの人間だと思う。	1	2	3	4	5	6
10. 私は短所よりも長所の方が多いと思う。	1	2	3	4	5	6

APPENDIX G
L2 SPEAKING ANXIETY SCALE (ENGLISH VERSION)
(adapted from Horwitz et al., 1986)

Indicate your level of agreement with each of the following statements by circling one number on the rating scale that best describe your attitude toward speaking English. Use the following scale as your guide.

	1	2	3	4	5	6
	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. I worry that other students think my English speaking ability is low.	1	2	3	4	5	6
2. I feel that other students speak English better than me.	1	2	3	4	5	6
3. I would feel nervous speaking English with native speakers of English.	1	2	3	4	5	6
4. I feel nervous about speaking English in class activities.	1	2	3	4	5	6
5. I lack confidence in my English speaking abilities.	1	2	3	4	5	6
6. I worry that my English teacher thinks that my English speaking level is low.	1	2	3	4	5	6
7. I feel nervous speaking English.	1	2	3	4	5	6
8. I worry that I will make mistakes when I speak English.	1	2	3	4	5	6
9. I feel nervous having a conversation in English.	1	2	3	4	5	6
10. I feel self-conscious when I speak English.	1	2	3	4	5	6

APPENDIX H
L2 SPEAKING ANXIETY SCALE (JAPANESE VERSION)
(adapted from Horwitz et al., 1986)

次の各項目について、あなた自身にどの程度当てはまるか、尺度上の該当する項目に○をつけてください。

	1	2	3	4	5	6
	全くそう 思わない	そう 思わない	あまりそう 思わない	ややそう 思う	そう思う	強くそう 思う
1. 他の学生に私の英語スピーキング能力が低いと思われることが心配である。	1	2	3	4	5	6
2. 自分よりも他の学生の方が英語を上手に話すと思っている。	1	2	3	4	5	6
3. ネイティブスピーカーと英語を話すとき緊張するだろうと思う。	1	2	3	4	5	6
4. 授業内活動で英語を話すとき、緊張する。	1	2	3	4	5	6
5. 自分のスピーキング能力に自信が持てない。	1	2	3	4	5	6
6. 英語の先生に、自分の英語スピーキング能力が低いと思われることが心配である。	1	2	3	4	5	6
7. 英語を話すとき緊張する。	1	2	3	4	5	6
8. 英語を話すとき、間違いをしないかを心配してしまう。	1	2	3	4	5	6
9. 英語で会話すると緊張する。	1	2	3	4	5	6
10. 英語を話すとき、自意識過剰になってしまう。	1	2	3	4	5	6

APPENDIX I
L2 WILLINGNESS TO COMMUNICATE SCALE (ENGLISH VERSION)
(adapted from Sick and Nagasaka, 2000)

Indicate your level of agreement with each of the following statements by circling one number on the rating scale that best describes your willingness to speak English. Use the following scale as your guide.

	1	2	3	4	5	6
	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1. I would be willing to answer a question from my teacher in English class.	1	2	3	4	5	6
2. I would be willing to start a conversation in English with an international student.	1	2	3	4	5	6
3. I would be willing to start an English conversation with my English teacher outside of class.	1	2	3	4	5	6
4. I would be willing to tell my native teacher about my weekend.	1	2	3	4	5	6
5. I would be willing to ask my teacher about his or her hometown in English.	1	2	3	4	5	6
6. I would be willing to interview a teacher in English.	1	2	3	4	5	6
7. I would be willing to participate in an English discussion with three or four students in English class.	1	2	3	4	5	6
8. I would be willing to give a welcoming speech to a group of foreign students.	1	2	3	4	5	6
9. I would be willing to talk about common topics (e.g., hobbies and vacation) in English with my pair partner in English class.	1	2	3	4	5	6
10. I would be willing to talk about academic topics (e.g., social issues and the environment) in English with my pair partner in English class.	1	2	3	4	5	6
11. I would be willing to talk about what I did during the summer vacation in front of the class.	1	2	3	4	5	6
12. I would be willing to guide a group of three Canadian students around Tokyo.	1	2	3	4	5	6

APPENDIX J
L2 WILLINGNESS TO COMMUNICATE SCALE (JAPANESE VERSION)
(adapted from Sick and Nagasaka, 2000)

次のような英語を使用する機会があるとして、それぞれについてのあなたのやる気度を答えてください。

	1	2	3	4	5	6
	全くそう 思わない	そう 思わない	あまりそう 思わない	ややそう 思う	そう思う	強くそう 思う
1. 英語のクラスで先生の英語の質問に答えたいと思う。	1	2	3	4	5	6
2. 留学生と英語で会話をしてみたいと思う。	1	2	3	4	5	6
3. 授業外で英語の先生と英語で会話してみたいと思う。	1	2	3	4	5	6
4. 自分の週末について、先生に英語で話してみたいと思う。	1	2	3	4	5	6
5. ネイティブの先生の故郷について英語で聞いてみたいと思う。	1	2	3	4	5	6
6. 英語で先生にインタビューしてみたいと思う。	1	2	3	4	5	6
7. 英語の授業で、3~4名のクラスメートと英語でディスカッションをしてみたいと思う。	1	2	3	4	5	6
8. 外国人の学生の集団に対して、歓迎のスピーチを英語で述べてみたいと思う。	1	2	3	4	5	6
9. 英語の授業で、ペアワークの相手と英語で一般的なトピック（趣味や休暇など）について英語で話してみたいと思う。	1	2	3	4	5	6
10. 英語の授業で、ペアワークの相手と英語でアカデミックなトピック（社会問題や環境問題など）について英語で話してみたいと思う。	1	2	3	4	5	6
11. クラスの前で、自分の夏休みの思い出について英語でスピーチしてみたいと思う。	1	2	3	4	5	6
12. 3人のカナダ人のグループを東京観光に連れて行ってみたいと思う。	1	2	3	4	5	6

APPENDIX K
L2 SPEAKING MOTIVATION SCALE(ENGLISH VERSION)
(adapted from Gardner, 1985)

Indicate your level of agreement with each of the following statements by circling one number on the rating scale that best describe your attitude toward speaking English. Use the following scale as your guide.

1	2	3	4	5	6
Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree

Attitude Toward Learning to Speak English

1. I enjoy speaking English.	1	2	3	4	5	6
2. I enjoy speaking English more than reading English.	1	2	3	4	5	6
3. I enjoy speaking English more than writing English.	1	2	3	4	5	6
4. I am very interested in learning to speak English.	1	2	3	4	5	6
5. I admire Japanese students who can speak English well.	1	2	3	4	5	6
6. I would enjoy talking with native English teachers.	1	2	3	4	5	6
7. I enjoy speaking English more than listening to English.	1	2	3	4	5	6
8. I look forward to my English speaking classes.	1	2	3	4	5	6
9. I enjoy English speaking classes more than other classes.	1	2	3	4	5	6
10. I look forward to opportunities to speak English.	1	2	3	4	5	6
11. I think that English is the most important subject in school.	1	2	3	4	5	6
12. Speaking English is important for engineers.	1	2	3	4	5	6
13. I consider speaking English to be one of the most important skills to learn in school.	1	2	3	4	5	6

L2 Speaking Motivational Intensity

1. I concentrate well when I speak English.	1	2	3	4	5	6
2. I think I try to speak English more than other students.	1	2	3	4	5	6
3. I speak English as much as possible in class.	1	2	3	4	5	6
4. I look for opportunities to speak English outside of class.	1	2	3	4	5	6
5. I make an effort not to make grammatical mistakes when I speak English.	1	2	3	4	5	6
6. I spend a long time studying English.	1	2	3	4	5	6
7. I study English more than most of my classmates.	1	2	3	4	5	6
8. I often think about how I can improve my English speaking skills.	1	2	3	4	5	6
9. I work hard to become an excellent speaker of English.	1	2	3	4	5	6
10. I plan to keep improving my English speaking skills even after graduating from college.	1	2	3	4	5	6

Desire to Learn to Speak English

1.	I would take an English conversation course in school, even if it were not required.	1	2	3	4	5	6
2.	I wish I had more classes in which I could speak English.	1	2	3	4	5	6
3.	I really want to learn to speak English better.	1	2	3	4	5	6
4.	Learning to speak English is more important than learning to read English.	1	2	3	4	5	6
5.	Learning to speak English is more important than learning to write English.	1	2	3	4	5	6
6.	I seek out opportunities to speak English.	1	2	3	4	5	6
7.	I study English speaking on my own through radio or TV language program.	1	2	3	4	5	6
8.	I believe that Japanese students should be taught to speak English at school.	1	2	3	4	5	6
9.	My desire to learn speak English is increasing.	1	2	3	4	5	6
10.	I wish I could speak English perfectly.	1	2	3	4	5	6

APPENDIX L
L2 SPEAKING MOTIVATION SCALE(JAPANESE VERSION)
(adapted from Gardner, 1985)

次の各項目について、あなた自身にどの程度当てはまるか、尺度上の該当する項目に○をつけてください。

1	2	3	4	5	6
全くそう 思わない	そう 思わない	あまりそう 思わない	ややそう 思う	そう思う	強くそう 思う

Attitude Toward Learning to Speak English

1. 英語を話すことは楽しい。	1	2	3	4	5	6
2. 英語のリーディングよりもスピーキングの方が楽しい。	1	2	3	4	5	6
3. 英語のライティングよりもスピーキングの方が楽しい。	1	2	3	4	5	6
4. 英語のスピーキングを学ぶことに興味がある。	1	2	3	4	5	6
5. 英語を上手に話せる日本人の学生は優秀だと思う。	1	2	3	4	5	6
6. ネイティブの英語の先生と英語で話せたら楽しいだろうと思う。	1	2	3	4	5	6
7. 英語のリスニングよりもスピーキングの方が楽しい。	1	2	3	4	5	6
8. 英語のスピーキングのクラスを楽しみにしている。	1	2	3	4	5	6
9. 他のクラスよりも英語のスピーキングのクラスの方が楽しいと思う。	1	2	3	4	5	6
10. 英語を話す機会を楽しみにしている。	1	2	3	4	5	6
11. 英語は学校で最も重要な科目である。	1	2	3	4	5	6
12. エンジニアにとって英語を話すことは重要である。	1	2	3	4	5	6
13. 英語のスピーキングは学校で学ぶスキルの中で最も重要なものの一つである。	1	2	3	4	5	6

L2 Speaking Motivational Intensity

- | | | | | | | |
|---|---|---|---|---|---|---|
| 1. 英語を話す時にはとても集中している。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2. 私は他の学生よりも英語を話すようにしている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 3. 授業ではできるだけたくさん英語を話したいと思っている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 4. 授業外で英語を話す機会を探している。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 5. 私は英語を話す時、文法を間違えないように努力している。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 6. 英語の勉強に長時間費やしている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7. 私は大部分のクラスメートよりも英語を勉強していると思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 8. 自分の英語のスピーキング能力をどうやって伸ばすことができるかについてよく考える。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 9. 英語を上手に話せるようになるために一生懸命勉強している。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 10. 大学卒業後も英語のスピーキング能力を伸ばしていくつもりである。 | 1 | 2 | 3 | 4 | 5 | 6 |

Desire to Learn to Speak English

- | | | | | | | |
|-------------------------------------|---|---|---|---|---|---|
| 1. 必修でなかったとしても、英会話のクラスを受講すると思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2. 英語を話すことができるクラスがもっとあればいいのにとと思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 3. 英語をもっと上手に話せるようになる方法を学びたいと強く思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 4. 英語のリーディングよりもスピーキングを学ぶほうが重要だと思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 5. 英語のライティングよりもスピーキングを学ぶほうが重要だと思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 6. 英語を話す機会を捜し求めている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7. ラジオやテレビの英語番組を使って、自分で英会話の勉強をしている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 8. 日本の学生は学校で英語のスピーキング学ぶべきだと思う。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 9. 英語のスピーキング学びたいという気持ちが大きくなっている。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 10. 完ぺきに英語を話すことができればいいのにとと思う。 | 1 | 2 | 3 | 4 | 5 | 6 |

APPENDIX M
L2 SPEAKING SELF-CONFIDENCE SCALE(ENGLISH VERSION)

Indicate your level of agreement with each of the following statements by circling one number on the rating scale that best describe your self-confidence toward speaking English. Use the following scale as your guide.

1 = Strongly disagree
4 = Slightly agree

2 =Disagree
5 =Agree

3 =Slightly disagree
6 = Strongly agree

- | | | | | | | |
|--|---|---|---|---|---|---|
| 1. I can hold a 5-minute conversation with my teacher in English | 1 | 2 | 3 | 4 | 5 | 6 |
| 2. I can participate in a 10-minute small group discussion with 3-4 other students in English class. | 1 | 2 | 3 | 4 | 5 | 6 |
| 3. I can hold a 5-minute conversation in English with pair partner. | 1 | 2 | 3 | 4 | 5 | 6 |
| 4. I can participate in a 20-minute group discussion with 3-4 other students in English class. | 1 | 2 | 3 | 4 | 5 | 6 |
| 5. I can talk about common topics (e.g., hobbies and vacation) in English. | 1 | 2 | 3 | 4 | 5 | 6 |
| 6. I can express my opinion about common topics in English. | 1 | 2 | 3 | 4 | 5 | 6 |
| 7. I can talk about academic topics (e.g., social issues and the environment) in English. | 1 | 2 | 3 | 4 | 5 | 6 |
| 8. I can give directions from this classroom to my home in English | 1 | 2 | 3 | 4 | 5 | 6 |
| 9. I can give a 3-minute speech in English (with the topic I chose) | 1 | 2 | 3 | 4 | 5 | 6 |
| 10. I can talk about what I did last weekend in English | 1 | 2 | 3 | 4 | 5 | 6 |
| 11. I can tell the time to a foreigner in English | 1 | 2 | 3 | 4 | 5 | 6 |
| 12. I can use English to phone a foreign student to invite him/her to dinner. | 1 | 2 | 3 | 4 | 5 | 6 |
| 13. I can give directions to a certain place in English using a map to a pair partner. | 1 | 2 | 3 | 4 | 5 | 6 |
| 14. I can exchange greetings in English. | 1 | 2 | 3 | 4 | 5 | 6 |
| 15. I can introduce myself in English. | 1 | 2 | 3 | 4 | 5 | 6 |
| 16. I can use English to phone a hotel in a foreign country to make a reservation. | 1 | 2 | 3 | 4 | 5 | 6 |

APPENDIX N
L2 SPEAKING SELF-CONFIDENCE SCALE(JAPANESE VERSION)

次の各項目について、あなた自身にどの程度当てはまるか、尺度上の該当する項目に○をつけてください。

1 = 全くそう思わない 2 = そう思わない 3 = あまりそう思わない
4 = ややそう思う 5 = そう思う 6 = 強くそう思う

- | | | | | | | |
|---|---|---|---|---|---|---|
| 1. 先生と5分間英語で会話ができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 2. 英語のクラスで、3～4人の学生と10分間のグループディスカッションに参加できる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 3. 英語のクラスでペアワークのパートナーと英語で5分間話すことができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 4. 英語のクラスで、3～4人の学生と20分間のグループディスカッションに参加できる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 5. 一般的なトピック（趣味や休暇など）について英語で話すことができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 6. 一般的なトピックについて英語で自分の意見を述べることができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 7. アカデミックなトピック（社会問題や環境問題など）について英語で話すことができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 8. この教室から自分の家までの道順を英語で説明できる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 9. 3分間英語でスピーチできる（自分の好きなトピックで）。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 10. 先週末にしたことを英語で話すことができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 11. 外国人に英語で時間を教えてあげることができる。 | 1 | 2 | 3 | 4 | 5 | |
| 12. 外国人の友人に電話をかけて、英語で夕食に招待することができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 13. ペアワークの相手に地図を使って、ある場所への行き方を英語で説明できる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 14. 英語でかんたんなあいさつができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 15. 英語でかんたんな自己紹介ができる。 | 1 | 2 | 3 | 4 | 5 | 6 |
| 16. 海外のホテルに電話をして、英語で部屋を予約できる。 | 1 | 2 | 3 | 4 | 5 | 6 |

APPENDIX O
ORAL INTERVIEW TRANSCRIPTS FOR EACH LEVEL PARTICIPANT

1. Transcript of Low Proficiency Participant

Part I

Interviewer: What is your hobby?

Student: My hobby is soccer.

I: What did you do last weekend?

S: I played. bowling.

I: How is your university life?

S:

I: Do you enjoy your university life?

S: yes I do

I: Please tell me difference between high school life and university life.

S: chigaidesuka?

I: Yes

S: (long pause) class. type.

I: What do you want to do after you graduate from this university? Please tell me about your future plans.

S: I (long pause) syusyoku work

What do you think about learning English as a required subject at university?

S: it is important

Part II

S: Etto (small laugh) nante ieba iidesuka?

She is talk jyudotai? she is talked. salesman. drug (long pause)

There drug is sanjyugo thirty five sonotoki te nandesitakke?

I: Then

S: Then It is twenty thousand yen. . . . She buy it.

Later that day, she look drug nedan te nan desitakke?

I: Price

S: Price is ten thousand. . . . yen

She go to the salesman but he he don't stay here

2. Transcript of Intermediate Proficiency Participant

Part I

I: What is your hobby?

S: My hobby is to play a basketball.

I: What did you do last weekend?

S: Last weekend. . . . I have a arubaito ah part-time job.

I: How is your university life?

S: It's very enjoyable.

I: Please tell me difference between high school life and university life.

S: Ah...I have many free time.

I: What do you want to do after you graduate from this university? Please tell me about your future plans.

S: I want to go graduatining...tion...school.

I: What do you think about learning English as a required subject at university?

S: Learning English is very important.

Part II

S: The woman...the man wants to sell...some thing..thing...woman buy these things. Later that day, she...find... these things cheaper...but the salesman disappeared...was disappeared

3. Transcript of High Proficiency Participant

Part I

I: What is your hobby?

S: My hobby is playing clarinet, my clarinet.

I: What did you do last weekend?

S: I performed on stage on... my wind orchestra.

I: How is your university life?

S: It's very fine.

I: Please tell me difference between high school life and university life.

S: University life is . . . more more exciting and free.

I: What do you want to do after you graduate from this university? Please tell me about your future plans.

S: I...I like to study more in graduate school.

I: What do you think about learning English as a required subject at university?

S: I think English is necessary to ...for my future. So I have to study hard.

Part II

S: A salesman asked ...a woman to buy some cosmetic items...in a high price...and she...bought them...but later she found ...she paid more money than its ... real cost. A woman searched for the businessman, but he is gone.