AN ITEM BANK DEVELOPMENT TO INCLUDE TEST ITEMS ASSESSING ORTHOGRAPHIC AND PHONOLOGICAL PROCESSING SKILLS AT THE FACULTY OF NURSING

ICHIYAMA, YOKO

Abstract

In recent years, the importance of assessing students' orthographic and phonological processing skills has been acknowledged, especially in L1 reading. Moreover, the development of an item bank for an in-house English placement test would enable faculty to assign students to appropriate English-language classes, which in turn would likely lead to students' successful completion of English-language programs in the tertiary-level institution. Little has been reported, however, on the L2/FL reading contexts. This study thus investigated the process of developing an item bank with orthographic and phonological processing skills for the Faculty of Nursing. The study involved identification of the orthographic and phonological features of the faculty's English curriculum and materials. It also explored the orthographic and phonological features of two commercially produced English proficiency tests, the TOEFL and the TOEIC, and determined whether these tests correspond to the Faculty of Nursing curriculum requirements. The study also used Rasch analysis to validate the development of test items to assess orthographic and phonological skills, and explored whether these test items correspond to the requirements of the faculty's English curriculum.

Analysis of the faculty's curriculum and the commercially produced English proficiency tests revealed that the two tests may not be appropriate tools to measure students' orthographic and phonological processing skills. The Rasch analysis—including separation, reliability, test targeting, and unidimensionality for a total of 147 items—yielded 90 equated test items. Moreover, the test items showed sufficient spreads: 9 (10%) were grouped at the beginner level, 74 (82%) at the intermediate level, and 7 (8%) at the advanced level.

AN ITEM BANK DEVELOPMENT TO INCLUDE TEST ITEMS ASSESSING ORTHOGRAPHIC AND PHONOLOGICAL PROCESSING SKILLS AT THE FACULTY OF NURSING

Yoko ICHIYAMA

Doctor of Education (EdD)

School of Education

Durham University

2018

List of Abbreviations

AWL	The 570 most frequently used academic words
CBT	Computer-based test
CTT	Classical Test Theory
CV	Consonant Vowel
CVV	Consonant Vowel Vowel
CVCV	Consonant Vowel Consonant Vowel
EC	English for Communication
EFL	English as a Foreign Language
ETS	English Testing Service
FL	Foreign Language
G	Grapheme
iBT	Internet-based Test
IELTS	The International English-language Testing System
IRT	Item Response Theory
JE	Phonetic symbol that has same grapheme as the Japanese one-to-one correspondences between graphemes and phonemes
K1	The 1,000 most frequently used words
K2	The second 1,000 most frequently used words
L1	First Language
L2	Second Language
M	Mean
ME	Medical English
MEXT	The Ministry of Education, Culture, Sports, Science and Technology
MNSQ	Mean square statistic
NE	Phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation
NNE	Phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese
OECD	The Organization for Economic Cooperation and Development
OFF	Off-listed words

P	Phoneme
PISA	Program for International Student Assessment
S	1-standard deviation
T	2-standard deviation
TLU	Target Language Use
TOEIC	The Test of English for International Communication
TOEFL	The Test of English as a Foreign Language
ZSTD	Z-standardized
14SF	A 2014 Spring term final test
14FF	A 2014 Fall term final test
15SF	A 2015 Spring term final test
15FF	A 2015 Fall term final test
15P	A 2015 placement test

Declaration

The work presented in this thesis is entirely my own and has not been previously submitted to any other university for a degree or a qualification.

Copyright

Acknowledgements

This dissertation would not have been possible without the support of many people and institutions. I would like to thank students and staff who participated in this study as test participants. I would also like to thank my dissertation committee, whose work over many years and in many different lectures inspired this study. Finally, I would like to thank my family for their constant support and encouragement on completing my dissertation.

Dedication

With great respect, love, and appreciation for my father, Kazumi Ichiyama, my mother, Keiko Ichiyama, and my brother, Shinichiro Ichiyama, to whom I am deeply indebted and grateful.

List of Tables

Table 1.1 *Situation at the Time of the Establishment of the Fundamental Law of Education Compared to the Present (MHLW, 2014; Portal Site of Official Statistics of Japan (e-Stat), 2014a, 2014b)* 17

Table 1.2 *The World's Rankings 2013-2014 Top 400 (THES, 2013)* 18

Table 1.3 *PISA Results on Reading Literacy, Mathematics, Science in 2000, 2003, 2006 and 2009 (OECD, 2007, 2009, 2010a, 2010b, 2010c).* 24

Table 1.4 *The Three Versions of Course of Study by Their Year of Introduction, the Year of Implementation, and Major Changes* 27

Table 1.5 *The English Textbooks Adopted in 11 Wards and Islands in Tokyo Since 2012* 32

Table 1.6 *Studies on the Number of Universities that Use English-language Placement Tests* 34

Table 1.7 *Similarities and Differences between Norm-referenced Tests and Criterion-Referenced Tests (Montgomery & Connolly 1987)* 36

Table 1.8 *Comparison of the IELTS, the TOEFL, TOEIC, the TOEIC Retrieved from IETLS and ETS Homepage* 38

Table 1.9 *Matching Tests to Decision Purposes (Brown, 1996, p.7)* 43

Table 2.1 *Japanese Hiragana, Katakana, and Romaji with International Phonetic Alphabet (Suski, 1931)* 70

Table 2.2 *English Consonant Chart (Rogerson-Revell, 2011 p. 47)* 73

Table 2.3 *Japanese Consonant Chart (Rogerson- Revell, 2011, p. 282)* 73

Table 2.4 *English Vowel Chart (Rogerson-Revell, 2011, p. 67)* 75

Table 2.5 *Japanese Vowel Chart (Rogerson- Revell, 2011, p.283)* 75

Table 2.6 *The Modified Version of Narita's (2009) Table of English Phonemes that Correspond with Grapheme* 78

Table 2.7 *Progressive Matrixes for Facets of Validity (Messick, 1989, p. 20)* 81

Table 2.8 *Six Components of Validity and Analysis Procedures (Koizumi, 2005; adapted from Bachman et al. 1997; Banerjee et al. 1997; Chapelle 1999; Cheng et al. 2004; Messick 1989, 1996)* 81

Table 3.1 *Textbooks Used in the Faculty of Nursing at Toho University in 2015* 92

Table 3.2 *An Example Of The List Of Medical Terms* 94

Table 3.3 *Entrance Examination Subjects for Postgraduate Studies* 97

Table 4.1 *TOEFL Reading Questions Types (The Official Guide to the TOEFL Test, 4th Edition, 2012)* 103

Table 4.2 *An Example of the Vocabulary List of the TOEFL Sample Test* 108

Table 4.3 *An Example of the Vocabulary List of the TOEIC Sample Test* 108

Table 4.4 *A Guide to Understanding TOEFL IBT Scores (ETS, 2008)* 111

Table 4.5 *TOEIC Reading Score Descriptors (ETS, 2014)* 114

Table 4.6 *Percentages of TOEIC Test Takers, by Reading Score Level, Who Indicated that They could Perform Various English-language Reading Tasks Either Easily or with Little Difficulty (ETS, 2008)* 115

Table 4.7 *Types of Test Items Included in the Reading Section of TOEFL* 118

Table 4.8 *Types of Test Questions in the TOEIC's Reading Section* 118

Table 5.1 *Tests by Type and Number of Participants and Items* 130

Table 5.2 *Number and Types of Words in Original Test Items* 132

Table 5.3 *Revision Processes of Tests Regarding the Number of Participants and Items* 136

Table 5.4 *Misfitting Persons and Items by Revisions of the Spring 2014 Final Test* 139

Table 5.5 *Standardized Residual Variance (in Eigen-value Units) for the Spring 2014 Final Test* 146

Table 5.6 *Misfitting Persons and Items by Revisions of the Fall 2014 Final Test* 148

Table 5.7 *Standardized Residual Variance (in Eigen-value Units) for the Fall 2014 Final Test* 158

Table 5.8 *Misfitting Persons and Items by Revisions of the 2015 Placement Test* 160

Table 5.9 *Standardized Residual Variance (in Eigen-value Units) for the 2015 Placement Test* 166

Table 5.10 *Misfitting Persons and Items by Revisions of the Spring 2015 Final Test* 168

Table 5.11 *Standardized Residual Variance (in Eigen-value Units) for the Spring 2015 Final Test* 179

Table 5.12 *Misfitting Persons and Items by Revisions of the Fall 2015 Final Test* 181

Table 5.13 *Standardized Residual Variance (in Eigen-value Units) for the Fall 2015 Final Test* 186

Table 5.14 *Misfitting Persons and Items by Revisions of the Equated Tests* 194

Table 5.15 *Standardized Residual Variance (in Eigen-value Units) for the equated tests* 213

Table 5.16 *Number of Items Based on Test Type* 214

Table 5.17 *Number of Items Based on Japanese Loanwords* 215

Table 5.18 *Number, Average Item Difficulty Measure, and S.D.* 216

List of Figures

Figure 1.1. The number of 18-year-old students and the number of universities in Japan (MEXT, 2014a). 20

Figure 1.2. The percentage of private universities that do not meet intake quotas 20-

(Kawaijyuku Educational Institution, 2014). 21

Figure 1.3. Degree of implementation of AO Nyushi at Japanese universities 2000-2011 (MEXT, 2012a). 22

Figure 1.4. Percentages of types of entrance examination used at private universities in 2000 and 2012 (MEXT 2014a). 23

Figure 1.5. Number of universities offering supplementary lessons before the first semester starts in April (MEXT, 2012b). 24

Figure 1.6. Test construction phases (Carroll, 1985). 48

Figure 1.7. Specification for a test (retrieved from Hughes, 2003, p. 59-61, 140 and adapted for a reading test). 50

Figure 1.8. Test development (Bachman & Palmer, 1996, p.87). 52

Figure 2.1. Simple view of reading model (Gough & Wren, 1999; Kadota, 2012). 64

Figure 2.2. Reading processes that are activated when we read (Adapted from Grabe & Stoller, 2002; Grabe, 2009). 65

Figure 2.3. Novice reader’s text comprehension (Samuels, 2006, p.37). 66

Figure 2.4. Fluent reading in L1 (Samuels, 2006, p. 38). 66

Figure 2.5. Fluent reading in L2 (Kadota, Noro, & Shiki, 2010, p. 337). 66

Figure 2.6. Revised version of 2007 dual access model of word processing (Kadota, 2012, p.133). 77

Figure 2.7. Factors that affect language test scores (Bachman, 1990, p.165). 80

Figure 2.8. Gap between the number-right scores and logit scores. 86

Figure 2.9. A generic construct map in construct “X” (adapted from Lynch, 2003, p. 93 and Wilson, 2005, p. 27). 87

Figure 3.1. Number of words by syllables for words in the reference book. 98

Figure 3.2. Percentage of graphemes /a/, /e/, /i/, /o/, and /u/ in the list of English medical terms in the reference book. 99

Figure 3.3. The number of JE, NE, and NNE phonemes of the graphemes /a/, /e/, /i/, /o/, and /u/ in the English medical terms of the reference book. 100

Figure 4.1. The number of syllables in K1, K2, AWL, OFF words on the TOEFL. 120

Figure 4.2. Number of syllables per word on the TOEIC. 121

Figure 4.3. Percentage of graphemes /a/, /e/, /i/, /o/, and /u/ in the TOEFL and TOEIC reading sections. 121

Figure 4.4. The number of JE, NE, and NNE phonemes of the grapheme /a/ in the TOEFL and TOEIC reading sections. 122

Figure 4.5. The number of JE, NE, and NNE phonemes of a grapheme /e/ in the TOEFL and TOEIC reading sections. 123

Figure 4.6. The number of JE, NE, and NNE phonemes of the grapheme /i/ in the TOEFL and TOEIC reading sections. 123

Figure 4.7. The number of JE, NE, and NNE phonemes of the grapheme /o/ in the TOEFL and TOEIC reading sections. 124

Figure 4.8. The number of JE, NE, and NNE phonemes of the grapheme /u/ in the TOEFL and TOEIC reading sections. 124

Figure 4.9. Number of syllables per word in the reference books' vocabulary list, on the TOEIC and on the TOEFL. 125

Figure 4.10. The percentage of JE, NE, and NNE phonemes of graphemes /a/, /e/, /i/, /o/, and /u/ in the English medical terms of the reference book, the TOEFL, and the TOEIC. 126

Figure 5.1. Mean square statistic (MNSQ) values and their meanings (Linacre, 2013) 135

Figure 5.2. Item and person map for the Spring 2014 final test (original analysis*).* 141

Figure 5.3. Item and person map for the Spring 2014 final test (1st revision*).* 143

Figure 5.4. Item and person map for the Spring 2014 final test (2nd revision*).* 144

Figure 5.5. Item and person map for the Fall 2014 final test (original analysis*).* 150

Figure 5.6. Item and person map for the Fall 2014 final test (1st revision*).* 152

Figure 5.7. Item and person map for the Fall 2014 final test (2nd revision*).* 154

Figure 5.8. Item and person map for the Fall 2014 final test (3rd revision*).* 155

Figure 5.9. Item and person map for the Fall 2014 final test (4th revision*).* 157

Figure 5.10. Item and person map for the 2015 placement test (original analysis*).* 162

Figure 5.11. Item and person map for the 2015 placement test (1st revision*).* 164

Figure 5.12. Item and person map for the 2015 placement test (2nd revision*).* 165

Figure 5.13. Item and person map for the Spring 2015 final test (original analysis*).* 170

Figure 5.14. Item and person map for the Spring 2015 final test (1st revision*).* 172

Figure 5.15. Item and person map for the Spring 2015 final test (2nd revision*).* 174

Figure 5.16. Item and person map for the Spring 2015 final test (3rd revision*).* 176

Figure 5.17. Item and person map for the Spring 2015 final test (4th revision*).* 178

Figure 5.18. Item and person map for the Fall 2015 final test (original analysis*).* 183

Figure 5.19. Item and person map for the Fall 2015 final test (1st revision*).* 185

Figure 5.20. Differential Test Functioning for 14SF and 15SF (1st) 187

Figure 5.21. Differential Test Functioning for 14SF and 15FF (1st) 188

Figure 5.22. Differential Test Functioning for 15SF and 15FF (1st) 189

Figure 5.23. Differential Test Functioning for 14SF and 15SF (2nd) 190

Figure 5.24. Differential Test Functioning for 14SF and 15FF (2nd) 191

Figure 5.25. Differential Test Functioning for 15SF and 15FF (2nd) 192

Figure 5.26. Item and person map for the equated tests (original analysis*).* 196

Figure 5.27. Item and person map for the equated tests (1st revision*).* 198

Figure 5.28. Item and person map for the equated tests (2nd revision*).* 200

Figure 5.29. Item and person map for the equated tests (3rd revision*)*. 202
Figure 5.30. Item and person map for the equated tests (4th revision*)*. 204
Figure 5.31. Item and person map for the equated tests (5th revision*)*. 206
Figure 5.32 Item and person map for the equated tests (6th revision*)*. 208
Figure 5.33 Item and person map for the equated tests (7th revision*)*. 210
Figure 5.34 Item and person map for the equated tests (8th revision*)*. 212
Figure 5.35 Item difficulty by test type on equated tests. 215
Figure 5.36 Item difficulty by loanwords on the equated tests. 216
Figure 5.37 Beginner level items' ratio of test type. 217
Figure 5.38 Medium level items' ratio of test type. 217
Figure 5.39 Advanced-level items' ratio of test type. 218

Table of Contents

Abstract............ 1
List of Abbreviations 3
Declaration............ 5
Copyright............ 6
Acknowledgements............ 7
Dedication............ 8
List of Tables............ 9
List of Figures............ 10
Chapter 1: Introduction............ 16
1.1 The purpose of the study............ 16
1.2 Statement of the problems in tertiary-level institutions in Japan............ 17
1.2.1 Changes of educational environment in Japanese tertiary-level institutions............ 17
1.2.2 Two problems at tertiary-level institutions in Japan: admission policy and the decline of students' academic literacy............ 21
1.2.3 English-language education at tertiary-level institutions in Japan............ 30
1.2.4 Commercially produced, norm-referenced, English-language proficiency tests for placement purposes............ 35
An overview of commercially produced, norm-referenced, English-language proficiency tests.. 37
1.2.5 The development of an in-house placement test and item bank............ 46
1.3 Statement of problems of an English-language programme in a Japanese tertiary-level institution............ 53
Chapter 2: Literature Review............ 57
2.1 Fundamental consideration of the characteristics of measurement............ 57
2.1.1 The limitations of specification............ 57
2.1.2 The limitations of observation and quantifications............ 59
2.2 The nature of reading process............ 61
2.2.1 Process models of reading............ 62
2.2.2 Component model............ 64
2.3 Assessing the usefulness of a language test............ 79
2.3.1 Reliability............ 79
2.3.2 Validity............ 80
2.4 Test theories............ 82
2.4.1 Classical test theory............ 82
2.4.2 Item response theory............ 85

2.4.3 IRT models and implications for use in individual tertiary-level institutions 88
2.4.4 The difference between the Rasch and the one-parameter IRT model 89
Chapter 3: Curriculum Analysis..91
3.1. Materials..91
3.2. Research process..92
3.3 Results..96
3.3.2 List of English medical terms in the reference book 98
3.4 Summary findings and conclusion 100
Chapter 4: Commercially Produced English-language Tests...102
4.1 Materials 102
4.2 Research process...103
4.3 Results..110
4.3.1 The official guide...110
4.3.2 Orthographic and phonological features of the TOEFL and the TOEIC.................119
4.3.3 Syllables..120
4.3.4 Graphemes...121
4.3.5 Comparison with L1 grapheme–phoneme correspondence.....................................121
4.3.6 Differences between the reference book's vocabulary, the TOEFL's and the TOEIC's 124
4.4. Summary findings and conclusion..127
Chapter 5: Item Bank Development 128
5.1 Test development 128
5.1.1 Materials and participants 129
5.1.2 Analyses 133
5.2.1 The Spring 2014 final test...137
5.2.2 Basic feature of the Spring 2014 final test...137
5.2.3 Revision process of the Spring 2014 final test 138
5.2.4 Unidimensionality 145
5.3.1 The Fall 2014 final test 145
5.3.2 Basic feature of the Fall 2014 final test 146
5.3.3 Revision process of the Fall 2014 final test...146
5.3.4 Unidimensionality 157
5.4.1 The 2015 placement test 158
5.4.2 Basic feature of the 2015 placement test 159
5.4.3 Revision process of the 2015 placement test...159
5.4.4 Unidimensionality 166
5.5.1 The Spring 2015 final test...166
5.5.2 Basic feature of the Spring 2015 final test...167
5.5.3 Revision process of the Spring 2015 final test 167
5.5.4 Unidimensionality 178
5.6.1 The Fall 2015 final test 179
5.6.2 Basic feature of the Fall 2015 final test 179
5.6.3 Revision process of the Fall 2015 final test...180
5.6.4 Unidimensionality 185
5.7.1 Equating test items..186
5.7.2 Materials and procedure 186
5.7.3 First analysis...187
5.7.4 Second analysis 189
5.8.1 The equated tests..192
5.8.2 Basic feature of the equated tests..192
5.8.3 Revision process of the equated tests 193
5.8.4 Unidimensionality 212
5.9.1 Pre-test prediction...213
5.9.2 Empirical results...215
5.10 Item banking...216

5.11 Summary findings and conclusion 218
Chapter 6: Discussion 220
6.1 Research question 1 220
6.2 Research question 2 222
6.3 Research question 3 223
6.4 Overall discussion 224
Chapter 7 Conclusion 225
7.1 Contribution of the thesis 225
7.2 Implications 226
7.3 Limitations 226
7.4 Suggestions for future study 227
Bibliography 229

Chapter 1: Introduction

1.1 The purpose of the study

The primary purpose of the present study is to validate the development of an item bank that includes an array of test items designed to assess neglected areas of proficiencies such as orthographical processing skills and phonological processing skills in commercially produced English proficiency tests but which are crucial for a successful learning at the Faculty of Nursing. By developing such an item bank, the paper aims to systematize the process of developing an in-house English placement test administered at a Faculty of Nursing at the beginning of students' university careers. Such an item bank for an in-house English placement test would enable the faculty to assign students to appropriate English-language classes, which would lead to students' successful completion of the English-language program in the tertiary-level institution.

The paper will first explore the orthographic and phonological features of the faculty's English curriculum and materials. Next, the paper will explore the features that are infrequently addressed in commercially produced English proficiency tests and determine whether these features are included in the faculty's English curriculum and materials. Finally, the paper will describe the process of developing suitable in-house placement test items, especially for orthographic and phonological skills, and explore whether these test items correspond to the requirements of the faculty's English curriculum.

As Cronbach (1988) pointed out, "Validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to guard against adverse consequences from meanings of the word validation, but you cannot deny the obligation" (p. 6). Therefore, this paper is written in response to the following research questions:

1. What kinds of abilities are required in the faculty's curriculum and English materials regarding orthographic knowledge and phonological awareness?
2. What kinds of test items are used and constructs measured in practice tests of commercially produced English-language proficiency tests and how well do

those abilities reflect the content of the faculty's curriculum?

3. Whether the development of an item bank for the English-language placement test a valid process and how well the test items reflect the content of the faculty's curriculum?

1.2 Statement of the problems in tertiary-level institutions in Japan

In this section, the paper will offer a brief overview of some of the major problems regarding English-language education at tertiary-level institutions in Japan. First, the paper will describe recent reforms in the educational environment in Japan, especially at tertiary-level institutions. Then the paper will report on the revisions and changes in admission policies in the past three decades as a result of the changes in the educational environment and will examine some of the problems created by such changes. The paper will then consider one of the most important issues regarding English-language education: the problems entailed in relying on commercially produced, norm-referenced, English-language proficiency tests and the feasibility of developing an item bank for in-house placement tests as an alternative. Finally, the paper will describe some of the problems a tertiary-level institution encounters when it uses a commercially produced, norm-referenced, English-language proficiency test as well as the ways such problems are being addressed.

1.2.1 Changes of educational environment in Japanese tertiary-level institutions

As in many other countries, with the advance of technology and the development of borderless societies in terms of language, culture, or economy, the environment surrounding Japanese universities is changing quite rapidly. The circumstances surrounding education have changed greatly from the time when the original Basic Act on Education was promulgated and put onto effect in March 1947.

Table 1.1
Situation at the Time of the Establishment of the Fundamental Law of Education Compared to the Present (MHLW, 2014; Portal Site of Official Statistics of Japan (e-Stat), 2014a, 2014b)

Average life expectancy		
Male	50.06 years (1964)	80.21 years (2013)
Female	53.96 years (1964)	86.61 years (2013)
Overall fertility rate	4.54 (1964)	1.43 (2013)
Percent of population aged 65 or above	5.72% (1960)	26.0% (2010)
High school attendance rate	42.5% (1950)	98.4% (2014)

University, etc., attendance rate	7.9% (1954)	51.5% (2014)

Even a brief look at Table 1.1 shows that there has been a demographic change in Japan within the past 50 years; an ageing population and popularization of higher education has increased the demands on high academic qualification, where approximately half the people enter universities. It also demonstrates that graduation from tertiary institutions no longer guarantees the elite track. Since the competition among university graduates has increased, there is more pressure on universities to provide a better education. The number is quite different from that of the students at the start of this century, when tertiary-level education was limited only to the leisured elite. According to Takemae (2009), university-level education, once recognised as a "privilege," came to be thought of as a "right" and now is perceived as "compulsory." Students are obliged to participate in compulsory classes to get the credits to be a university graduate, a minimal level certificate necessary to get a proper job in today's competitive employment front.

However, the increase in the percent of students enrolling universities does not correspond to the maintenance of the quality of education provided at Japan's tertiary-level institutions. On an international scale, the quality of education in Japanese universities is not sufficient. The *Times Higher Education* World University Rankings 2015-2016 ranked the world's top 400 universities (2016), and only two of Japan's nearly 800 universities are ranked within the top 100 universities in the world. This means that the mere increase in universities does not guarantee the improvement of the quality of education provided at the universities.

Table 1.2
The World's Rankings 2015-2016 Top 400 (THES, 2016)

Ranking	University	Country
1	California Institute of Technology	US
6	Harvard University	US
2	University of Oxford	US
3	Stanford University	US
5	Massachusetts Institute of Technology	US
7	Princeton University	US
4	University of Cambridge	UK
13	University of California, Berkeley	US
10	The University of Chicago	US
8	Imperial College London	UK

12	Yale University	US
16	University of California, Los Angeles	US
15	Columbia University	US
9	ETH Zürich – Swiss Federal Institute of Technology Zürich	Switzerland
11	Johns Hopkins University	US
17	University of Pennsylvania	US
20	Duke University	US
14	University College London	UK
18	Cornell University	US
19	University of Toronto	CA
	～～	～～
43	University of Tokyo	Japan
88	Kyoto University	Japan

The media tends to give the results huge coverage, resulting in national, societal, and economic pressure to improve the students' quality of learning provided at the Japanese tertiary educational level institutions. In order to respond to such pressure, several reforms, including the revision of the Basic Act on Education, have taken place. Among the rapid and successive reforms initiated predominantly by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), such as the privatization of national universities and the prioritized allocation of funding based on the superiority of research results, the deregulation of the Standards for the Establishment of Universities 1991 loosened the convention on higher-education enterprises and may have urged universities to work toward amendments in several aspects, including the loosening of admission policies. The drastic changes have a great impact on the number of universities in Japan, which increased by 34.8 percent between 1990 and 2014. The increase has more profound effect on the tertiary-level institutions when we compare this with the number of 18-year-old students that is dropping gradually as a result of a lower birth-rate (see Figure 1.1).

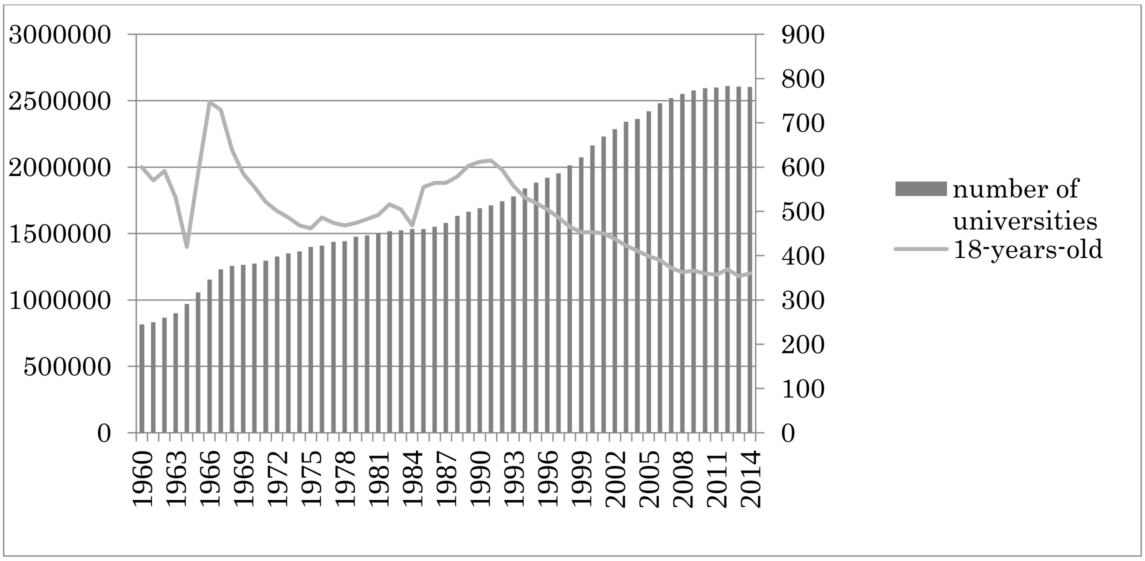


Figure 1.1. The number of 18-year-old students and the number of universities in Japan (MEXT, 2014a).

In 2006, the number of university applicants and the likely number of freshmen to be accepted had become almost the same, the situation named "*Zennyu-jidai*" (the age of universal university admissions). This meant that all high-school students were accepted into a university, if they are not particular about their choice of study. Although some universities, predominantly those with strong financial standings, decreased their enrolment numbers, it has been reported that more than 45 percent of Japan's 578 private universities did not meet their intake quotas in 2014. As the Figure 1.2. suggests, the rate of capacity utilization is dropping each year. This means that nearly half of the private universities that constitute nearly three-fourths of all four-year universities in Japan do not have enough students.

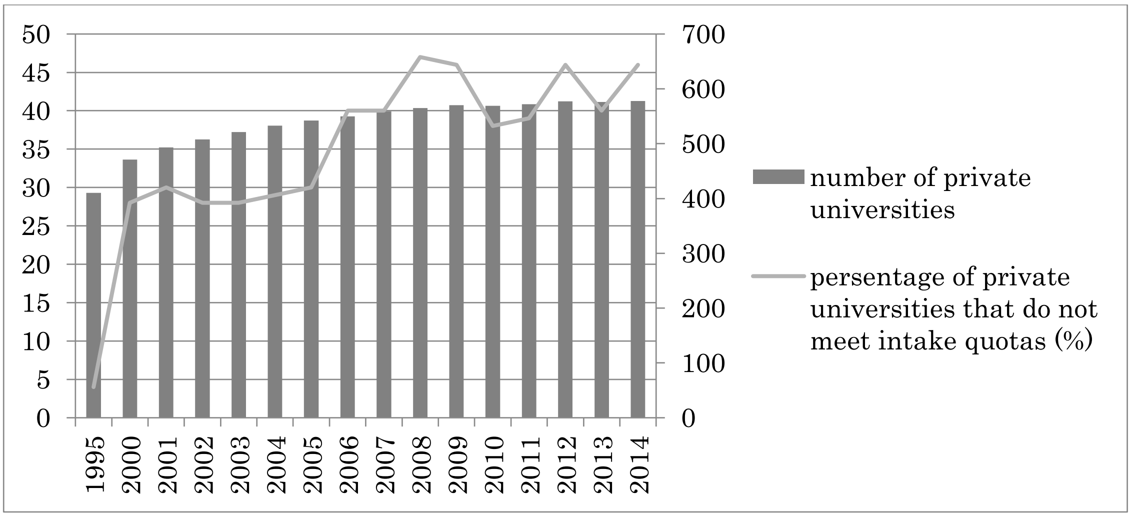


Figure 1.2. The percentage of private universities that do not meet intake quotas (Kawaijyuku Educational Institution, 2014).

1.2.2 Two problems at tertiary-level institutions in Japan: admission policy and the decline of students' academic literacy

In order to get through a difficult situation, many universities decided to change admissions selection systems to get hold of incoming students as a first step (Sugiyama, 2004). Some university staff had pointed out that, overall, students' academic skills would decline if the entrance exams were simplified just to keep up enrolment. However, many universities now are vigorously recruiting students through a new entrance examination system that exempts students from the traditional screening process, *AO Nyush,* introduced by MEXT in 1990. This system is an admissions office examination that corresponds to an admission system based on recommendations—*Suisen Nyushi*—but fundamentally different because it does not require recommendation letters from high schools.

Suisen Nyushi

Suisen Nyushi is one of the conventional admission systems in Japan that is widely used in many departments that educates students who are to become professionals in medical services. Since professionals in medical services are required to have high standard of ethical views, the departments need to evaluate not only scholastic but also vocational aptitude of students by drawing in the views of high school teachers regarding the personality of a candidate. Although the screening procedure becomes time-consuming because of the additional selection criteria, the department can obtain personality and motivation information about the candidate to determine his/her suitability to professional ethics.

AO Nyushi

While the conventional screening processes often involve written tests, *AO Nyushi,* introduced by MEXT in 1990, uses interviews and school reports as major sources of the assessment procedure. According to the admissions office at Hokkaido University, one of six former imperial universities that became famous for implementing *AO Nyushi* in 2006, such criteria enable admissions to identify "promising people at the entrance examination," people who may not have appropriate academic abilities but have other talents suiting the university's educational purposes. Sugiyama (2004) states that the original goal of *AO Nyushi* was to stimulate the university by admitting highly motivated students who lack satisfactory academic abilities instead of less-motivated students with an appropriate level of academic ability. An article in *The Japan Times* published February 16, 2008, goes even further by claiming that the conventional screening process at tertiary-level institutions has caused many innocent candidates to engage in meaningless examination preparation. The author argues that the current system that allows students to matriculate into universities without written tests is "somewhat more flexible and slightly more human." While faculty have raised several concerns regarding the academic ability of such students, nearly 70 percent of Japanese universities used *AO Nyushi* in 2011.

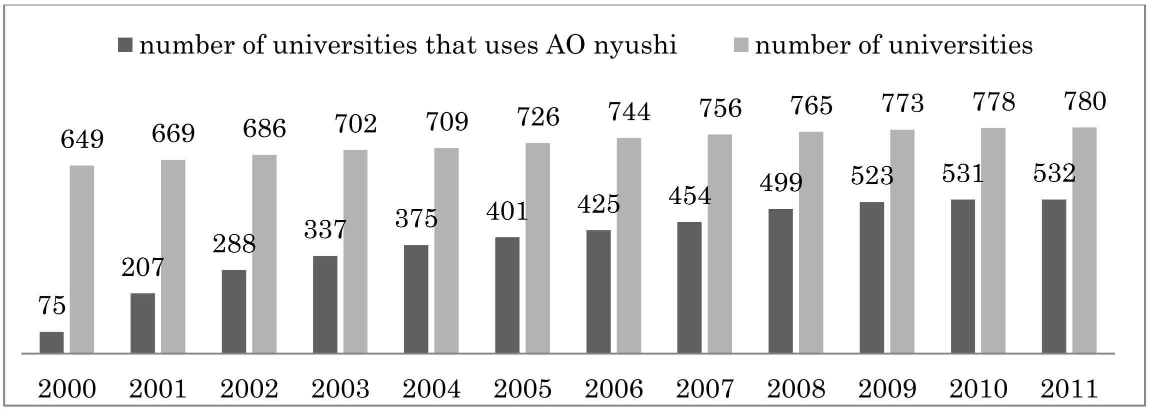


Figure 1.3. Degree of implementation of AO Nyushi at Japanese universities 2000-2011 (MEXT, 2012a).

Criticisms of AO Nyushi

According to a report by MEXT (2014a), approximately 250,000 students enter Japanese universities via either *AO Nyushi* or *Suisen Nyushi* in 2012. This figure accounts for approximately 43 percent of total university enrolment in Japan.

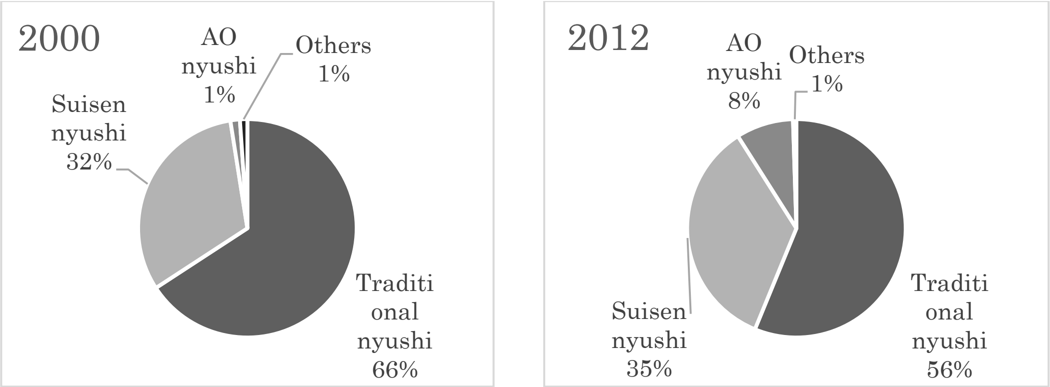


Figure 1.4. Percentages of types of entrance examination used at private universities in 2000 and 2012 (MEXT, 2014a).

Regarding this situation, MEXT has warned that overusing *AO Nyushi* and *Suisen Nyushi* should be refrained at the tertiary level, since students admitted to universities via *AO Nyushi* tend both to fail to earn enough credits and to stay on an extra year, or worse, drop out because they lack a basic level of academic ability. Students who entered university via *AO Nyushi* tend to show academic deficiency, especially in subjects included on traditional entrance examinations such as English, mathematics, and physics. The problem lies in the fact that universities often do not require *AO Nyushi* applicants to submit school records. Therefore, these students tend to avoid taking subjects fundamental to surviving collegiate life, but which seemed laborious in high school. As a result, there are students who have not taken differential and integral calculus, yet who major in mathematics, as well as students who have not taken Japanese classics, but who nevertheless major in Japanese literature. In order to cope with such situations, several universities now offer supplementary lessons for candidates who matriculated into university using *AO Nyushi* before the first semester starts in April. According to MEXT (2014a), approximately 46% of universities are offering supplementary lessons for *AO Nyushi* entrants, an increase of 210% since 2001.

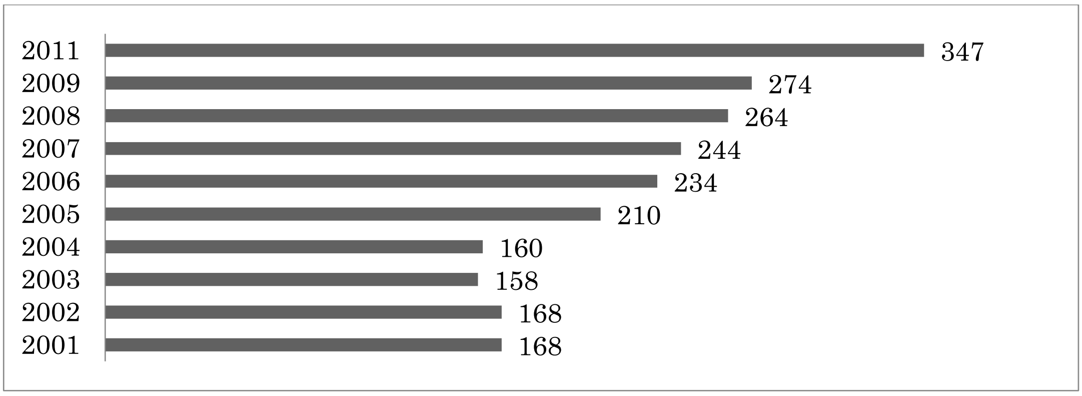


Figure 1.5. Number of universities offering supplementary lessons before the first semester starts in April (MEXT, 2012b).

Along with the emasculation of university admission, there came another attack to the university's teaching staff regarding the academic literacy of incoming students. The decline of Japanese students' academic literacy became apparent when the results of two international academic literacy surveys, Program for International Student Assessment (PISA) conducted by the Organization for Economic Cooperation and Development (OECD) (2007) were announced. Introduced in 2000, PISA is a worldwide test that assesses 15-year-old students' scholastic performance in reading, mathematics, and science literacy every three years.

As shown in Table 1.3, Japan's rankings always dropped in all three subjects until a slight recovery especially in a reading literacy section in 2009. Although the results of two surveys in 2000 and 2003 had shown the symptoms of a decline in Japanese students' academic performance, both the government and ordinary citizens appeared to have been taking optimistic views about the results. The optimism may have resulted from the fact that Japanese high school students achieved first place in the mathematics test ranking in 2000.

Table 1.3

PISA Results on Reading Literacy, Mathematics, Science in 2000, 2003, 2006 and 2009 (OECD, 2007, 2009, 2010a, 2010b, 2010c).

Reading Literacy

2000		2003		2006		2009	
1 Finland	546	1Finland	543	1 South Korea	556	1 Shanghai-China	556
2 Canada	534	2 South Korea	534	2 Finland	547	2 Korea	539
3 New Zealand	529	3 Canada	528	3 Hong Kong-China	536	3 Finland	536
4 Australia	528	4 Australia	525	4 Canada	527	4 Hong Kong-China	533
5 Ireland	527	4 Liechtenstein	525	5 New Zealand	521	5 Singapore	526
6 South Korea	525	6 New Zealand	522	6 Ireland	517	6 Canada	524
7 United Kingdom	523	7 Ireland	515	7 Australia	513	7 New Zealand	521
8 Japan	**522**	8 Sweden	514	8 Liechtenstein	510	**8 Japan**	**520**
9 Sweden	516	9 Netherlands	513	9 Netherlands	508	9 Australia	515
10 Austria	507	10 Hong Kong	510	10 Sweden	507	10 Netherland	508

Mathematics

2000		2003		2006		2009	
1 Japan	**557**	1 Finland	544	1 Taiwan	549	1 Shanghai-China	600
2 South Korea	547	2 South Korea	542	2 Finland	548	2 Singapore	562
3 New Zealand	537	3 Netherlands	538	3 Hong Kong-China	547	3 Hong Kong-China	555
4 Finland	536	**4 Japan**	**534**	3 South Korea	547	4 Korea	546
5 Australia	533	5 Canada	532	5 Netherlands	531	5 Chinese Taipei	543
6 Canada	533	6 Belgium	529	6 Switzerland	530	6 Finland	541
7 Switzerland	529	7 Switzerland	527	7 Canada	527	7 Liechtenstein	536
8 United Kingdom	529	8 Australia	524	8 Macao	525	8 Switzerland	534
9 Belgium	520	9 New Zealand	523	8 Liechtenstein	525	**9 Japan**	**525**
10 France	517	10 Czech Republic	516	**10 Japan**	**523**	10 Canada	527

Science

2000		2003		2006		2009	
1 Korea	552	1 Finland	548	1 Finland	563	1 Shanghai-China	575
2 Japan	**550**	**1 Japan**	**548**	2 Hong Kong	534	2 Finland	554
3 Finland	538	3 Hong Kong	539	3 Canada	534	3 Hong Kong- China	547
4 United Kingdom	532	4 South Korea	538	4 Taiwan	532	4 Singapore	542
5 Canada	529	5 Liechtenstein	525	5 Estonia	531	**5 Japan**	**539**
6 New Zealand	528	5 Australia	525	**5 Japan**	**531**	6 Korea	538
6 Australia	528	5 Macao	525	7 New Zealand	530	7 New Zealand	532
8 Austria	519	8 Netherlands	524	8 Australia	527	8 Canada	529
9 Ireland	513	9 Czech Republic	523	9 Netherlands	525	9 Estonia	528
10 Sweden	512	10 New Zealand	521	10 Liechtenstein	522	10Australia	527

However, the 2006 results have finally shown without a doubt that Japanese students' academic abilities have gradually but steadily been dropping. Some researchers argue that the drop of Japanese students' academic abilities can be attributed to the introduction of the two national curricula, the *Course of Study*, for junior high schools and senior high schools in 1989 and 1998/1999. The *Course of Study*, devised by the MEXT, is a national curriculum that has strong control over teaching and learning in primary and secondary schools in Japan. The textbooks used in the public schools need to pass MEXT's approval. The *Course of Study* is revised every 10 years which has an immense impact on teaching and learning in Japan. The Table 1.4 below offers a brief overview of the past three decades of changes to the *Course of Study* in Japan.

Table 1.4

The Three Versions of Course of Study by Their Year of Introduction, the Year of Implementation, and Major Changes

	Course of Study								
	1992 version			20002 version			2012 version		
	Primary	Junior high	High school	Primary	Junior high	High school	Primary	Junior high	High school
The year of introduction	1989			1998		1999	2008		2009
The year of implementation	1992	1993	1994	2002		2003	2012		2013
Major changes	Reduction of content by 30%			Reduction of content and class hours by approximately 30 %			Addition of content and class hours by approximately 30 %		

The two *Course of Study* have been notorious for their drastic reduction of content and class hours under the phrase of *yutori-kyoiku (yutori education)*, relaxed education or pressure-free education. As the phrase suggests, the main focus of these two *Course of Study* was to reduce student workload so they have more spare time for activities outside the classroom. According to Otsu (2005), approximately 30 percent of contents, especially the amount of vocabulary used and grammar points, were deleted from English as well as other subjects. Although pressure-free education was favoured enthusiastically by the government and parents who had criticised the traditional teaching practices as overemphasizing rote learning or cramming, some researchers have criticised the reduction as the renunciation of public education (Takahashi, 2005). The results also imply that students can no longer expect to receive sufficient education from public education systems and are expected to seek alternative ways to receive an education with extra monetary obligations if they are to succeed in entrance examinations and job hunting.

While the PISA survey results described above are only concerned with students below the tertiary level, the decline of academic literacy at junior and senior high schools has a profound effect on teaching and learning at Japanese universities.

The results of international surveys have been taken seriously by both parties, and in 2008, the ministry announced a change in educational policy, partially retracting some of the educational reforms from the past decade. For example, several primary schools and junior high schools have started to give regular classes on Saturdays and more importantly, many schools have started to teach some of the items removed from the compulsory curriculum.

While benefits of the abrupt turnabout in the educational system in primary and secondary schools has started to be acknowledged at primary and secondary education, the tertiary-level institutions still have to remain in the present conditions at least for a decade to get the results of the change. Meanwhile, the tertiary-level institutions are accepting the influx of students who have received the reduced curriculum. Several educators and researchers at tertiary-level institutions claim that the academic skills among Japanese university students have been declining (White, Eguchi, Kawanaka,

& Henneberry, 2005). According to White, Eguchi, Kawanaka, and Henneberry (op. cit.), universities are no longer limiting their support services to psychology and career counselling, but also to study support services. Several measures have been taken to provide academic support to students.

1.2.3 English-language education at tertiary-level institutions in Japan

English is a compulsory subject for Japanese students in secondary education and most university entrants have received the minimum of six years of English-language instruction when they enter universities. Moreover, the MEXT has decided on a plan to facilitate English abilities so they can use English at work after they graduate. In 2003, the MEXT proposed the Action Plan to Cultivate Japanese with English Abilities. The report says that, "For children living in the 21st century, it is essential for them to acquire communication abilities in English as a common international language… (but) due to the lack of sufficient ability, many Japanese are restricted in their exchanges with foreigners and their ideas or opinions are not evaluated appropriately… in order to make such improvements bear fruit, it is necessary to carry out simultaneously a number of different measures" (p. i). These include improving teaching methods, the teaching ability of teachers, and the selection system for school and university applicants, as well as creating better curricula (MEXT, op. cit.). The goal of English-language education at higher level is that "graduates can use English in their work" (p. i). In order to do so, the action plan states "each university should establish attainment targets from the viewpoint of fostering personnel who can use English in their work" (MEXT, op. cit.).

Contrary to the government's expectation, Japanese students were ranked 40th amongst 48 countries that had taken the Test of English for International Communication (TOEIC) in 2013 and 10th amongst 15 Asian countries (ETS, 2013b). Several researchers and practitioners, including Ford (2009) and Takemae (2009), have argued that the English ability of incoming students is not an exception to the trend described above where a gradual decline of academic literacy among secondary students is taking place. Several practitioners and researchers also claim that the university-level students' English proficiency level is not only declining, but also that the student discrepancies between the high-achievers and low-achievers are becoming great (Sato, Nakagawa, & Yamana, 2007). Sato, Nakagawa, and Yamana, (op. cit.) for example, report that of the 104 test items that assess knowledge of junior high -school level English grammar, 24 items show less than 50% correctness. Moreover, the discrepancy of the percentage of correct answers is large between advanced-level classes and elementary-level classes.

As has been described in the earlier section, the increase in number of private universities that use *AO Nyushi* and *Suisen Nyushi* implies that approximately 43 % of total university enrolment in Japan may have been admitted without taking English examinations The problem with this number lies not only with entrance examination but also the reality that these students had not taken English classes seriously in their high-school years. Reducing the number of subjects in an entrance examination is a serious problem in Japan. This is because even high schools encourage students to focus their attention on studying subjects that are included in the entrance examinations. In 2006, nearly 20% of senior year students in private high schools and 8% in public high schools (about 80,000 students) had been accused of not completing the compulsory subject of world history. Those high schools had not set up the subject, although *the Course of Study* (2003) states that all high school students are to be taught 35 hours of world history. This event caused a big discussion involving the MEXT and the board of education, which had been accused of implementing a “yutori” type of *Course of Study* (MEXT, 2003a). The major criticism had to do with a contradiction in the *Course of Study*’s requirements, which forced high school to raise the ratio of successful applicants accepted into good universities. In order to achieve the goal, the high schools needed to reduce teaching a number of “unnecessary” compulsory subjects, such as world history and earth science, and focus more time on subjects that were often included in entrance examinations. The scandal of dismissing the teaching of compulsory subjects in high school also implied a drop in the relative role English plays in classrooms in Japan.

Regarding the decline of incoming students’ English-language proficiency, Takemae (2009) argued that many students face difficulties when asked to read textbooks with orthographically and phonologically difficult words. Although the 2012 version of *Course of Study* shows (MEXT, 2011) a drastic increase in its content, for example, students learn more words at junior high schools than in the 2002 version (1,200 words vs. 900 words), the actual teaching of phonological awareness and orthographic processing skills is relatively limited. The teaching of phonological awareness and orthographic processing skills appears to be, in fact, encouraged to a certain extent in the *Course of Study*. In the “reading activity” section, for example, students are encouraged to acknowledge the shapes and the meanings of alphabetic letters and to pronounce the words based on the rules of grapheme–phoneme correspondence. Students also need to acknowledge that English vowel and consonants differ in number and type than in Japanese. Moreover, a word may have a sequence of consonants (e.g. “street”) or end with a consonant (e.g. “school”), which does not happen in Japanese (e.g. /sutori:to/). The *Course of Study* further states that students should not only acquire knowledge of the pronunciation of individual words but also sequences of words, such as liaison, in which the sound of a consonant becomes unpronounced at the end of a word due to a vowel at the beginning of the next word. However, as Kameyama (1992) pointed out, Japanese students have fewer opportunities to learn the relationships between English writings and pronunciation. While researchers like Smith (2012) and Noguchi (2014) demonstrated the need to teach the orthographical and phonological differences between English and Japanese, the people involved in English in Japan lack the materials to teach the relationships between English orthography and phonology. This is reinforced by the fact that no textbook adopted by the public junior high schools in Tokyo since 2012 include a section on relationships on English orthography and phonology. A survey of the textbooks adopted by the junior high schools in 11 wards and districts of Tokyo, (including Chiyoda, Chuo, Minato, Bunkyo, Taito, Shinagawa, Oota, Setagaya , as well as the Izu islands and the Ogasawara islands) provided by the Japan Textbook Distributors Association (http://www.text-kyoukyuu.or.jp/kaiin/tokuyaku13.html) found that only four English textbooks are used in these areas. Table 1.5 shows this list of textbook and publishers.

Table 1.5

The English Textbooks Adopted in 11 Wards and Islands in Tokyo Since 2012

Title	Publisher	Year of Publication
New Horizon	Tousho	2011
Sunshine English Course	Kairyudo	2012
New Crown	Sunseido	2012
One World	Kyosyutu	2012

Most activities and exercises encouraged in the textbook are to motivate students to speak in several communicative situations, such as asking questions in hotels, restaurants, and movie theatres; ordering products on the phone; introducing their family members to a friend; and talking with friends about international issues as well as listening to authentic English dialogues about these topics. There are no sections that deal with the relationship between English orthography and pronunciation in these four textbooks, although all textbooks encourage readers to purchase reference materials like CDs, DVDs, and flash cards with English words on one side and a Japanese translation on the other. In the classroom, a teacher shuffles a bundle of these cards, chooses one, and then asks a student to correctly pronounce and define the word. These cards do not have phonetic symbols but only a Japanese translation. If students want to know whether their pronunciation is correct, they must listen to the attached CDs. The teaching of phonetic symbols was not encouraged in the previous *Course of Study* because learners would focus more on consulting dictionaries to check the pronunciation rather than listening to native speakers' authentic pronunciation, a habit which can demotivate students to speak in English (MEXT, 2011).

Moreover, the introduction of romaji, the foreign style of writing Japanese language that uses an alphabet (Suski, 1931), was moved up from the fourth year to the third year of primary school in the new *Course of Study*. Since the introduction of English reading education (especially, the reading of the English alphabet) still begins in junior high school, Japanese students learn to read in the romaji alphabet prior to the English alphabet. While the relationships between the romaji alphabet and English alphabets will be described in detail in the next chapter, there is one crucial problem about teaching of the romaji alphabet prior to the English alphabet: Japanese learners of English can be considerably confused by the fact that the romaji alphabet has one-to-one relationships between grapheme and phoneme while English does not (Kay, 1995; Ohata, 2004).

The diversification of screening methods of entrance examinations into tertiary-level institutions as well as the diversification of students' academic abilities at the beginning of collegiate life in Japan also has created the need to provide various forms of English-language support. There is a relatively large variety of students, especially of natural and life science majors, who have been admitted to an institution but whose

English proficiency is not sufficient to meet the demands of their degree studies. Without any moves to make English tests mandatory for entry, many English programs at tertiary-level institutions needed to provide various forms of ongoing language support for those students who do not meet the required level.

Although various forms of learning support may be available to the students, the issue is how to identify the students who need such assistance and to what extent they should be required to take advantage of it. One way to address the condition is to introduce some form of diagnostic assessment comparable to placement tests. For example, Fulcher's report (1997) that described the development of English-language placement tests for all incoming students at the University of Surrey in the UK gives insightful accounts into the design of an investigation into the reliability and validity of in-house placement tests.

The Table 1.6 below shows a brief overview of research done on the degree of implementation of English-language placement tests at Japanese universities. One of the earlier studies was performed by Koike (1990). In his study, only 3.8 % of respondents answered that they had used an English placement test in their institution. The number makes a good contrast to that of Shimizu (2003) in which 48% of universities answered that they used an English-language placement test at their institution. While the number of universities that use an English-language placement test appears to have increased in a substantially rapid manner over the last two decades, there seems to be a trend in the type of English-language test tertiary-level institutions use for placement purposes. Otani, Yokoyama, and Bradford-Watts (2014) have organised a more in-depth study on the type of test the organizations use, and the pros and cons of those tests. According to their study, of the 16 universities investigated, four universities reported that they are using an in-house English-language placement test, while ten universities responded that they are using a commercially produced English-language placement test.

Table 1.6

Studies on the Number of Universities that Use English-language Placement Tests

Researcher	Number of universities /	Number of universities /	Number of universities /

	people addressed	respondents responding to the research	respondents that use English placement test
Koike (1990)	NA	959 people	36 (3.8%)
Shimizu (2003)	616 universities	200 universities	96 (48%)
Sugimori (2003)	400 people	208 people	131 (63%)
Otani, et al. (2014)	16 universities	16 universities	12 (75%)

The literature on placement testing provides two ways to approach the issue of selecting more appropriate assessment tools for each institution. One is the use of commercially produced , norm-referenced, English-language proficiency tests such as the International English-Language Testing System (IELTS), the Test of English as a Foreign Language (TOEFL), and TOEIC, and the other is to adopt a test developed inside the individual institution. Each test has its own advantages and disadvantages, making the classification of students into the appropriate level even more complex. The following section will briefly describe the basic characteristics of some of the tests frequently used in Japanese universities and argues the appropriateness of each test using their advantages and disadvantages.

1.2.4 Commercially produced, norm-referenced, English-language proficiency tests for placement purposes

Before discussing problems of using commercially produced, norm-referenced, English-language proficiency tests for placement purposes, the section will first briefly describe what the literature has said about the similarities and differences between norm-referenced tests and criterion-referenced tests. It will also offer an overview on some of the commercially produced, norm-referenced, English-language proficiency tests frequently used in Japanese tertiary-level institutions. The paper will then describe some of the problems related to using norm-referenced, English-language tests for placement purposes in language programs.

Norm-referenced tests and criterion-referenced tests

The major differences between the two types of tests involve their purposes, test construction, administration, and scoring (see Table 1.7). Regarding purposes, norm-referenced tests examine a student's relative position within the examined group, while criterion-referenced tests assess the student's level of achievement in relation to criteria or external standards. Likewise, the test items used in norm-referenced tests are not related to the course objectives, while criterion-referenced test items need to be related to the objectives that require the test makers to analyse the type of task in which examiners engage within the course. Since norm-referenced tests are often administered on commercial bases, and the results of the tests are used for making important judgements such as the entrance to schools and companies, the tests' administration should be based on standard procedure. Criterion-referenced tests, by way of contrast, are used to provide teachers and students as means to assess the students' levels of achievement within limited criteria so that they can improve the way they learn or teach. They are often administered within a limited group of students, such as a single class or a course, and therefore, the administrative procedures tend to be less rigorous than those of norm-referenced tests. In terms of scoring, norm-referenced tests use relative evaluation where the means and standard deviation of the group are used to indicate the examinee's relative position in the group, while criterion-referenced tests adopt absolute assessment standards where scores are given based on the levels of achievement in the target criteria.

Table 1.7

Similarities and Differences between Norm-referenced Tests and Criterion-Referenced Tests (Montgomery & Connolly, 1987)

	Norm-referenced tests	Criterion-referenced tests
Purposes	To examine individual performance in relation to a representative group; can be used to establish age levels; used for diagnosis and placement	To examine individual performances in relation to a criterion or external standard; cannot be used to establish age levels unless normed; used for program and evaluation because items are sensitive to the effects of instruction (intervention).
Test construction	Items usually not developed from task analysis; test items may or may not be related to the objectives of instruction (intervention)	Items developed from task analysis; test items are related to the objectives of instruction (intervention)
Administration	Must be administered in a standard manner	May or may not be administered in a standard manner
Scoring	Based on standards relative to a group; variability of scores (i.e., means and standard deviations) is with normal	Based on absolute standards; variability of scores is not obtained because perfect or near-perfect scores are desired

Psychometric properties	Test should demonstrate reliability and validity	Test should demonstrate reliability and validity

An overview of commercially produced, norm-referenced, English-language proficiency tests

The commercially produced , norm-referenced, English-language proficiency tests frequently used in Japanese universities include one of the most influential language tests, the International English-language Testing System (IELTS), which is jointly managed by the British Council, IDP: IELTS Australia, and Cambridge English Language Assessment; the Test of English as a Foreign Language (TOEFL); and the TOEIC and the TOEIC Bridge, which were developed by the world’s leading test development organization, Educational Testing Service (ETS). These tests are used for placement in the appropriate course levels for further education, credit authorisations, and entrance examinations (MEXT, 2014b). In the following section, the paper will outline the formats of the three most influential commercially produced, norm-referenced, English-language proficiency tests, the IELTS, TOEFL, and TOEIC. This is followed by a review of previous studies that focused on these three tests.

Table 1.8

Comparison of the IELTS, the TOEFL, TOEIC, the TOEIC Retrieved from IETLS and ETS Homepage

	IELTS		TOEFL CBT + iBT	TOEIC
Test	A secure, global, authentic and customer-focused test which measures true to life ability to communicate in English. It measures ability to communicate in English across all four language skills—listening, reading, writing and speaking—for people who intend to study or work where English is the language of communication.		A required test to demonstrate the ability to communicate in English in academic settings.	Global standard test for the assessment of ability to communicate in English
Target examinee	The Academic Module Those who wish to enrol in universities and other institutions of higher education where English is used	The General Training Module Those planning to undertake non-academic training or to gain work experience or for immigration purposes	Students in secondary schools, and adult schools, especially prospective students who are planning to enter the US universities and other higher educational institutions	All ages and occupations
Measurement level	The full range of ability from non-user to expert user		The levels of the examinee's readiness for each academic program	From beginners to near-native level speakers
Content of questions	NA		Real-life English-language usage in university lectures, classes, and laboratories. The reading passages are from real textbooks and course materials.	Wide-ranging, from everyday to business topics General English used in the workplace and does not contain

academic language.

	IELTS				TOEFL CBT + iBT						TOEIC		
					CBT			iBT					
Score range	9 Expert user ~		0 No original English Used		0 ~ 300			0 ~120			5~990		
		Number of it/t		M		Number of it/t	M		Number of it/t	M		Number of it/t	M
Test format /time	L	40 it		30	L	30 to 49 it	40 to 60	L	35 to 51 it	60 to 90	L	100 it	45
	R	40 it	40 it	60	R	44 to 55 it	70 to 90	R	30 to 70 it	60 to 100	R	100 it	75
	W	2 t (150 to 250 w)	2 t (150 to 250 w)	60	W S	20 to 25 it	15 to 20	S	6 t	30			
	S	IN	IN	14	W	1 t	30	W	2 t	50			
	T	80 it 2 Wt 1 IN		165	T	94 to129 it 1 Wt	155 to 200	T	70 to 121 it 6 S t 2 W t	190 to 260	T	200 it	120

L=Listening, R=Reading, W=Writing, S=Speaking, WS=Written and Structure, T=Total, M=minute(s), IN=interview, it=item(s), t= tasks(s), l=lecture(s)

As shown in Table 1.8, the basic features of the IELTS, TOEFL, and TOEIC are different in several aspects. The IELTS was one of the first tests that measured the four language skills, i.e., reading, listening, writing, and speaking. Recently, the TOEFL Internet-based test (iBT) was introduced that also includes a speaking section. The IELTS includes listening and speaking sections that all examinees are required to take. It also contains the following two options in reading and writing modules: general training and academic. Examinees who wish to use their test scores to demonstrate their general English proficiency for the global workplace may choose the general training version, while those who wish to use their scores to prove their readiness for academic programmes may choose the academic version. One of the most noticeable characteristics of the IELTS is that it is scored on a scale ranging from one to nine. For example, nine is equivalent to "expert user (who) has a fully operational command of the language: appropriate, accurate and fluent with complex understanding," while one means "No original English used. No assessable information provided. Candidate may have failed to sit for the test" (see http://www.ielts.org for more details).

The TOEFL has undergone a recent revision which transformed it from the TOEFL computer-based test (CBT) to the TOEFL iBT. Details of the change are available on the ETS official Web page. One of the biggest changes in the exam is that it now includes a speaking section comprising six tasks, two of which are independent and four of which are integrated. Examinees' responses are recorded through a microphone and sent to the ETS online scoring network. A TOEFL brochure, *TOEFL iBT Tips—How to prepare for the TOEFL iBT* (ETS, 2007), stated that the new Internet-based test includes "new questions [that] involve integrated [combined] language skills [asking the examinee to] read, listen, and then speak in response to a question" (p. 7) in order to determine the "ability to use English to communicate effectively and determine if they have the language skills needed for academic success" (p. 7).

While the TOEFL test is used primarily for academic purposes, the TOEIC was designed theoretically to evaluate "real-life" English used in everyday conversation and business contexts. ETS claimed that the TOEIC test was originally designed "to meet the need for a measure of English-language skills outside of the traditional academic context" (Woodford, 1982, p. 4). The *TOEIC User Guide* (ETS, 2013b) stated that "the primary purpose of the test was to determine the proficiency level of employees, or potential employees, for human resource planning and development in the contexts of business, industry, and commerce" (p. 2). Therefore, the text content ranges from corporate development and finance budgeting to entertainment and dining out. The reading section is made up of the following three task types: 40 questions of incomplete sentences, 12 questions of text completion, and 48 reading comprehension questions using single and double passages.

Potential problems of using norm-referenced, English-language proficiency tests for placement purposes in Japan

To date, it appears that the use of commercially produced, norm-referenced, English-language proficiency tests for placement purposes is increasing worldwide. Japan is among those countries relying on such tests, due to their perceived utility as a tool for programme coordinators to assign students in appropriate classes according to their English abilities. The major reason why these commercially produced , norm-referenced, English-language proficiency tests are frequently used as placement tests in Japanese institutions is because they are assumed be highly reliable with fast and objective scoring. Moreover, the fact that the tests are easy to administer is also appealing to administrators. Therefore, despite Bachman and Palmer's statement (1996, p. 6) that "there is no such thing as the one 'best' test, even for a specific context," the number of universities that use commercially produced English proficiency tests is steadily increasing in Japan. However, several issues have been raised regarding the use of commercially produced, norm-referenced, English-language proficiency tests in ESL/EFL contexts; the paper will describe them in detail in the following section.

Several concerns have been raised regarding the use of commercially produced , norm-referenced, English-language proficiency tests (Chapman & Newfields, 2008; Hirai, 2010). For example, Chapman and Newfields (op. cit.) question the construct validity of the TOEIC, claiming that the test may not be an appropriate tool to measure "a complex, multifaceted construct such as communication proficiency…through the testing of only receptive language skills" (p. 32). The importance of selecting the right type of test for the placement is, however, rarely perceived in many universities. For example, Yoshida (2009) reports that many universities are using the wrong types of

English test for placement. She argues that universities have seldom reflected upon the appropriateness of the test for their institution in terms of its relevance to a programme or its purpose, except when they need to produce the end-of-term report on their students' progress in English. In her view, universities are adopting these commercially produced tests, because other universities are using them. Researchers such as Culligan and Gorsuch (1999) and Lee, Yoshizawa, and Shimabayashi (2006) have criticised the discrepancy between the purpose of the commercially produced, norm-referenced, English-language proficiency tests and a university program's curriculum. They found that there tends to be a mismatch between the design of the commercially produced, norm-referenced, English-language proficiency tests and the institutions' educational objectives. They also claim that the tests are not flexible enough to fit the target students' proficiency levels. Researchers such as Westrick (2005) and Nakamura (2007) propose the use of their own placement tests. Among the proponents of in-house placement tests, Nakamura (2007) claims that "an institution's placement test should be closely linked with its curriculum" (p. 97). He suggests that "the content, level and purpose of those tests" (p.98) should be related closely to the institution's curriculum. Moreover, the tests may affect various aspects of teaching and learning in schools and the curriculum.

The criticisms raised by Nakamura (2007) and Yoshida (2009) are closely related to the issue of the type of test appropriate for individual language programmes. Brown (1996) proposes four types of tests for program-testing contexts: proficiency tests, achievement tests, placement tests, and diagnostic tests. This gives a clear distinction between norm-referenced tests and criterion-referenced tests (see Table 1.7). Table 1.9 provides insightful information about what type of test is most appropriate for different purposes and contexts of test use.

Table 1.9

Matching Tests to Decision Purposes (Brown, 1996, p.7)

Test qualities	Type of decision			
	Norm-referenced		Criterion-referenced	
	Proficiency	Placement	Achievement	Diagnostic
Detail of information	Very general	General	Specific	Very specific
Focus	Usually general skills prerequisite to entry	Learning points from all levels & skills of program	Terminal objectives of course or program	Terminal and enabling objectives of courses
Purpose of decision	To compare an individual's overall ability with other individuals	To find each student's appropriate level	To determine the degree of learning for advancement or graduation	To inform students and teachers of objectives needing more work
Relationship to program	Comparisons with other institutions or programs	Comparisons within program	Directly related to objectives	Directly related to objectives still needing work
When administered	Before entry and sometimes at exit	Beginning of program	End of courses	Beginning and/or middle of courses
Interpretation of scores	Spread of wide range of scores	Spread of narrower, program-specific range of scores	Overall number and percentage of objectives learned	Percentage of each objective in terms of strengths and weaknesses

According to Hughes (2003), proficiency tests are designed to measure global language skills or abilities. Each student’s score is relative to the scores of all the other students who took the test. Hughes claims that the greatest difference between proficiency and normal class tests is that the students may encounter test items they have never been taught, since proficiency tests are not based on what was taught in the classroom. On the other hand, achievement tests are directed toward classroom lessons and curriculum. Test content is limited to specific material and items addressed in a curriculum. Therefore, if the test includes tasks that have not been taught in the classroom, such tasks should be deleted from the test. Placement tests are used to “place students into a particular level or section of language curriculum or school” (Brown, 2004, p. 45). Based on the results of placement tests, administrators decide which class level is most appropriate for students. Brown (2004) argues that a diagnostic test “is designed to diagnose specified aspects of a language...that are difficult for learners and should therefore become part of a curriculum” (p. 46).

As Brown (2004) admit, the tests described above have indistinguishable aspects. For example, Brown (2004) argues that placement tests tend to have diagnostic aspects, because “any placement test that offers information beyond simply designating a course level may also serve diagnostic purposes” (p. 47). This, however, does not apply to many of the commercially produced, norm-referenced, English-language proficiency tests that are used for placement in tertiary-level institution in Japan. For example, the TOEIC result is comprised of five discrete parts to determine reading abilities (referred to as “abilities measured” in the TOEIC): inferring ability, ability to locate and match information in texts, ability to connect information across single/multiple sentences(s) in a single/multiple passage text(s), knowledge of vocabulary, and knowledge of grammar (Schedl, 2010). Therefore, teachers are provided only with an overall score for the knowledge of grammar but do not receive more precise information regarding the students’ level of achievement within certain criteria (Culligan & Gorsuch, 1999; Wistner & Sakai, 2007). For example, Minai (2000) argues that non-finite verbs, especially past participles, are difficult for Japanese students because many students are still unaware of the fact that an English sentence cannot have two finite verbs unless the sentence has at least one coordinate conjunction in between the two verbs, and participles that modify nouns are often misinterpreted as finite verbs. If there were a grammar test comprised of several subordinate tests, the results would be useful in helping teachers to improve the way they teach in the classroom by giving more information on less achieved criterion (Minai, 2000, 2003). Moreover, there are several orthographic and phonological points Japanese students find difficult (Takebayashi & Saito, 2008; Narita, 2007, 2009). For example, Narita (2009) argues that the letter /s/ as the indicator of the plural form of some nouns can be read in two different ways: /s/ and /z/ confuse novice readers in Japanese junior high schools. Many students spend more time identifying which words are read /s/ and which are read /z/. Not only novice readers but also relatively fluent readers tend to demonstrate the inclusion of vowels while reading, such as /tek**u**nology/ for technology and /d**o**rama/ for drama (Ichiyama, 2014). If there were an English proficiency test that had orthographic and phonologic processing skills sections, the results would be useful in helping teachers improve the way they teach in the classroom by giving them more information on criterion with less achievement (Ishikawa, 2008; Narita, 2009).

Another disadvantage of using norm-referenced tests in the classroom would be their negative effects on teaching and learning. The problem is that some teachers believe that these tests are used to measure the effectiveness of the teaching in courses. While teachers say that they do not want to teach for the test, their practices might be influenced by the national curriculum or the school syllabus that requires end-of-term examinations to measure the students' and teachers' performance. If the content and format of the tests are given in advance—which happens quite often, since the past test papers are attainable from most bookstores—teachers may unconsciously or not prepare students to become used to such forms and contents to expect test-based accountability. This may result in dismissing the content of curricula that have been judged as unrelated to the content of norm-referenced tests. Even students expect teachers to teach test-specific English. A survey done by Tokunaga (2007) on the use of the TOEIC test in universities revealed that the majority (97%) of respondents who had taken TOEIC tests at their universities answered that they agree or strongly agree with the question of "whether they need to study TOEIC-specific books to raise the TOEIC scores." This result sharply contrasts with the result of 47% of respondents who answered that they agree or strongly agree with the question of "whether they need to study general English material to raise the TOEIC scores."

1.2.5 The development of an in-house placement test and item bank

In this section, the paper will firstly describe the pros and cons of the use of an in-house English-language placement test, referring to the degree of implementation of English-language placement tests in Japanese universities. Then, the paper will describe the benefits of developing an item bank to facilitate the introduction of in-house English-language placement tests. Finally, the paper will review the prior studies on the development of English-language tests.

In-house placement tests

The most appealing advantage of in-house placement tests is that the test content matches the level of the students. This is described in Hugh's arguments (2003) on kinds of tests. According to him, the placement tests that are most successful are those constructed particularly for a situation. The test is designed as a result of the identification of the key features in different levels of teaching in the institution. Therefore, they are tailor-made rather than "bought off the peg." This means that they have been produced in-house (Hughes, op. cit., p. 17). The benefits of in-house placement test are felt in Japan and the number of universities that use tests developed in their own institutions is increasing (Obermeier, 2009; Otani, Yokoyama, & Bradford-Watts, 2014).

However, several disadvantages also exist for the use of in-house placement tests. One of the biggest obstacles is that it takes time and labour for administrators and teachers to produce a reliable and valid test (Nakamura, 2007). There is a need to first identify students' needs in the target course and then devise test tasks that reflect the content of the course and measure the knowledge and skills the course requires of students (Rian, 2010). The process of developing an in-house placement test will be described in detail later in this chapter.

Item bank

The accurate measurement of a student's ability to read in English is one of the most important aspects of a test for teaching English as a foreign language (EFL) (Brown, 1996, 2004; Bachman & Palmer, 1996). Without accurate assessments and an understanding of students' abilities, teachers are unable to provide students with effective and efficient instruction tailored to their needs (Nakamura & Ohtomo, 2002; Nakamura, 2007; Yoshida, 2009). Moreover, tests should not be limited to high-stakes tests, such as entrance examinations to a university or class-placement tests at the beginning of a semester. As Brown (1996) points out, the benefits of formal and informal classroom assessments are considerable. Teachers of English and their students can benefit from weaknesses multiple opportunities to identify students' weaknesses that may have gone unnoticed before, such as confusion about the use of the articles "a" and "the" in a writing class. The teacher would not know whether the confusion exists only for that particular student or whether many students have the same confusion, and thus, special attention should be given to clarifying the concept. The earlier the teacher assesses his or her students' abilities, the more effective the instruction can be. In order to accurately evaluate students' abilities, the teacher needs to be equipped with effective media and tools for conducting appropriate assessments. While several assessment tools are available, storing their own test items in an "item bank" is one useful option for teachers (Hirai, 2010).

A "test bank," sometimes referred to as an "item pool," is a "bank" of test questions that have been developed and assessed at various levels and categories before the administration of the actual test (Beeston, 2000). These pre-selected test items are often assessed thoroughly using statistical analysis, such as Rasch analysis, a test theory that measures the item's difficulty based on the examinees' responses to the item, to ensure that the items function properly (Aline & Churchill, 2005; Abe, Wistner, & Sakai, 2008). Rasch scaling also enables the placement of different test items on the same scale—a process described as "equating, a direct linking procedure, [that] maps measures from two or more test forms onto each other for the purpose of allowing the measures from each test to be used interchangeably" (Skaggs & Wolfe, 2010).

The use of an instructor's own item bank appears to be beneficial for English teachers for several reasons. Firstly, not all the commercially produced assessment tools meet their students' levels of English proficiency (Culligan & Gorsuch, 1999). Storing test items with various formats and levels would benefit many teachers who can develop their own tests, retrieving test items from the item bank as needed. Secondly, because testing companies do not disclose test items or the results of individual test items, teachers do not know their students’ responses to the particular test items. Using a teacher-developed test allows instructors to obtain the results of individual test items so that they can fine-tune their teaching to address students7 specific confusions or misunderstandings. Moreover, for educational institutions with time and financial constraints, the use of in-house tests would be beneficial because of the relatively lower cost of administration and time constraints (Hirai, 2010; Koyama, 2013).

Test development

Useful suggestions have been proposed by several test developers. One of the earlier attempts was the framework provided by Carroll (1985), the prominent researcher in the field of English-language testing. He divided the test construction phases into four stages: design, development, operation, monitoring (D-D-O-M).

Phase 1 : Design	Description of testee(s)
	Specifications of settings, needs
	Statement of test tasks, topics
Phase 2: Development	Construction of draft test
	Trials of test
	Analysis of trial and test revision
Phase 3: Operation	Introduction of test for practical use
	Making decisions on test information
Phase 4: Monitoring	Survey of test administration
	Establishment of test measurement of characteristics
	Preparation of test revision schedule

Figure 1.6. Test construction phases (Carroll, 1985).

More detailed guidelines were provided by Hughes (2003, p. 58), who described the procedures of language test development predominantly for ESL/EFL test developers:

- Make a full and clear statement of the testing ‘problem’.
- Write complete specifications for the test.
- Write and moderate items.
- Trial the items informally on native speakers and reject or modify problematic ones as necessary.
- Trial the test on a group of non-native speakers similar to those for

whom the test is intended.

- Analyse the results of the trial and make any necessary changes.
- Calibrate scales.
- Validate.
- Write handbooks for test takers, test users and staff.
- Train any necessary staff (interviewers, raters, etc.)

Hughes' framework provided practical administration of the test design with emphasis on writing detailed specifications of a test. Specification, a "generative blueprint or design documents for a test," (Fulcher & Davidson, 2007) in Hugh's view, includes information regarding content, test structure, timing, medium/channel, techniques to be used, critical levels of performance, and scoring procedures which are broadly categorized into four groups, each with subcategories, which may differ depending on the type of tests. Alderson (2000) argues that test specification can be a useful source of information regarding the construct of test measures, an issue fully developed in the following section.

Figure 1.7 shows a specification of a reading suggested by Hughes.

i) *Content*	
Operations	The tasks which the test takers are required to engage, including skimming, scanning, guessing etc.
Type of texts	May vary depending on the test type. Examples include English-language educations, handouts, articles in newspapers, journals and magazines, letters, poems, leaflets, advertisement, novels, times, etc.
Text forms	Description, exposition, argumentation, instruction, narration.
Length of text(s)	The number of word
Topics	May vary depending on test types
Readability	The Flesch Reading Ease score or the Flesch-Kincaid Grade Level Score available for text in Microsoft Word
Structural range	List of structures which may/may not occur in texts
Vocabulary range	Either a list of words with frequencies and levels or just a few wordings, such as non-technical
Grammar range	Either a list of structure or a reference in a course book
ii) *Structure, timing, medium/channel and techniques*	
Test structure	The number of sections
Number of items	In total and in the various sections
Number of passages	In total and in the various sections
Medium/channel	Paper/pencil, face-to-face etc.
Timing	Each section and for entire test
Techniques	What techniques will be used to measure what skills or sub-skills?
iii) *Criterial levels of performance*	
Statement on what is scored a “success” or a “failure”	
iv) Scoring Procedure	
How or who will score a piece of work, etc.	

Figure 1.7. Specification for a test (retrieved from Hughes, 2003, p. 59-61, 140 and adapted for a reading test).

Bachman and Palmer (1996) elaborated on test development, especially the theoretical aspect of test specification. The framework, which was originally provided by Bachman (1990) and later elaborated on by Bachman and Palmer *(*op. cit.) gives useful checklists for the development of the test tasks that reflect the notion of target language use domain (TLU domain) analysis. Bachman and Palmer’s test development framework is comprised of three stages: design, operationalisation, and administration.

One of the aspects that makes Bachman and Palmer’s model different from the earlier models is that it includes the notion of TLU domain and target language use tasks (TLU tasks). The TLU domain is an idea closely related to the issue of authenticity. Bachman and Palmer (1996) state that "(authenticity) is the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (p. 23). According to their description, the term TLU domain refers to the domain where students use the target language to communicate in real life; TLU tasks are described as "A set of specific language use tasks that the test takers is likely to encounter outside of the test itself, and to which we want to make inferences about the language ability to generalize" (p. 44). In order to design test tasks that correspond with the TLU domain, therefore, we need to specify the "critical features" of TLU domain, the process which related to the issue of authenticity of test tasks. The more the TLU tasks resemble the use of language in TLU domain, the more the interpretation of such test scores becomes meaningful and effective and one can claim that the score represents how well the test takers might succeed in real-life situations. Another important aspect of the model is the theoretical definition of the construct to be measured that is also related to the TLU analysis phase. Bachman and Palmer (1996 p. 116) argue that the construct to be tested is defined through the TLU analysis phase, and based on this analysis, a test and specifications will be designed. They give three reasons why the definition of the constructs are needed: "1. to provide a basis for using test scores for their intended purpose(s), 2. to guide test development efforts, and 3. to enable the test developers and users to demonstrate the construct validity of these interpretations."

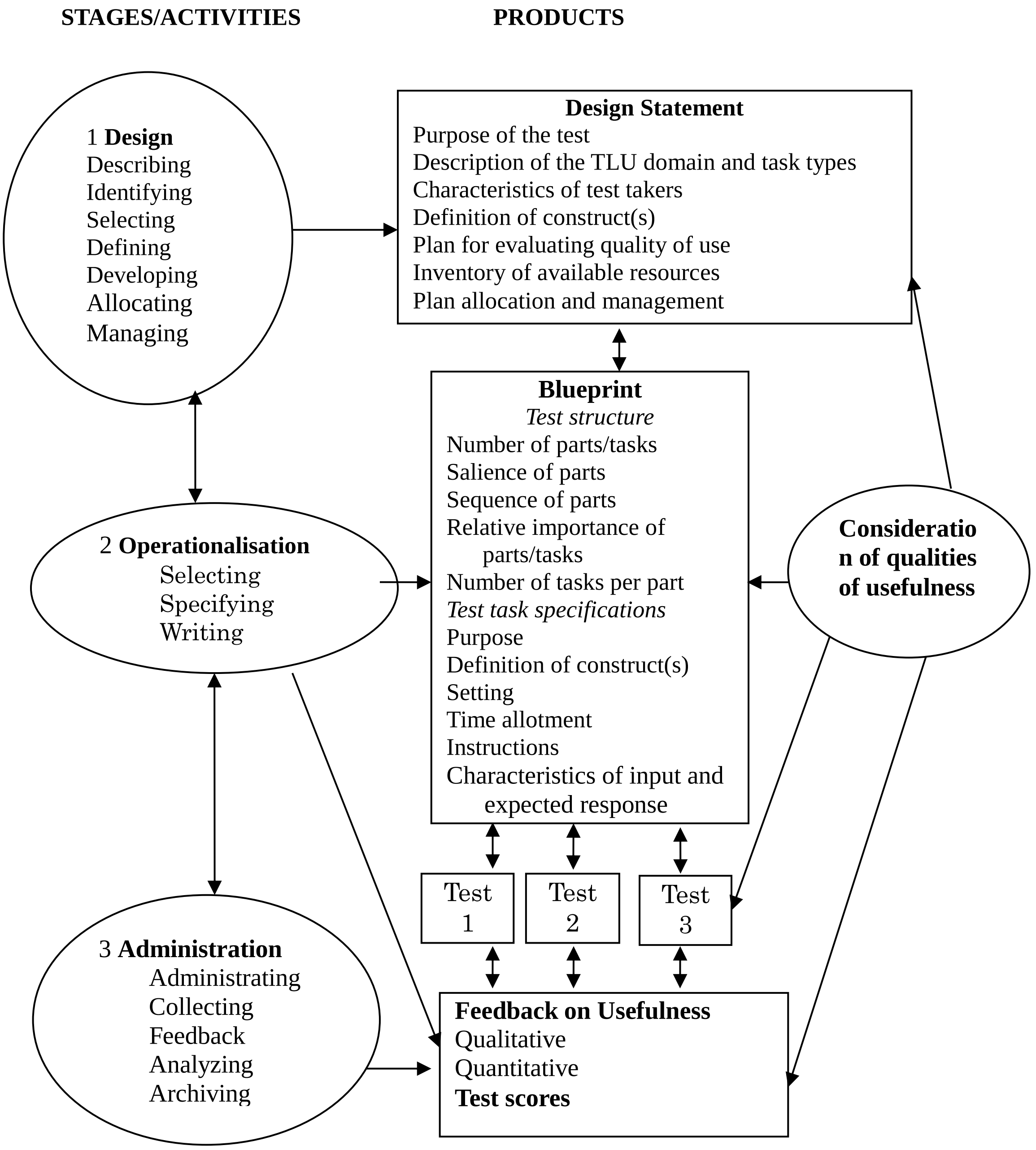


Figure 1.8. Test development (Bachman & Palmer, 1996, p.87).

1.3 Statement of problems of an English-language programme in a Japanese tertiary-level institution

The writer works for the university that was originally founded as the Imperial Women's Medical College in 1925 and later changed to Toho University in 1950. The university is comprised of four faculties: Medicine, Pharmacy, Science, and Nursing. Each faculty aims to graduate students with a national qualification, such as, medical licence, nursing licence, pharmacist licence, and medical technologist licence. Therefore, most of the university's graduates work primarily in medical fields in Japan and around the globe. In the Tokyo district alone, the faculty of Medicine has four medical centres, which can cater approximately 2000 inpatients a day.

The Faculty of Nursing is encompasses thirteen departments, ranging from the Department of Humanities to the Department of International Heath and Nursing. There are approximately 450 undergraduate and postgraduate students. During their final year of collegiate life, students take the national examination to become a licensed nurse. Some students also take the test of public health nurse or a midwife. The faculty encourages students to work outside of Japan and therefore contribute to the support and development of nursing care in Asian countries. Consequently, the students of the Department of International Health and Nursing are required to participate in practicums in Thailand and Laos. Moreover, students must take at least one foreign language subject besides English. This requirement is described as innovative since many of the nursing faculties struggle to plan a syllabus within the limited framework. The Japanese government sets a rigorous standard for granting nursing diplomas.

Foreign language education, including English-language education, at the Faculty of Nursing is administered by the Foreign Language Department, which consists of one full-time lecturer and fifteen part-time lecturers. The writer is the full-time lecturer and is responsible for the administration of all the foreign language subjects. Most of the part-time lecturers are native speakers of a foreign language and have lengthy experience teaching English in Japan. Many have taught English at the faculty for ten years.

The English education at the Faculty of Nursing is comprised of eight subjects and five selective subjects. Each student is required to take at least two English compulsory subjects every year. Each compulsory subject requires students to participate in a one-and-a-half hour class a week that lasts for 30 weeks. Therefore, first-year students, for example, take one English class on Tuesday morning and another class on Wednesday morning throughout the year. A native speaker of English teaches at least one of the classes. Therefore, some students participate in two English classes, which are both taught by a native speaker of English.

Every year, the Department of Foreign Language administers an English placement test to all the incoming students. After the scoring, the students are allocated to eight different classes, A1, A2, B1 B2, C1, C2, D1, and D2. Only the A1 and A2 classes consists of the highest achievers of the placement test and C1 to D2 classes are mixed-ability classes. Usually, the English placement test is comprised of two main sections, listening comprehension and reading comprehension. In the listening comprehension section, students listen to a short conversation / watch DVD and later answer several listening comprehension questions. In the reading comprehension section, students read approximately two to three 150-300 words passages and then answer the comprehension questions.

The primary reason why the tertiary-level institution administers an English placement test is that the faculty introduces different types of an entrance examination. To be more precise, there are three methods of admission into the faculty. Although various types of entrance examinations used in Japan have already been described substantially in the preceding section, there are three methods of admission into the faculty: requires the examinees to take two subjects including English as the compulsory part of their entrance exam, requires the examinees to take tests of mathematics, biology, chemistry, sociology, a current question and Japanese, and requires the examinees to provide high school records and an interview. Because of the extensive competition among the nursing faculty, the admission office decided to encourage more students to enter without taking conventional paper-and-pencil writing tests. As a result, the number of students entering the university without taking an English exam has drastically increased over the past few decades. Although the exact number of students is not disclosed by the institution’s admission office because of confidentiality, the prospectus provided by the university shows that in 2014 approximately 40 percent of students entered the university without having taken English tests. Therefore, it is necessary to develop an in-house test items that would classify those students appropriately so they are able to benefit the most from English-language programmes early in their collegiate careers.

There seems to have been, however, several problems with the faculty's placement testing system:

- Until I started working for the faculty, the Foreign Language Department had developed a new set of placement test each year. This process required a great deal of time and energy. As an only full-time English lecturer, I am required to develop seven sets of English entrance examination for undergraduate and postgraduate programs of the faculty, each of which is administered between October and March respectively. The types of an entrance examination in Japan will be fully described in the next chapter. Moreover, as a member of the faculty's entrance examination board, I also have to grade and make final decisions on which applicant should be allowed to enter the university. Shortly after the final entrance examination in the middle of February, the department was forced to finalize decisions on several administrative issues concerning the new academic year, set to start only a month and half later, on April 1st.
- The analyses of placement test results are few and far between, especially regarding the appropriateness of test items relative to the incoming students' proficiency levels. Such neglect may stem from the fact that after placement tests have been administered, faculty schedules quickly become full. Moreover, the department must not only distribute students into classes, but also explain to students which textbooks to buy and where to go on the first day of the school. Since these administrative tasks have been done by a full-time lecturer, there is no time for any analysis of results once the semester starts.
- One of the most significant problems is that, since the test items are always changing, there is no way to compare students by entrance year which could, however, beneficial for several reasons. Fujita (2004) asserts that the neglect of "analysing the test results means wasting data and losing the opportunity to gradually create more reliable, valid, and effective tests."
- These tests are quite informative regarding students' relative English proficiency

levels. However, neither the overall score nor sub-score of any section provides to teachers which areas of English-language education should be given more attention and those that should not. As Bloom, Hastings, and Madaus (1971) have pointed out, testing used to make initial evaluations should "diagnose strengths and weaknesses" of students, so that teachers can pay increased attention to teaching topics that have proven difficult for many underachieving students.

Chapter 2: Literature Review

This chapter presents a literature review to justify this study's use of particular research models, frameworks, and theories, including the Rasch model, several reading models for testing reading comprehension and some lower-level reading components, an orthographical processing skill and a phonological processing skill, and a framework for developing a valid in-house placement test. Before discussing the theories, this section will briefly summarize some fundamental considerations of the characteristics of measurement, after which the following section will focus on considerations of language ability and reading processes. The third section will describe the design and development of an in-house placement test tailored to students in the program; these topics involve some of the most important research issues of test development, including the reliability and validity and test development procedures. Lastly, an argument for the using the Rasch model in placement testing will be articulated.

2.1 Fundamental consideration of the characteristics of measurement

In developing tools that successfully measure language ability, test developers need to be aware of several concerns, mainly regarding the limitations of measurements and the interpretation of test scores, which are closely related to the characteristics of measurements themselves (Bachman, 1990). Bachman, in the preeminent *Fundamental Considerations in Language Testing*, argues that measurements are limited by three principal characteristics: the limitations of specification, observation, and quantification.

2.1.1 The limitations of specification

The limitations of specification are closely related to the theorization of language ability, described in the following section. Test administrators are interested in assessing not only direct observational abilities, but also language abilities that are based on abstract psychological attributes or abilities, sometimes referred to as "traits" or "constructs" (Wilson, 2005). Because researchers cannot actually see what is happening inside the human brain, they can only infer from the observed behaviours of examinees. We assume that surface behaviour is an indication of the ability to perform an intended activity taking place inside the brain. If, for example, a test developer designs a test that requires examinees to answer 20 multiple-choice items that test the examinees' linguistic knowledge of English their reading ability in English as a foreign language, the developer has to clearly theorize what constitutes "reading ability," as well as demonstrate that the linguistic knowledge tested is, in fact, a component of reading ability (Jung, 2010).

Moreover, language testers need to provide valid evidence that test scores are not affected by other factors, such as vocabulary ability, in order to avoid violating unidimensionality (Shizuka, 2007), which Bond and Fox (2007) describe as the most fundamental aspect of measurement. They define unidimensionality as "a focus on one attribute or dimension at a time" in order to make "meaningful estimates of the object" (p. 32). Therefore, an item that is claimed to assess listening comprehension ability cannot also be claimed to assess reading comprehension ability. Under "single attribute at one time" restrictions, language test developers are required not to combine a number of attributes to obtain single scores. This is why some of the listening sections of commercially produced English proficiency tests are criticised for being multidimensional, (Kenneth, 2000; Kluitmann, 2008). For example, multiple-choice questions in the listening sections of tests ask test takers to choose one right "written" answer. Test takers' listening ability scores are influenced by their ability to "read" written information. There are moves to compensate for such deficits. For example, Obermeir (2009) reports on the development of a listening test by Kyoto University of Education's general English requirement course. He claims that their test items are "purely a listening item in the sense that there is no reading involved" (p.108) because students are required to listen to both the question and the choices of response.

The definition of language ability, however, is a long-standing problem not only for language test developers, but for researchers in education. The literature on what constitutes language ability suggests that there are two strands of thought in the field of language testing: a single general factor approach and a componential approach (Sang, 2005; Shiotsu & Weir, 2007). One of the most influential proponents of the single general factor approach is Oller (1979), who has proposed the unitary trait hypothesis. His argument, however, has been strongly criticised by researchers who critique his flawed methodologies (Vollmer & Sang, 1983). The multi-component approach takes the position that language proficiency consists of one higher-order factor and several first-order factors (Carroll, 1993; Sasaki, 1996). To define language ability, as Bachman and Palmer (1996) observe, there is a need to know exactly what language ability is and what its components are. Moreover, there is a need to specify factors that are not part of language ability but might affect test performance, such as individual characteristics including cognitive, affective, and physical characteristics (Bachman, 1990; Jung, 2010).

2.1.2 The limitations of observation and quantifications

The limitations of observation are also related to the specification of an abstract notion of ability. Regardless of how much objective observation test developers intend to achieve, many assessment procedures include subjective judgments (Bachman & Palmer, 1996). This might be the case, for example, when an interviewer judges examinees' ability to successfully complete a hotel manager's responsibilities in an environment where English is the communication medium. Even if the criteria needed to make an objective assessment have been clearly stated, the interviewer's subjective views may ultimately decide whether examinees pass or fail. Moreover, because what testers see in testing contexts is limited in scope, there is no guarantee that it represents testees' real abilities. Therefore, as Montgomery and Connolly (1987) have noted, scoring should be "based on absolute standards" (p. 1874).

Perhaps more important to measurement than quantification and unidimensionality, however, is whether an attribute is quantitative (Bachman, 1990). Cowles' (1989) general understanding of measurement in the social sciences is instructive:

> Measurement is the application of mathematics to events. We use numbers to designate objects and events and the relationships that are obtained between them. On occasion, the objects are quite real, and the relationships are immediately comprehensible…At other times, we may be dealing with intangibles, such as intelligence… In these cases our measurements are descriptions of behaviour that, we assume, reflect the underlying construct. But the critical concern is the hope that measurement will provide us with precise and economical descriptions of events in a manner that is readily

> communicated to others. Whatever one's view of mathematics with regard to its complexities and difficulty, it is generally regarded as a discipline that is clear, orderly, and rational. The scientist attempts to add clarity, under rationality to the world about us by measurement (p. 35).

Moreover, the assignment of numbers depends critically on the type of scale used. For example, the classification of people by race requires the use of a nominal scale, based on names or labels, with test scores based on numbers. Right answers demand an interval scale, based on regular interval divisions and without an absolute zero.

In response to such arguments, however, researchers like Mitchell (1999) have proposed more vigorous definitions of measurement. Mitchell (op. cit.) alleges that measurement is "the attempt to discover real numerical relations (ratios) between things (magnitudes of attributes), and not the attempt to construct convenient numerical relations where they do not otherwise exist" (p. 17). He continues,

> …only attributes which possess quantitative structure are measurable. This is because only quantitative structures sustain ratios. Unless every attribute really is quantitative, to conclude that, because one can make numerical assignments to things, the attribute involved must be measureable is to presume upon nature (Mitchell, 1999, p. 19).

Mitchell's argument about the quantifiability of an attribute leads to an important point that will be discussed in the next section of this paper. Particular statistical methods, such as Rasch analysis, which transforms unquantifiable attributes into quantitative structures, will be discussed (Shizuka, 2007).

To summarize the above section, as Bachman (1990) claims, the procedures to connect language knowledge or construct the test developers that try to assess and the observed language performance should be logical. According to his claim, there are three stages to the procedure of relating the two aspects, theoretical constructs and performance: theoretical definition of constructs to be measured, operational definition of constructs, and quantification of the observation. Therefore, the following section will review the literature on language ability, the reading process, test development and assessing of the usefulness of a language test and the test theories. The paper will describe how the literature on language testers as well as second language acquisition (SLA) researchers have strived to relate and theorize grammatical knowledge within the model of language ability. The paper will also describe some important aspects of the Rasch model.

2.2 The nature of reading process

What happens in a reader's mind while reading—that is, the cognitive aspect of reading— is one of the most important issues that language testing explores. As Davies (1995) notes, research into the reading process consists of "a systematic set of guesses or prediction about a hidden process." According to Koike, Kinoshita, Terauchi, and Narita (2004), several cognitive reading processes take place which are interrelated with each other. In the same way as language ability, the recent literature on reading process does not consist of one single factor process but a multivariate factor process including a complex combination and integration of various cognitive, linguistic, and meta-cognitive skills (Nassaji, 2003). As Hudson (1998) points out, because the reading process is the integration of numerous reading processes, there are several views on what constitutes the reading process. For example, to the question of what kind of processing is included in the reading process, Grabe (2000, p. 230) proposes,

> …orthographic processing, phonological coding, word recognition (lexical access), working memory activation, sentence parsing, propositional integration, propositional text-model formation, comprehension strategy use, inference-making, text-model development, and the development of an appropriate situation model (or mental model).

It is Urquhart and Weir (1998) who categorise the reading models into two types: process models and component models. According to their view, a process model explains the dynamic process of reading so as to explain how each process interrelates with the others, while a component model describes the combinations of constituents of the reading in order to reach a comprehensive view of reading. Both have proposed several models, claiming the relative importance over the other. Therefore, this paper will first outline the models that have been proposed in the recent literature.

2.2.1 Process models of reading

Broadly speaking, three types of process models of reading can be identified in the literature (Grabe, 2009). These are a bottom-up, a top-down, an interactive model which attempts to synthesise the characteristics of the bottom-up and top-down model and neo-bottom-up processing models. Although the models have been criticised for lacking evidences to support their views, they provide several basic ideas including word recognition, automaticity and fluency (Stanovich, 1990, 1991; Grabe ,1991).

Bottom-up processing model

The bottom-up model was developed by Gough (1972). He suggests that comprehension takes place in the linear order of processing from individual letters, words, phrases and sentences, discourse, and finally to comprehending the message conveyed by the writer through the text. Cairnely (1990) follows this view by suggesting that readers extract meaning from print by processing the text in a linear way, permitting them to transfer meaning from the page to their minds. Here, reading is assumed to be primarily a decoding process of reconstructing the printed letters and words and building up meaning from a text from the smallest units. This type of information processing is often called text-based or data-driven processing because processing is inspired by linguistic input from the text of the incoming data (Silberstein, 1994). One aspect that the bottom-up processing approach attempts to account for is the development of automaticity in word recognition (see LaBerge & Samuels, 1974; Eskey & Grabe, 1988). Leading research indicates the importance of accurate and rapid word/phrase recognition to the development of fluency in reading. This has been observed especially with “good readers who can be distinguished from those who read less well by means of nothing more than their skills in recognising individual words in context-free settings both more rapidly and more accurately” (Eskey & Grabe, 1988, p. 232). The same view is held by Kim and Goetz (1994) who claim that the ability to do this will distinguish a good reader from a weak one. Conversely, research indicates that “comprehension deficits can at least in part be traced to deficiencies within the word recognition process” (Chabot, Zehr, Prinzo, & Petors, 1984, p. 148). Major criticism of the bottom-up model derives from its linearity of process. Rayner and Pollatsek (1989) argue that the process does not explain some of the important features of reading, such as weaker readers re-reading the same words, or the mechanism of inferences.

Top-down processing model

In contrast with Gough’s bottom-up model, scholars such as Goodman (1967) and Smith (1971) focus on the active and cognitive aspects of reading processes that use psycholinguistic models of reading as underlying theory. Researchers such as Carrell (1988), Nuttall (1996), and Carinely (1990) argue that reading needs a higher level of cognitive ability than that of the relatively simple bottom-up decoding. They emphasise that the knowledge and experience that a reader brings to the reading process play a significant role in predicting and inferring the meaning of a text. The top-down approach may also be termed “content/knowledge processing” (Cornish, 1992, p. 725), where the emphasis is on the reader’s use of pre-existing knowledge and information. This idea of background knowledge in reading comprehension has been formalised as “schema theory” (Rumelhart, 1986). According to schema theory, readers retrieve or construct meaning based on their own previously acquired knowledge. This previously acquired knowledge is called the reader’s background knowledge and previously acquired structures are called schemata (Rumelhart, 1980). Davies (1995, p. 66) claims that without the reader’s prior knowledge and experience, it is difficult to interpret visual information and words. Wray and Lewis (1997, p. 31) also suggest that, “Learning which does not make connections with our previous knowledge is learning at the level of rote only and is soon forgotten.” However, the reading research on L1 has shown that readers actually understand most of the words while reading (Kadota, 2001). Just and Carpenter (1987), for example, report that readers observed more than 80% of content words and 40% of functional words in their eye fixation study. As Stanovich (1991) points out, good readers read effectively without spending cognitive resources on the perceptional process while reading. Weak readers, in contrast, depend more on contextual information than the good readers (Biemiller, 1970).

Interactive processing model

More recently an interactive model of reading has been proposed as a way of accommodating aspects of bottom-up and top-down models of reading, which are both seen as important in processing and interpreting text. The approach was developed from the view that both approaches described have deficiencies in their effectiveness and efficacy. The two approaches run into difficulty because they assume that processing must proceed exclusively from top to bottom or from bottom to top, or that reading is only a decoding or only a cognitive process (LaBerge & Samuels, 1974). Rumelhart (1977) first proposed the approach to reading that incorporates features of both bottom-up and top-down models. Rumelhart (op. cit.) argues that the process of reading, "begins with a flutter of patterns on the retina and ends when successful with a definite idea about the author's intended message" (p. 573-574). For him, reading is "at once a perceptual and cognitive process," and that the various sources of information appear to interact in many complex ways during the process of reading.

2.2.2 Component model

As has been described in the earlier section, the component approach aims to model the combination of reading constituents in order to explain the differences between good and poor readers. One of the most prominent models is the simple view of reading.

Simple view of reading model

According to Hoover and Gough (1990), the simple view of reading model refers to a simple combination of decoding (D) and comprehension (C) processes. Therefore, reading comprehension (R) can be predicted either by multiplication of D and C or by addition of D and C. He argues that the score of decoding correlates well with language comprehension. Kadota (2012) claims that the model assumes the success of decoding as the basis of further reading comprehension.

Reading = Decoding × Comprehension
Reading= Decoding + Comprehension

Figure 2.1. Simple view of reading model (Gough & Wren, 1999; Kadota, 2012).

Grabe and Stoller's view of reading process

Researchers, such as, Grabe and Stoller (2002) and Grabe (2009) offer a comprehensive set of reading constituents which explores the ideas of the simple view of reading model. They also categorise the constituents into two levels: lower-level and higher-level.

Lower-level

- word recognition
- syntactic parsing
- semantic proposition formation
- working memory activation

Higher-level

- text model of comprehension
- situation model of interpretation
- background knowledge use and inferring
- a set of reading skills and resources under the command of the executive control mechanism in working memory

Figure 2.2. Reading processes that are activated when we read (Adapted from Grabe & Stoller, 2002; Grabe, 2009).

Samuels (2006) developed the idea of automaticity by comparing the reading processes of novice and fluent readers. Grabe (2009) states that "automaticity is seen as a critical way for readers to engage in multiple processes more or less simultaneously (or in parallel)" (p. 28). The idea of automatization derived from automatic theory proposed by LaBerge and Samuels (1974). According to their theory, fluent readers can spare more of their cognitive resource of working memory, in this case, attention, to comprehension rather than decoding because the decoding process has been automatized.

The dotted lines in the following figures (*Figure 2.3. and 2.4.*) indicate the process being automatized while the solid lines indicate that the process is consciously operated by paying attention, which leads to the consumption of cognitive resources. One of the greatest differences between fluent and novice readers is that many of the reading processes, especially that of decoding, are automatized so that less attention has been paid in fluent reading. This enables the fluent readers to spare more cognitive resources to comprehending the meaning of the text, which leads to fast reading.

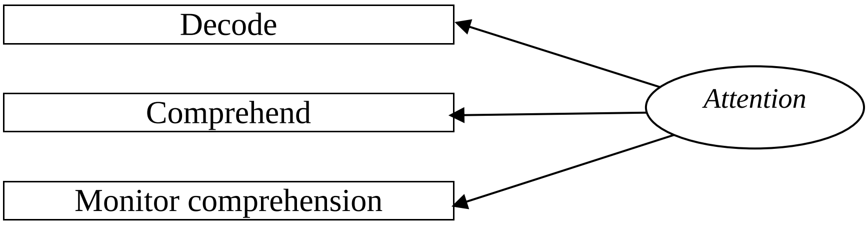


Figure 2.3. Novice reader's text comprehension (Samuels, 2006, p.37).

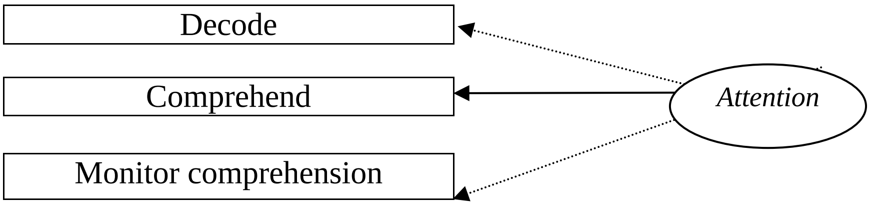


Figure 2.4. Fluent reading in L1 (Samuels, 2006, p. 38).

Based on Samuel's fluent reading process, Kadota, Noro, and Shiki (2010) developed the process of fluent reading in L2 (*Figure 2.5*). Similar to the fluent readers in L1, L2 readers have automatized the lower-level reading processes by paying less attention, which enables them to spend more attention on higher-level processes.

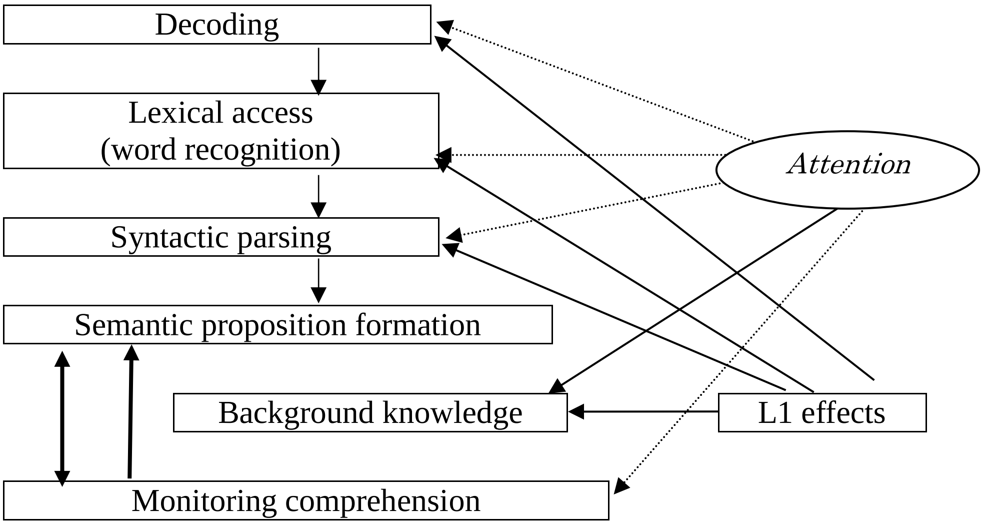


Figure 2.5. Fluent reading in L2 (Kadota, Noro, & Shiki, 2010, p. 337).

Working memory is also an important constituent of fluent reading and the automatization of lower-level processes: in this case, word recognition and syntactic processing. Working memory is a memory system that enables readers to engage in a complicated intellectual activity, such as reading. According to Funahashi (1996), working memory is a mechanism that actively retains information needed to accomplish an activity or task for a required period of time. Grabe (2009) states that the role of working memory in lower-level memory is “direct and well established.”

Working memory is memory that processes information by temporarily retaining information in order to integrate it with the next information to be input. The term working memory develops the idea of short-time memory that is believed to be a static memory system and only retains memory temporarily. The theoretical model of working memory, on the other hand, is based on the idea that it has both retaining and processing functions (Kadota, 2012).

Some researchers like Just and Carpenter (1992) and Funahashi (1996) proposed a “trade-off” relationship between the two operations of working memory, “processing” and “retaining.” According to Just and Carpenter (1992), when more cognitive resources have been consumed to “process” the incoming information, working memory cannot spare enough resources for another important operation, “retaining,” and therefore a learner cannot achieve successful understanding.

In the area of native language (L1) acquisition, several research projects have studied with regards to the relationship between the working memory capacity and reading comprehension. Although the number of studies that target L2/FL language learners is limited, several studies have also found a positive relationship between scores on working memory tests and the above categories. Researchers, such as Kato (2003) and Ikeno (2004) have found that there is a correlation between the size of a reader’s working memory and reading comprehension. Miyasako and Takatsuka (2004) has found that the more difficult the syntactic structures of the sentence were, the higher respondents scored on the working memory test. Takano (1995) argues that because Learners of English as a foreign language spend more of the cognitive resources on processing the spoken language they may not be able to consider the content of the speech because of the shortage of retaining the information given in the speech. He calls this the “foreign language effect.”

What can be understood from these arguments is that EFL reading is partially influenced by the capacity of working memory, and in order to better develop the reading skills of inefficient readers, there is a need to improve lower-level processing skills, including word recognition, so that more cognitive resources can be used to retain information and use it for more higher-level processing.

Word recognition: Orthographic processing skills and phonological processing skills

According to Grabe (2009), word recognition is comprised of two important subskills, orthographic processing and phonological processing. Orthographic processing refers to using knowledge about the writings and spelling of a printed language. Barker, Torgesen, and Wagner (1992) refer to orthographic knowledge as "memory for specific visual/spelling patterns" (p. 335). Katzir, Kim, Wolf, Kennedy, Morris, and Lovett (2006) define orthographic processing as "a visually mediated ability to analyse and recognise letter[s] and letter strings" (p. 846). In their definition, the knowledge of letter sequences is primarily comprised of knowledge about grapheme–phoneme correspondences, word structures (including morphological rules), and letter distribution. L1 studies of orthographic processing skills claim that the skill can be an accurate predicator of reading comprehension skills (Furnes & Samuelson, 2009; O'Brien, Wolf, Miller, Lovett, & Morris, 2011). In regard to L2/EFL, many studies have identified that a learner's success or failure in reading comprehension later in life can be attributed to differences in orthographic depth in L1 and L2/EFL (Miller, 2005a, 2005b) .

Orthographic depth (Frost, 1994; Frost, Katz, & Bentin, 1987) refers to the degree of correspondence between graphemes and phonemes (Koda, 2005). If a language has one-to-one relationships in grapheme–phoneme correspondence, the language is referred to as having "shallow orthography," and if the one-to-one correspondence does not exist, it has "deep orthography." The English alphabet is assumed to have a deep orthographic structure because only fifty percent of all English words have one-to-one correspondence between graphemes and phonemes (Hanna, Hanna, Hodges, & Rudorf, 1966). An example is the pronunciation of the English letter /a/, which has at least eight different possible pronunciation, as illustrated in the following words: "bad" [æ], "age" [ei], "father" [ɑ:], "many" [e], "all" [ɔ:], "about" [ə], and "usage" [I] (Narita, 2009). Moreover, as Moats (2005) admits, English words' one-to-one correspondences between graphemes and phonemes are complex and therefore challenging for L1 learners to acquire. Therefore, when not only EFL but L1 readers are asked to read aloud, they tend to read “very slowly” so that they do not pronounce words in the text inaccurately (Chang, 2012).

Japanese, by contrast, has both deep and shallow orthography because there are four different systems of characters in Japanese: Kanji, hiragana, katakana, and romaji (“the foreign style of writing Japanese language” (Suski, 1931). Kanji, a Chinese-derived character system, is a logogram in which a grapheme represents a meaning (Suski, 1931; Smith, 2012). Kanji characters can be transcribed into hiragana/katakana/romaji characters by their sound.

Hiragana, katakana, and romaji are phonograms where a grapheme represents a phoneme (an abstract sound unit that distinguishes one word from another, e.g., [p] in “pit” and [b] in “bit”), a syllable (a pronounceable unit consisting of one vowel with or without consonants), or a combination of phonemes/syllables. Phonograms have one-to-one relationships between the grapheme(s) and the phoneme(s).

Hiragana and katakana, simplified forms of Kanji characters, are equivalent to the English alphabet. They are comprised of 46 standard characters. Table 2.1 shows 46 hiragana and katakana characters with romaji and International Phonetic Alphabet (IPA) transcription, a standardized representation of sounds.

Table 2.1

Japanese Hiragana, Katakana, and Romaji with International Phonetic Alphabet (Suski, 1931)

H	K	R	I	H	K	R	I	H	K	R	I	H	K	R	I	H	K	R	I
あ	ア	a	[a]	い	イ	i	[u]	う	ウ	u	[ɯ]	え	エ	e	[e]	お	オ	o	[o]
か	カ	ka	[ka]	き	キ	ki	[ki]	く	ク	ku	[kɯ]	け	ケ	ke	[ke]	こ	コ	ko	[ko]
さ	サ	sa	[sa]	し	シ	shi	[ɕi]	す	ス	su	[su]	せ	セ	se	[se]	そ	ソ	so	[so]
た	タ	ta	[ta]	ち	チ	chi	[tɕi]	つ	ツ	tsu	[tsɯ]	て	テ	te	[te]	と	ト	to	[to]
な	ナ	na	[na]	に	ニ	ni	[ni]	ぬ	ヌ	nu	[nɯ]	ね	ネ	ne	[ne]	の	ノ	no	[no]
は	ハ	ha	[ha]	ひ	ヒ	hi	[çi]	ふ	フ	fu	[ɸu]	へ	ヘ	he	[he]	ほ	ホ	ho	[ho]
ま	マ	ma	[ma]	み	ミ	mi	[mi]	む	ム	mu	[mɯ]	め	メ	me	[me]	も	モ	mo	[mo]
や	ヤ	ya	[ja]					ゆ	ユ	yu	[jɯ]					よ	ヨ	yo	[jo]
ら	ラ	ra	[ɾa]	り	リ	ri	[ɾi]	る	ル	ru	[ɾɯ]	れ	レ		[ɾe]	ろ	ロ		[ɾo]
わ	ワ	wa	[wa]	ゐ		wi	[wi]					ゑ		we	[we]	を	ヲ		[wo]
ん	ン	n	[n]																

As shown in Table 2.1, each hiragana/katakana character can be transcribed into romaji based on its sound (Kay, 1995). Romaji, especially modern romaji, refers to characters using the "Hepburn" system of Romanization developed by J. C. Hepburn in the nineteenth century.

Although Japanese speakers rarely use romaji to write or read on actual paper, the major use of romaji is for inputting Japanese into a computer and for the transcription of loanwords. There are two ways to input Japanese into a computer: romaji or hiragana. Romaji input is convenient since people can use the English keyboard layout rather than the hiragana keyboard layout. One does not have to remember both the alphabet keyboard layout and the hiragana keyboard layout. Loanwords refer to English-derived words or words taken from another language used in their original forms, which have been difficult test items for EFL readers (Morita, 2010). Loanwords are sometimes found to be beneficial for learning EFL (Daulton, 2008). Students can infer the meaning of an unknown word from the loanwords in their first language (L1), especially when the meaning of the loanword is similar to that of the original word in English.

Some researchers argue that readers of EFL can learn more original English words (Kimura 1989; Brown & Williams, 1985; Benthusysen, 2005). Nation (2002) argues that "encouraging learners to notice this borrowing and to use the loanwords to help the learning of English is a very effective vocabulary expansion strategy." In the study of pronouncing loanwords, however, this is rarely found to be beneficial for EFL learners (Hei, 2009; Shepherd, 1996). Shepherd (op. cit.) argues that there are several "pitfalls" of loanwords, including the excessive omission of the meaning of the original words in English. He states that the original word in English usually has more than one meaning; however, loanwords tend to have only one meaning. For example, "accessory" refers to any kind of item of equipment that is not essential but can be used to make something more decorative and/or efficient, while loanwords refer only to "artificial or costume jewellery."

One important issue in the transcription of English loanwords in Japanese into katakana and romaji, however, is the phonological and orthographical "adaptation" of

English words. As Kay (1995) points out, the difficulty of reproducing foreign words leads to modification of the original pronunciation and spelling. For example, /ti/ becomes /chi/ (e.g. English "ticket" is transcribed as "chiketto" in romaji.), /th/ becomes /s/ (e.g. "thrill" to "suriru") or /di/ becomes /ji/ (e.g. "radio" to "rajio").

Moreover, although there is some variation such as abbreviations ("kiro" for "kilometre") in the pronunciation of loanwords, some researchers, such as Hei (2009) points out, that Japanese readers tend to put vowels in between all of the consonants, which is the pattern found in L1 reading. Japanese has Consonant Vowel (CV), Consonant Vowel Vowel (CVV), and Consonant Vowel Consonant Vowel (CVCV) syllable structure (Rogerson-Revell, 2011) due to the limited amount of vowels. Smith (2012) for example, states that because Japanese never finishes the final position of word in consonants, such as /d/, /t/, /k/, Japanese speakers pronounce "god" as [gada], "foot" as [futa]/, and "cook" as [kuku], all of which are Japanese loanwords listed in Japanese dictionaries in katakana. While fluent EFL readers read unknown English words based on the phonological rules, novice readers tend to put vowels between all the consonants (Shepherd, 1996). Because of the regularities of letter-sound correspondences, as Kawasaki (2013) notes, many students can already read substantially large amounts of hiragana characters when they enter a primary school. At primary school, the students learn approximately 1000 Kanji characters. Researchers such as Hamada and Koda (2011) claim that L2/EFL learners tend to utilize L1 strategies to process L2/EFL written text.

Phonological processing refers to the use of a set of knowledge about sounds in spoken language. Wei and Zhouh (2013) define phonological awareness as recognition of a group of sounds, phonemes, and syllables. There are 24 consonants and 20 vowels in English, while only 20 consonants and five vowels in Japanese (Rogerson-Revell, 2011).

Table 2.2

English Consonant Chart (Rogerson-Revell, 2011 p. 47)

	Place of articulation							
Manner of articulation	bilabial	labiodental	dental	alveolar	post-alveolar	palatal	velar	glottal
Plosive	p b			t d			k g	
Tap or Php								
Fricative		f **v**	**θ ð**	s z	ʃ **ʒ**			h
Affricate					tʃ dʒ			
Nasal	m			n			ŋ	
Lateral				**l**				
Approximant	**w**				**r**	j		

If there are two symbols in one column, the symbol on the right indicates a voiced consonant.

Table 2.3

Japanese Consonant Chart (Rogerson- Revell, 2011, p. 282)

	Place of articulation							
Manner of articulation	bilabial	labiodental	dental	alveolar	post-alveolar	palatal	velar	glottal
Plosive	p b			t d			k g	
Tap or Php				*ɾ*				
Fricative	*φ*			s z	ʃ	ç		h
Affricate				*ts*	tʃ dʒ			
Nasal	m			n			ŋ	
Lateral								

Approximant	j

Table 2.4

English Vowel Chart (Rogerson-Revell, 2011, p. 67)

	Front	Central	Back
High	i(:)		u:
Mid-high	**ɪ(:)**		**ʊ̈**
Mid-low		**ɘ**	
	e	**ɜ(:)**	ɔ(:)
Low	**æ**	**ʌ**	a / ɐ

Table 2.5

Japanese Vowel Chart (Rogerson- Revell, 2011, p. 283)

	Front	Central	Back
Close	i		<u>ɯ</u>
Close-mid			
Open-mid	ɛ		ɔ
Open	a		

Notes: Bold letters indicate that there are no corresponding phonetic symbols in Japanese, while italic letters indicate that there are no corresponding phonetic symbols in English. Underlined letters indicate that there is an equivalent phonetic symbol from one language to the other.

Regarding how English and Japanese consonants compare, as Ohata (2004) points out, Japanese speakers may be confused by the /b/v/ (e.g., “ban” and “van”) difference in spoken but not written language, since /b/ is pronounced either [b] (e.g., “berry”) or [ø] (e.g., “comb”), while /v/ is usually pronounced [v] (e.g., “vehicle”) or [f] (e.g., “leitmotiv”) (Narita, 2009). The same theory applies to /r/l/ (e.g.., “arrive” and “alive”). The most problematic of all is [θ/ð] because, Japanese speakers replace [θ] with [s] or [t] (e.g., “theory” is read as /seori:/) —and [ð] with [z] or [d] (e.g., “they” is read as [zei]). Both “theory” and “they” share the same spelling /th/ though it is pronounced differently.

With regards to how English and Japanese vowels compare, because the number of Japanese vowels is quite limited compared to that of English, several ‘replacement’ take place. For example, the absence of the phonetic symbol of [æ] in

Japanese may cause some confusion since /a/ is usually replaced by [a] (e.g., “want” and “fat”). (Kameyama, 1992).

Studies of L1 reading indicate that there is a correlation between proficiency in phonological processing and reading comprehension skills (Badian, 2001; Vellutino, Tunmer, Jaccard, & Chen, 2007). Several L2 studies on relationships between phonological processing and reading comprehension skills have also revealed that proficiency in phonological processing is positively related to reading comprehension (Kojima, 2010; Koshimizu, 2010). Koshimizu (op. cit.) argues that fluent readers use non-lexical route to process words while unskilled readers use lexical-routes to process words, which takes more time than non-lexical routes. Therefore, when unskilled readers are asked to read aloud, they often read very slowly and pay more attention to pronouncing the words correctly. Ishikawa (2008) points out that there is a discrepancy between the ability to process a word phonetically and semantically. He argues that the discrepancy between the ability to process a word phonetically and semantically was observed, especially with lower-level EFL readers. When high-level readers read written text, phonetic and semantic processing occur simultaneously, while in lower-level readers’ reading processes, the phonological processing occurs after the semantic processing (Ishikawa & Ishikawa, 2008).

Kadota (2012) argues that there are two routes for word identification for second-language learners. One is a route that transfers visual input into an orthographic representation, and then transfers it into a phonological representation before changing it into a semantic representation (Route A). The other route is omits a phonological representation and reaches semantic representation directly from orthographical input (Route B). Fluent readers tend to use this first route by automatizing the phonological process; when they fail to make the transfer, they still have the second route as a backup strategy.

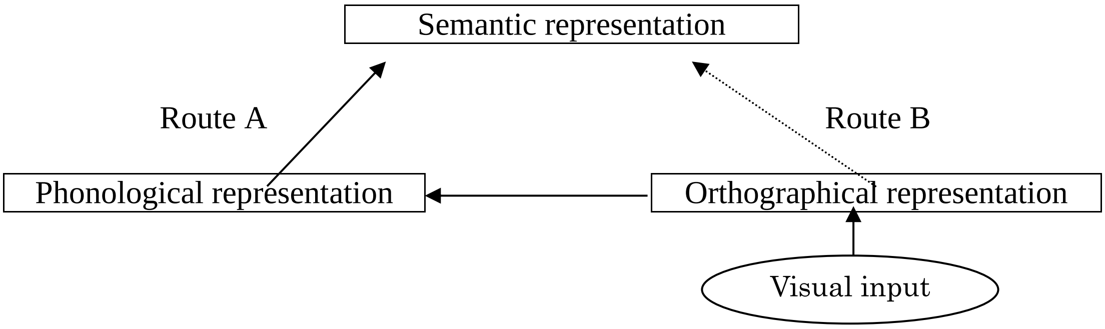


Figure 2.6. Revised version of 2007 dual access model of word processing (Kadota, 2012, p.133).

Measurement of orthographic and phonological awareness of Japanese learners of English

The measurement of orthographic and phonological processing skills of Japanese learners of English should focus on several issues concerning L1 and FL differences in orthography and phonology. First, assessment should take into account whether a grapheme of vocabulary has one-to-one correspondence with a phoneme and whether the phoneme exists in Japanese. As described above, Japanese vowels and consonants have one-to-one relationships with their graphemes and phonemes, even in romaji: /a/ is exclusively pronounced as [a], /e/ as [ɛ], /i/ as [i], /o/as [ɔ], and /u/ as [ɯ̲]. Therefore, when phonemic symbols such as [ɪ(:)], [ʊ], [ɘ], [ɜ(:)], [æ], and [ʌ] do correspond with the graphemes /a/, /e/, /i/, /o/, and /u/, reading becomes challenging for Japanese learners of English. Table 2.6 shows the modified version of Narita's (2009) table of English phonemes that correspond with graphemes. To clarify the differences between Japanese and English grapheme–phoneme relationships, particularly in terms of how letters /a/, /e/, /i/, /o/, and /u/ can be pronounced differently between the two languages, English phonemes are divided into three categories: one, phonetic symbol that has same grapheme as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese (NNE). For Japanese learners of English, the reading of a grapheme /a/ as [æ] or [ɔ:] is more difficult than reading of /a/ as [ei] because the pronunciation of [ei] exists in Japanese language.

Table 2.6

The Modified Version of Narita's (2009) table of English Phonemes that Correspond with Grapheme

G		/a/			/e/			/i/			/u/			/o/	
P	JE	NE	NNE	JE	NE	NNE	JE	NE	NNE	JE	NE	NNE	JE	NE	NNE
	[a]	[aː]	[æ]	[e]	[a]	[φ]	[i]	[iː]	[æ]	[ɯ]	[e]	[ʌ]	[ɔ]	[ɔː]	[əʊ]
		[ɪ]	[ə]		[iː]	[ə]		[ɪ]	[ə]		[w]	[φ]		[a]	[ʌ]
		[ɔː]			[ɪ]			[aɪ]			[ju]	[ə]		[ʊ]	[ɜː]
		[eɪ]			[eɪ]						[ʊ]			[uː]	[ə]
											[ɪ]			[ɪ]	

G=Grapheme, P=Phoneme

JE= phonetic symbol that has same grapheme as the Japanese one-to-one correspondences between graphemes and phonemes

NE= phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation

NNE= phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese

Second, assessment should consider whether the target word is a loanword. As described in the earlier paragraph, loanwords tend to undergo orthographic and/or phonological adaptation, which if done regarding the particular term, makes reading the term in English challenging for Japanese readers who will be affected by romaji spellings and pronunciation.

Third, whether the syllable ends with a consonant should also be considered. As described in the earlier paragraph, Japanese syllables usually end with a vowel, meaning that a syllable is CV, CVV, or CVCV. By contrast, English syllables can end with a consonant, which may cause Japanese readers to add a vowel to a syllable without a vowel.

2.3 Assessing the usefulness of a language test

Bachman and Palmer (1996) argue that reliability and validity are the two characteristics that have a profound significance when considering a test's usefulness. Unlike measurement of height and weight, the measurement of language ability is based on indirect observation of invisible traits. Therefore, investigation into the reliability and validity provides evidence in determining whether a test actually assesses that which it claims to measure.

2.3.1 Reliability

Reliability of a test refers to the level of consistency of measurement. Language test scores or testees' performance on language tests are affected by several factors dubbed "measurement error" by Bachman (1990, p.160). He categorized factors that affect language performance into four groups: communicative language ability, test method facets, personal attributes, and random factors. He argues that personal attributes, such as prior knowledge of the test contents, and random factors including emotions and affects, are impossible to completely control. There is a need to minimize the effect of such factors, and therefore, there is a need to identify the source of error which is one of the most difficult issues. This relates to the issues proposed in the preceding section on the limitations of measurement.

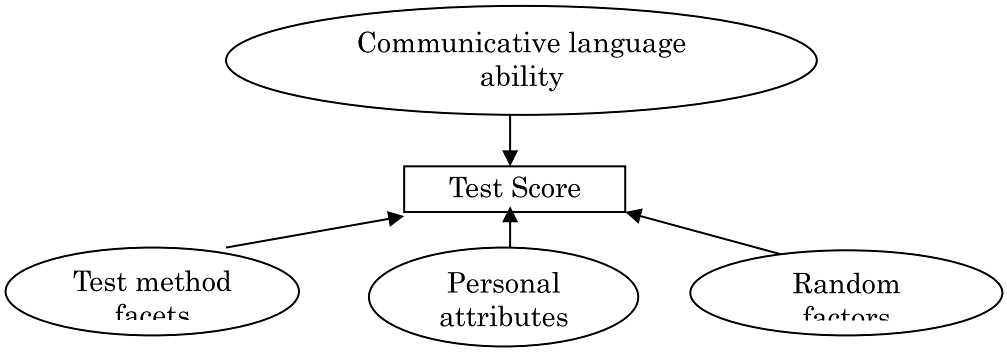


Figure 2.7. Factors that affect language test scores (Bachman, 1990, p.165).

2.3.2 Validity

Fulcher and Davidson (2007) indicate that validity is the "central concept in testing and assessment" (p. 3). They define validity as a process of collecting "evidence to support specific interpretations of test scores" (p. 159). The Standards for Educational and Psychological Testing (AERA et al., 1999, p. 9) states:

> Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kind of evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. As validation proceeds and new evidence about the meaning of a test's score becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test.

The best known theoretical concept of validity was proposed by Messick (1989) who defined validity as a unitary concept. He said, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p.13). Traditionally, validity has been assumed to be composed of content-, criterion-, and construct-related validity. Content validity refers to the level of correspondence between the test items and its specification. Because specifications should include detailed accounts of skills and items that the test intends to measure, the investigation into the relationship between test items and specifications provide evidence for content validity. Criterion-related validity refers to the relationship between the test and the other independent measurement on the same subject to determine whether the test scores correlate with each other. If the two scores show positive correlation, the target test is assumed to possess criterion-related validity. However, along with the gradual emphasis on construct validity, Cronbach (1988) proposed the notion that all validation is a single quality. Following this view, Messick (1980, 1989) proposed two interconnected facets of the unitary validity concept and offered a fourfold classification system in his progressive matrix.

Table 2.7

Progressive Matrixes for Facets of Validity (Messick, 1989, p. 20)

	Function of outcome	
Source of justification	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Value implications	Social consequence

According to his view of the unitary validity concept:

> One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of the testing, being either interpretation or use. If the facet for source of justification (that is, either an evidential basis or a consequential basis) is crossed with the facet for function or outcome of the testing (that is, either test interpretation or test use), we obtain a four-fold classification. (Messick, 1989, p.20)

Following Messick’s theoretical view of validity as a unitary concept, a framework for the systematic validation of the placement test was proposed by Koizumi (2005). She proposed a systematic validation that includes six components of test validation and research methods which will be used as the baseline framework in this study.

Table 2.8

Six components of validity and analysis procedures (Koizumi, 2005; adapted from

Bachman et al., 1997; Banerjee et al., 1997; Chapelle, 1999; Cheng et al. 2004; Messick, 1989, 1996)

Components	Items	Analysis methods
Content validity	Whether the test contents correspond to the purposes of measurement.	▪ Expert judgement ▪ Task analysis
Face validity	Whether the test items appear to have assessed the abilities to be measured.	▪ Questionnaire ▪ Observation ▪ Interview ▪ Discourse analysis
Construct validity	Whether the test items relate to the abilities the test aims to assess.	▪ Factor analysis ▪ Item response analysis
Reliability	Whether a test score can be generalizable or not.	▪ Generalizability theory ▪ Reliability ▪ (DIF) ▪ Analysis of variance (ANOVA)
Criterion	Whether the test scores correlate with the results of other measurements.	▪ Correlation ▪ Structural Equation Modelling (SEM)
Washback	Whether the test use/ interpretation has positive or negative effect on people who use tests.	▪ Observation ▪ Interview ▪ Questionnaire

2.4 Test theories

This section presents a literature review that justifies the use of a particular measurement theory, the Rasch model, to test language ability in language programs. Before discussing the Rasch model, this paper will first summarize some of the fundamental differences between classical test theory (CTT) and item response theory (IRT). It will then describe one of the most important research issues related to test development and offer arguments for the use of the Rasch model in language testing. The discussion will cover how item response theories and the Rasch model differ from classic test theory and how the Rasch model is different from item response theories, which will lead to why the Rasch model is suitable for developing an instrument that tests language ability in relatively small groups of testees where test administrators are not language testing professionals.

2.4.1 Classical test theory

Classical test theory (CTT), sometimes referred to as “classical reliability theory” or “true score theory”, is one of the most frequently used approaches in test analysis that uses correlation coefficients (Bachman, 1990). The analysis of test data relies on several frequently used terms, including “mean,” “standard deviation,” and “variance” which are the components of descriptive statistics; “correlation coefficient,” which is used to provide information regarding items’ reliability; and “item difficulty,” “facility value,” and “(item) discrimination power index,” which are used to provide information on the validity of test items (Nakamura & Ohtomo, 2002). These terms are all broadly categorized as part of CTT, under which estimates fall into three broad categories: fundamental statistics, item analysis, and reliability. One major characteristic of CTT is that it bases data analyses on raw or number-right scores.

Compared to more recent innovative and sophisticated types of statistical analysis, CTT has been found suitable for the identification of individual examinees’ relative rankings within groups. For example, if an institution decides to rank applicants according to their level of English proficiency, so that it can decide which students to admit, it needs to know exactly which students are most suitable. For this reason, several educational institutions use CTT to help them make enrolment decisions (Nakano, Ueda, Oya, & Tsutui, 2004; Maeda, 2003). Moreover, one of the biggest advantages of CTT for educational practitioners is that it can be carried out using basic computer software, such as Microsoft Excel. This is especially beneficial in Japan, for example, where many English-language teachers are professionals but are reluctant to use highly-advanced statistical tools for data analysis. Therefore, a relatively large number of studies regarding test data analysis have been done using CTT in Japan (Shimizu, Kimura, Sugino, Yamakawa, Ohba, & Nakano, 2003; Nakano, Ueda, Ohya, & Tsutui, 2004; Hirose, 2004, 2005).

However, in the context of CTT, several problems exist in relation to the characteristics of measurement. First, as Bond and Fox (2007) have posited “…the scales to which we routinely ascribe that measurement status in the human sciences are often merely presumed to have interval-level measurement properties; those measurement properties are almost never tested empirically. It is not good enough to allocate numbers to human behaviours and then merely to assert that this is measurement in the social sciences” (p.4). A fundamental problem concerning the validity evidence for in-house test items is that collected data have been analysed with the assumption that they are interval (i.e. linear and ratio measures) although they are in fact ordinal (i.e. linear measures). For example, student A got 10, student B got 9, and student C got 8 in the English vocabulary test respectively. The difference of

English vocabulary proficiency levels between student A and student B and student B and student C are not the same since the scores of the English vocabulary test are based on an ordinal scale and not on a ratio scale.

Second, test scores are innately test-dependent (McNamara, 2000; Bond & Fox, 2007). If an individual test taker completes two different sets of test items designed to measure the same abilities, his/her test score might change. Because test score is essentially the product of a particular set of test items, the comparison of different test scores is impossible. For example, when a person takes an English speaking test and receives a relatively low score, one cannot say that this person lacks the ability to speak English, because the test might have been unreasonably difficult. He/she might score higher on a different test.

Third, CTT item characteristic estimates are sample-dependent (McNamara, 2000; Nakamura & Ohtomo, 2002). When a group of examinees takes a test that is defined as difficult because the average score was 50%, we still cannot say that the test was difficult. Indeed, the level of difficulty might differ if another group of examinees, with higher proficiency levels, were to take the same test.

Fourth, CTT only provides a single global estimate for a group of examinees (Bond & Fox, 2007; McNamara, 2000; Nakamura & Ohtomo, 2002). This means that it does not produce information on individual items' difficulty. Therefore, if the difficulty of a particular test item is estimated to have a 0.50 p-value (probability value), this does not mean that other items on the same test are of similar difficulty. They may have p-values of 0.20 or 0.70.

The problems described above can be traced back to two fundamental limitations of CTT: lack of sample/item independence and lack of linear and ratio measures or lack of calibration using a common unit (Ohtomo, 1996; McNamara, 2000). Number-right scores are based on ordinal scales where the ranking of scores is based on the relative order of scores, but the interval between scores is arbitrary. In such cases, the meanings of scores are distorted. For example, an interval of 5 between the scores of 95 and 100 is different from an interval of 5 between 40 and 45. Because there is no score above 100, those students with the ability to score 120 or more are also categorized within this interval. This is an example of the “ceiling effect.” Wright (1997) points out that most statistical procedures require interval scales. Ordinal scale scores must be converted into interval scale scores, as with the Rasch model.

2.4.2 Item response theory

Item response theory (IRT) models, “probabilistic model[s] of test performance” (Lynch, 2003, p. 92), use probability theory to find the probability of examinees’ answering particular items correctly. While the IRT models most in use were invented and developed by two American researchers, Frederic M. Lord and Paul Lazarsfeld, in the 1950s, Rasch analysis, developed by Danish researcher George Rasch, is also important and shares some features with Lord and Lazarsfeld’s approach, including the use of a one-parameter logistic model. Therefore, several researchers, including Ohtomo (1996) and Ikeda (1994), have assumed that the Rasch model is an IRT model. This argument will be discussed in the next section. IRT models are claimed to have overcome many of the limitations of CTT and are becoming the most frequently used tool in data analysis (Bachman, 1990; Nakamura, 2007). This section will describe the basic characteristics of IRT, including the pros and cons of using it in an educational context.

Instead of using number-right scores, IRT uses logit scores converted from natural logarithms. As noted above, number-right scores do not have absolute zeroes or regular interval scales, whereas logit scores do have interval scales. Logit scores are calculated using the natural logarithm of a number that is equal to the percentage of correct answers divided by the percentage of incorrect answers, as given in the formula below:

Logit score = $\ln (p/(1 - p))$

(*ln refers to the natural logarithm, p refers to the percentage of correct answers)

Applying this formula to a group of students whose number-right scores are 45, 55, 85, and 95, figure 2.8., below, shows the logit scores converted from the number-right scores and gaps.

	number-right score	logit score	number-right score	logit score
	55	0.202	95	2.994
	45	−0.201	85	1.735
gap	10	0.402	10	1.209

Figure 2.8. Gap between the number-right scores and logit scores.

As seen in the figure above, in the number-right score category, the gaps between 55 and 45, and 95 and 85, are 10, while the logit scores are 0.402 and 1.209. This clearly shows that the gaps between the number-right scores are not regular intervals.

IRT predicts the probability of each examinee correctly answering each item. The formula used to make such a prediction is called a model. The model calculates the probability of an examinee answering an item correctly, according to the relationships between time characteristics and examinee ability. For example, if an item is “difficult,” the probability that it will be answered correctly is reduced, while the opposite is true for “easy items.” Similarly, the probability of an examinee with “low” ability answering items correctly is also reduced. The so-called “parameter” of items and examinee ability is based on an interval scale. Wilson (2005, 2008) has introduced a “developmental perspective” of students learning. He claims that assessment should be organised using sound measurement with four principles: a developmental perspective, a match between instruction and assessment, the generating of quality evidence, and management by instructors to allow appropriate feedback, feed forward and follow-up.

Direction of increasing “X”

Respondents	+Logit	Responses to Item
	+3.0	
Respondents with high “X”		Item response indicates highest level of “X”
	+2.0	
	+1.0	
	0.0	
	−1.0	
Respondents with low “X”	−2.0	Item response indicates lowest level of “X”
	−3.0	

Figure 2.9. A generic construct map in construct “X” (adapted from Lynch, 2003, p. 93 and Wilson, 2005, p. 27).

The aforementioned characteristics of IRT provide several benefits to testing research. First, the IRT model provides test-free estimates. Therefore, comparison of different sets of tests is possible. Second, the IRT model provides sample-free item calibration, because estimates of item difficulty are independent of sample ability. Third, the IRT model assesses individual test-takers’ performance on test items, according to item difficulty and examinee ability. In other words, test item and examinee ability parameters are “invariant” across tests. This means that these parameters have absolute values, no matter who takes a test or what test items are used. Nakano, Ueda, Ohya, and Tsutui (2004) have compared the results of both CTT and IRT English placement tests administered at Waseda University and concluded that because CTT tests are sample-dependent, 20% of the test items in one section were inappropriate. IRT analysis, however, being sample-independent, provided more precise information on the appropriateness of the test items.

One reason why IRT is widely used by test developers is that the difficulty of IRT test items is known before the tests are administered. This is closely related to the issue of item banking, where large sets of “equated” items from different tests are collected and categorized according to their contents and level of difficulty. The term “equation” refers to test items that have been collected from different test sources and placed on a common scale using item response theory. Hughes (2003) has stated that the benefits of item banking are that it contributes greatly to: saving time and costs, because test developers do not need to start from scratch, securing the quality of testing, because it has been used at least once and, therefore, been trialled, and maintaining the standards, fairness, and evaluation of teaching. Although Lawrence (1998, p. 4) has observed that users of item banks need to review test items for “technical quality, curriculum match and potential bias,” it is possible to accurately predict which items may cause difficulty.

2.4.3 IRT models and implications for use in individual tertiary-level institutions

There are three main IRT models: one-parameter, two-parameter, and three parameter. Each model relies on different formulas and assumptions regarding item properties and requires a different number of samples and items for estimations to be valid.

One parameter (Rasch) IRT model (1 PLM)

The one-parameter IRT model and Rasch model both use mathematically equivalent formulas and provide the *b* parameter–namely, item difficulty. While the range of the *b* parameter is theoretically infinite, it is normally between -3 (easiest) and +3 (most difficult). Because the model requires a relatively small amount of samples and items (100 and 20, respectively), it is assumed to be the easiest to administer for practitioners with limited statistical and computer processing skills (Nakamura & Ohtomo, 2002). The model is particularly popular in Japan, and much research has been done on its use (see Wistner, Sakai, & Abe, 2009; Nakamura, 2007).

Two-parameter IRT model (2PLM)

While the one-parameter model only provides parameter *b*, the two-parameter IRT model provides the *a* parameter (indicating item discrimination). The larger the *a* parameter, the larger an item’s level of discrimination at a particular level of difficulty. For purposes of accurate estimation, the model assumes 30 items for every 200-500 participants (Ohtomo, 1996).

Three-parameter IRT model (3PLM)

The three-parameter model estimates the *c* parameter (known as “the guessing parameter”) and is added to the *a* and *b* parameters of the two-parameter model. By identifying lower-ability students who answer correctly by guessing, this model makes it possible to better account for lower level participants' data. Such items would be quantified as misfits to the one- and two-parameter models, while they would not be identified at all by CTT. To achieve accurate estimations, the model requires 60-80 items and 1,000 participants (Ohtomo, 1996; Toyoda, 2002). The IRT model is assumed to be ideal for item banking, which refers to the storing of items that are calibrated according to individual test-taker ability, item difficulty, and powers of discrimination. The idea of item banking became popular in congruence with the development of the IRT model, which allows test developers to easily retrieve the test items best suited to their students' proficiency levels and also to compare the results of recent tests to tests used in the past. Most commercially-available testing services, such as ETS and Cambridge English for Speakers of Other Languages, use test items that have been calibrated using pilot testing and, therefore, claim invariance in the estimates obtained for all different versions of tests administered to different examinees and including different test items.

2.4.4 The difference between the Rasch and the one-parameter IRT model

This research has roughly defined two fundamentally differing models—the Rasch model and the one-parameter IRT model—as exchangeable concepts. Some researchers have defined the Rasch model as a synonym for the one-parameter IRT model and treated these as interchangeable concepts (see Watanabe & Noguchi, 1999). Others, however, claim that the models are fundamentally different, with regard to their basic assumptions. Smith, Linacre and Smith (2003) have stated, in the *Journal of Applied Measurement's* submission guidelines, "We do not encourage the use of 'Item Response Theory' as a term for Rasch measurement."

However, Shizuka (2007), one of the pioneers of Rasch analysis in Japan, has stated that the differences between the two models lies in their purpose. According to him, the Rasch model aims to design tests that make objective measurements, while IRT aims to develop models that best describe the data acquired (Wilson, 2005). He quotes Embreston and Hershberger (1999, p. 252):

> Individuals who strongly prefer particular IRT models place different values on two fundamental issues; empirical fit of test data

> to a model versus designing a test to fit a justifiable measurement model. In one case the data is considered fundamental, whereas in the other, the model is considered more fundamental.

Therefore, as Andrich (1989, p. 14) has observed, "Rasch's specifications are requirements for the data to produce measurements and not assumption about the data." The models provided by Rasch analysis are only valid for particular sets of data. This means that the Rasch model is an ideal that no data ever perfectly fits. Unlike IRT models that try to provide models that perfectly fit data, the Rasch model seeks to identify degree of misfit and determine whether it precludes meaningful measurement. The idea is well expressed in Wright and Masters (1982, p. 102): "When items do not fit, that signifies to us not the occasion for a looser model, but the need for better items."

This paper has presented a brief overview of fundamental considerations of the characteristics of measurement and a discussion of the use of Rasch-based analysis in language testing. It can be seen that Rasch-based analysis is suitable in language testing for several reasons, including its test-independence, sample-independence, quantifiability, and linearity, which enable test administrators to gain more precise information regarding testees. Moreover, the Rasch model provides us with data on misfit items, making the improvement of items more feasible (Wilson, 2005). As Souji (2007) has noted, Rasch-based analysis requires testers to rewrite test items, rather than modify test data. This is more appropriate, especially for those who are not used to sophisticated statistical tools.

For these reasons, it can be suggested that the use of the Rasch model is most appropriate for test development in individual educational programs, where tests are developed using small sample sizes and with relatively limited resources and where practitioners are not language testing professionals.

Chapter 3: Curriculum Analysis

The purpose of this chapter is to report on the processes and results of a preliminary study for Research Question (RQ) 1: What kinds of abilities are required in the Faculty of Nursing's curriculum regarding orthographic knowledge and phonological awareness? In practice, the chapter aims to examine the syllabus, textbook, and a reference book to identify the orthographic and phonological features of English education at the Faculty of Nursing.

3.1. Materials

Materials and rationales for the choices for RQ 1 are given below.

- syllabuses of the Faculty of Nursing's English subjects
- a list of textbooks used in the faculty's English classes
- a list of a medical vocabulary in a reference book used in the faculty's English classes.

As described in the chapter 1, the selection of English course textbooks is left to each lecturer as long as the textbooks reflect the aims of the faculty's curriculum. The list below shows the textbooks adopted by all lecturers in 2015. Of thirteen textbooks, five are for communicative purposes and the rest are for medicine, nursing, and culture.

Table 3.1

Textbooks Used in the Faculty of Nursing at Toho University in 2015

Title	Publisher	Major focus	The number of courses	Course names
English file elementary 3rd edition	Oxford University Press	Communication	2	EC*
New interchange	Cambridge University Press	Communication	2	EC*
Topic talk class	EFL Press	Communication	2	EC*
Touchstone	Cambridge University Press	Communication	2	EC*
Let's check out the UK!	Kinseido	Communication/ Culture	2	EC*
ESP for food literacy	Eihosha	Health	4	EC*
Health talk	Pearson Education	Health	6	ME*
Healthy living	Nanundo	Health	2	EC*
Caregiver: Reading the current medical world	Asahi Press	Medical	4	EC*
English for medicine	Kinseido	Medicine	2	EC*
Nursing 1	Oxford	Nursing	2	EC*
Nursing case studies	Seibido	Nursing/Health	2	ME*
TOEIC official collection of past questions	Educational Testing Service	TOEIC preparation	1	EC*

* EC=English for Communication, ME=Medical English

There are not any consistencies in the selection of textbooks regarding the content. While the focus areas of the textbook show variety, one of the reference books, a medical vocabulary book, *"Igakueitango [Medical English vocabulary*]," is recommended for purchase. The book has a list of 493 medical terms. Because all students are encouraged to purchase the book, the vocabulary list used in the book is included as material in this research.

3.2. Research process

For the analysis of English curriculum at the Faculty of Nursing, the author analysed syllabus's aims and purposes of each subject. The process of identifying the orthographic and phonological features of the vocabulary list used in the reference book is provided here.

- The author transcribes all words in alphabetical order in an Excel file
- Words were copied and pasted to the VocabProfiles of Compleat Lexical Tutor (Cobb, 2002; Heatley, Nation, & Coxhead, 2002), a computer program that categorizes English words into four categories by frequencies: the 1,000 most frequently used words (K1), the second 1,000 most frequently used words (K2), the 570 most frequently used academic words (AWL), and off-listed words (OFF).
- The phonetic symbol of each word was checked using *Taishukan's Unabridged Genius English-Japanese Dictionary*. When the phonetic symbol of the word was not included in the *Genius*, the *Shogakukan Random House English-Japanese Dictionary 2nd edition* was used.
- Phonetic symbols were separated into different rows of the Excel file based on the syllables.
- The author checked whether the syllable ended with a consonant and counted the number of occurrences.
- Each phoneme of vowels was matched with corresponding graphemes.
- All graphemes of syllables were rearranged by alphabetical order
- All graphemes of syllables were divided into three categories: one, phonetic symbol that has same grapheme as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese (NNE)

An example of the list of words with their phonetic symbols is shown below in Table 3.2.

Table 3.2

An Example Of The List Of Medical Terms

	Term	Phonetic symbol	Number of syllables	1st syllable' grapheme and phoneme	2nd	3rd	4th	5th	6th	7th	8th	9th	Number of non-open ended syllables
1	arthritis	arˈθraɪ. təs	2	a=a	i=ai	is=əs							2
2	cholic	kou lik	2	o=ou	i=i								1
3	cystic	sis tik	2	y=i	i=i								2
4	keloid	kí: lɔid	2	e=i:	oi=ɔi								1
5	mucoid	mjú: kɔid	2	u=ju:	oi=oi								1
6	mucous	mjú: kəs	2	u=ju:	o=ə								1
7	nasal	néi zəl	2	a=ei	a=ə								
8	neoplasm	ní:ə plæ`zm	2	eo=í:ə	a=æ								1
9	optics	ɑ'p tiks	2	o=a	i=i								1
10	ovoid	óu vɔid	2	o=ou	oi=ɔi								1
11	pelvic	pél vik	2	e=e	i=i								1

12	plasmid	plæ'z mid	2	a=æ	i=i	2
13	pleural	plúə rəl	2	eu=iə	a=ə	1
14	renal	rí: nl	2	e=i:	a=*	1

3.3 Results

This section describes the method and results of the curriculum analysis and then explains the method and results of the vocabulary analysis regarding orthographic and phonologic features.

3.3.1 Curriculum

Curriculum for English-language subjects at the Faculty of Nursing is relatively limited. Although all the liberal arts subjects at the faculty aim to “cultivate humanity and sensitivity… as well as broadening views,” there are no explicit curriculums for each subject. All the English-language subjects are categorized under a liberal arts education domain, and the purpose of the domain is “to broaden the perspectives of the students.” There are six compulsory English-language subjects for first- to fourth-year students: *English for Communication 1-4* and *Medical English 1/2*. The statements of purpose on each subjects’ syllabus are as follows:

> · *English for Communication 1/2*
>
> Students will acquire the basics of English communication skills by improving four basic language skills: reading, listening, writing, and speaking. In listening- and speaking-focused classrooms, English will be used as a medium of communication so that the students can acquire everyday conversation skills. In reading- and writing-focused classrooms, the ability to understand the organization and main ideas of English written texts will be stressed, as well as an ability to express ideas and thoughts clearly and concisely.
>
> · *English for Communication 3/4*
>
> Students will advance their English communication skills by engaging in four language skills: reading, listening, writing, and speaking. Moreover, students will learn the basic structures of medical English passages, as well as increase their medical English vocabulary.
>
> · *Medical English 1/2*
>
> Students will increase their medical English vocabulary and practice basic skills in order to read academic texts about nursing fluently and accurately. The class

also focuses on using clinical expressions so that students can prepare for communicating with English speakers at hospitals in the near future.

A close reading the syllabuses reveals a hidden curriculum, which can be summarized as having one chief purpose: to stress that students should acquire both communication and academic skills in English. Goals should include fluent reading and listening skills in both medical and everyday English, expanded vocabulary in medical English, and basic speaking and writing skills for self-expression. According to this purpose and its goals, curriculum should expand not only students' medical English vocabulary but also their fluency in reading medical texts.

In the Faculty of Nursing, English fluency in reading medical texts is a crucial skill, especially for students seeking to pursue graduate study. Since most graduates of postgraduate courses will become teachers at tertiary-level institutions, many entrance examinations for postgraduate courses at schools of nursing consider English fluency in reading medical texts to be a requisite skill. As shown in Table 3.3, of five prominent schools of nursing within the Tokyo district (St. Luke's International University, Jyuntendo University, Kitasato University, Toho University, and Tokyo Women's Medical University), all but one (Tokyo Women's Medical University) consider English skills to be a pivotal entrance condition. English fluency in reading medical texts can provide future tertiary-level lecturers with ways to absorb new information and skills in nursing practices, all of which are vital to becoming educators of future nurses.

Table 3.3

Entrance Examination Subjects for Postgraduate Studies

University	English	Major subject	Essay	Interview	Résumé
St. Luke's International	90 min test	75 min test	90 min essay	30 min interview	Required
Jyuntendo	60 min test	60 min test	N/A	60 min interview	Required
Kitasato	90 min test	90 min. test	N/A	30 min interview	Required
Toho	60 min. test	60 min. test	N/A	30 min interview	Required
Tokyo Women's	N/A	N/A	N/A	N/A	Required

As described in chapter 2, reading fluency requires automatic recognition of words, which is based on the orthographic knowledge and phonological awareness of English words. Moreover, as Miyoshi, Naito, and Tozawa (2011) point out, the vocabulary used in medical English is relatively “challenging” for novice readers in English with limited orthographic knowledge and phonological awareness. Therefore, the measurement of accurate orthographic knowledge and phonological awareness in English placement testing is a prerequisite for the students of the Faculty of Nursing.

3.3.2 List of English medical terms in the reference book

All 493 words are categorized as off-list words (OFF), which exceeds the average ratio of English written text, K1 (70%), K2(10%), AWL (10%), and OFF (10%). Since the list is not a sentence, more words are categorized as less frequent words, but this seems to suggest that words in the list are not frequently used words and, therefore, challenging for the students to acquire.

The average number of syllables per word in the reference book’s vocabulary list was 4.41. There were no single syllable words, while some words had more than seven syllables. The words with the most syllables was “otorhinolaryngology” [ou/tou/rai/nou/lær/iŋ/ gɑ'l/ə/ʤi] (slash indicating syllabic boundaries).

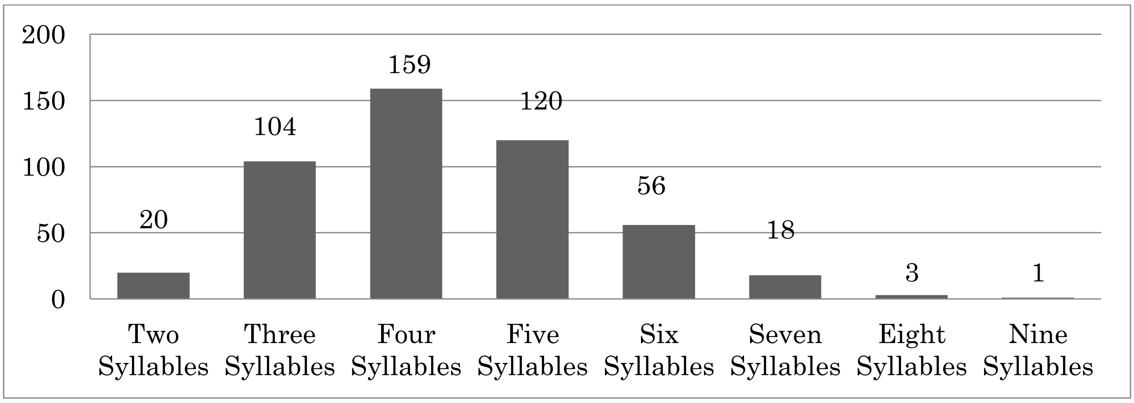


Figure 3.1. Number of words by syllables for words in the reference book.

Grapheme

As shown in Figure 3.2., all graphemes except /u/ (3%) comprise approximately a quarter of all graphemes in the list of English medical terms in the reference book respectively. This result will be compared with that of the two commercially produced English tests, the TOEFL and the Test of English for International Communication (TOEIC), in the next chapter.

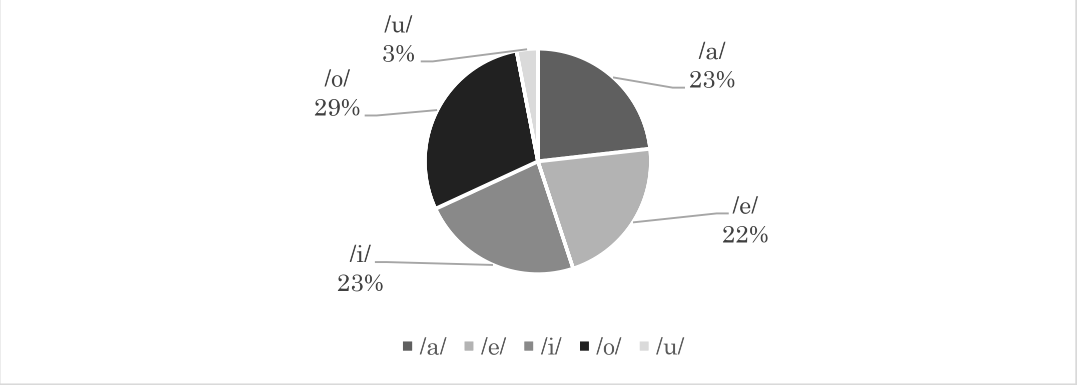


Figure 3.2. Percentage of graphemes /a/, /e/, /i/, /o/, and /u/ in the list of English medical terms in the reference book.

Comparison with L1 grapheme–phoneme correspondence

As described in chapter two, the author has divided the vowels of English phonemes into three categories: one, phonetic symbols that have same graphemes as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese (NNE). This categorization is used to identify the orthographic and phonological features of the two tests. For the grapheme /a/, the phonetic symbol [a] will be categorized as JE; [a:], [ai], [e], [i] and [o] are categorized as NE; and [ə], [ɔ(:)], [æ], and [ʌ] are categorized as NNE. Figure 4.3 shows the percentage of JE, NE, and NNE for the graphemes /a/, [e], [i], [o], and [u] in the English medical terms in the reference book. The results indicate that graphemes /a/ and /u/ are rarely pronounced in English in the way Japanese pronounce the graphemes /a/ and /u/. Moreover, the number of NNE vowels exceeds not only that of JE but also NE. Reading the graphemes /a/ and /u/, therefore, is deemed to be more challenging for Japanese learners of English. On the other hand, pronouncing the grapheme /i/ is less challenging, since majority of the grapheme /i/ will be read as [i], which is the same as the Japanese grapheme–phoneme correspondence.

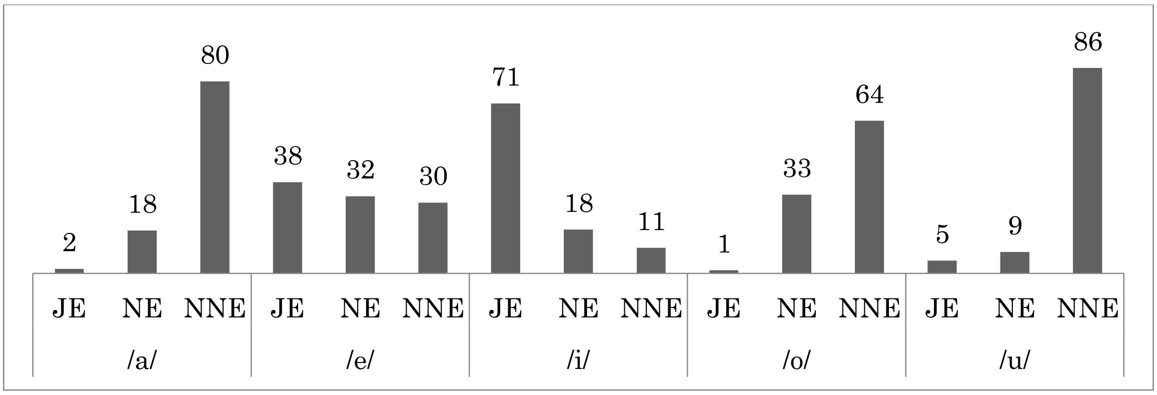


Figure 3.3. The number of JE, NE, and NNE phonemes of the graphemes /a/, /e/, /i/, /o/, and /u/ in the English medical terms of the reference book.

Open-ended syllables

Of 2083 syllables, 657 syllables (31%) ended with consonant. This result will also be compared with that of the TOEFL and the TOEIC in the next chapter to identify whether there is a difference between the English vocabulary that is encouraged in the faculty to acquire and the vocabulary of the reading section in the TOEFL and the TOEIC.

Loanwords

Of 481 words, five words (1%), homosexual, hypertension, keloid, mammography, and parasite are categorized as a loanwords. Although the number of loanwords are limited in the medical terms, all five terms are “adapted” from the original English term when pronounced in Japanese language. For example, homosexual becomes [homəsekʃuəlu], the grapheme/o/ is pronounced [o] (in English, [ou]) and an addition of a vowel [u] after the grapheme /l/.

3.4 Summary findings and conclusion

This chapter has thus far analysed the faculty’s syllabus requiring students’ development of orthographic and phonological awareness as a means of achieving reading fluency and expanding their medical vocabulary.

The analysis of vocabulary contained in the reference book has revealed that the vocabulary level of medical terms is not only high, but also demonstrates several orthographic and phonological features infrequently addressed in Japanese English education. Firstly, the terms are comprised of relatively large number of syllables. Moreover, several graphemes, especially /a/, /o/, and /u/ have more difficult grapheme–phoneme correspondences. Their phonemes do not exist in Japanese pronunciation. Thirdly, more than 30% of all syllables end with consonant, which cause a trouble with Japanese learners of English. Japanese syllables has open-ended syllables therefore, they put extra vowels after consonants. Finally, although limited in the number, there are loanwords in the vocabulary. Japanese learners of English tend to adapt the English terms into Japanese pronunciation. Proficiency in orthographic and phonological awareness would benefit the Japanese learners of English to read fluently. The faculty would therefore benefit from placing students in appropriate English classes based on their abilities with English orthography and phonology.

Chapter 4: Commercially Produced English-language Tests

The purpose of this chapter is to report the processes and results of a study conducted to answer Research Question (RQ) 2—that is, what kinds of test items are used and what kinds of constructs measured by practice tests of commercially produced English-language proficiency tests, and how well do the abilities assessed reflect the content of the faculty's curriculum? In practice, the chapter aims to examine the extent to which the test items and constructs of two commercially produced English-language proficiency tests frequently used for placement purposes in Japanese universities—Test of English as a Foreign Language (TOEFL) and Test of English for International Communication (TOEIC) —use and measure orthographic knowledge and phonological awareness required by the Faculty of Nursing. As such, the chapter first describes the materials and processes of assessment and second, reports the results of the assessment.

4.1 Materials

Materials and rationales chosen for RQ 2 appear below. Three materials were chosen for analysis: the official guide to understanding the test scores, the official collection of practice test items, and the results of Chapter 3's analysis of the Faculty of Nursing's English-language curriculum, each of which includes titles of the references.

The official guide to understanding the test scores:

- *A Guide to Understanding TOEFL iBT Scores*
- *TOEFL Monograph Series: TOEFL 2000 Reading Framework: A Working Paper*
- *TOEIC Reading Score Descriptors*
- *TOEIC Can-Do Guide Executive Summary Listening & Reading*

The official collection of practice test items:

- *TOEFL Official Guide to the TOEFL Test with CD-ROM, 4th Edition (2012)*
- *TOEIC Shin Koshiki Mondaishu Vol 5. [New Official Collection of Past Questions] (2012)*

The results of Chapter 3's analysis of the Faculty of Nursing's English-language curriculum:

- The results of Chapter 3's analysis of the syllabi of four English courses provided at the Faculty of Nursing
- The results of Chapter 3's analysis of the medical vocabulary list containing 481 words from the reference book for the Faculty of Nursing

Both the TOEFL and the TOEIC tests are designed by English Testing Services (ETS) which offers several guides to their existing and future test takers, education professionals, and media. This is so that not only can the test takers prepare for the tests but also so that people interested in the tests can understand the aims and purposes of the tests and receive feedback on their performance. Moreover, in order to assess students' English-language proficiency levels accurately, ETS provides the ways they have invented their tasks and questions. Therefore, the writer can compare the test items with the explanations of the test items and assess whether or not the test items are actually assessing what the test designers claim they assess; in this case, the construct of each test item. Moreover, to explore whether TOEFL and TOEIC test items reflect the content of the faculty's curriculum, the results of Chapter 3's analysis of the syllabus, the textbook, and the medical vocabulary list of 481 words from a reference book were included as material in this research.

4.2 Research process

This section describes the ways in which the research was done on official guides, official collections of practice test items and the results of Chapter 3's analysis on the faculty's curriculum. For the assessment of the official guide to understanding the test scores of the TOEFL and the TOEIC test, the writer reads the whole description in order to detect whether test questions and reading passages are designed to assess test taker's orthographic knowledge and phonological awareness. With regards to the official collection of practice test items, one actual full-length test of the reading section is chosen as research targets from each official collection. The reading section of a TOEFL test has three reading passages and 14 questions and they are categorized into eleven categories as Table 4.1 shows below:

Table 4.1

TOEFL Reading Questions Types (The Official Guide to the TOEFL Test, 4th Edition, *2012a)*

		Number of questions per set
(1)	Factual information questions	3 to 6
(2)	Negative factual information questions	0 to 2
(3)	Inference questions	1to 3
(4)	Rhetorical purposes questions	1 to 2
(5)	Vocabulary questions	3 to 5
(6)	Reference questions	0 to 2
(7)	Sentence simplification questions	0 or 1
(8)	Insert text questions	1
(9	Prose summary questions	1
(10)	Fill in a questions	1

After each set of practical test, there is an answer explanations' section where explanations on types of questions, right answers, and reasons are given. An example of TOEFL reading section's passage and question in *The Official Guide to the TOEFL Test 4th Edition* (2012a) is given below:

> Paragraph 1
> Architecture is the art and science of designing structures that organise and enclose space for practical and symbolic purposes. Because architecture grows out of human needs and aspirations, it clearly communicates cultural values. Of all the visual arts, architecture affects our lives most directly for it determines the character of the human environment in major ways.
>
> 1. According to paragraph 1, all the following statements about architecture are true EXCEPT:
> Architecture is a visual art.
> Architecture reflects the cultural values of its creators.
> Architecture has both artistic and scientific dimensions.
> Architecture has an indirect effect on life.
> (p.463)

According to the answer explanation, this is a negative factual information question because sentence three in the first paragraph states "Of all the visual arts, architecture affects our lives most directly for it determines the character of the human environment in major ways," which contradicts with choice 4's "architecture has an indirect effect on life."

TOEIC's reading section consists of 40 fill-in-the-blank questions and 60 comprehension questions based on 20 reading passages, for a total of which makes 100 test questions. The writer reads the passage, test questions, and explanations of test types if available, and then judges whether the test item assesses orthographic knowledge or phonological awareness.

The writer first affirms whether the explanations of the official collection correspond to the writer's categorization. When a disagreement is found between the categorization of the writer and the explanation of the official collection of practice tests regarding the type of skills and constructs the test item is measuring, the writer invites another evaluator to assess the test items. If the perspective of the second evaluator corresponds with the categorization of the official collection, then the official collection's categorization is adopted. Then, the writer will assess whether the vocabulary test questions measure orthographic knowledge or phonological awareness.

Unlike TOEFL's official practice test collection, the TOEIC's official practice test collection does not provide explanations about the type of skills the item is measuring. The reading section of the TOEIC test is composed of three parts: fill-in-the-blank of short sentence questions, fill-in-the-blank of short sentences within a text, and reading comprehension that includes 1-2 passages and 2-4 corresponding questions. Although the official collection states that part three aims to test reading comprehension skills, there are test items that appear to assess the test taker's vocabulary and grammar skills. (For example, "The word 'noted' in paragraph 4, line 8, is closest in meaning to (A) indicated, (B) well-known, (C) observed, or (D) knowledgeable.) Therefore, all the items in the reading section of TOEIC are the targets of the study. Of these 100 test items, the writer will first categorize all the test items into three categories: (1) grammar, (2) vocabulary, and (3) comprehension. The writer will perform a second categorization one week after the first categorization. If the two categorizations performed by the writer show inconsistencies, the writer will invite a second evaluator to categorize the test items. When the second evaluator has completed his or her categorization, the writer and the second evaluator will discuss the results and make final decisions. If it is not possible to reach an agreement, the writer will omit the test item from the research target. Then, the writer will assess whether the vocabulary test questions measure orthographic knowledge or phonological awareness.

Moreover, in order to compare the results of Chapter 3's analysis on identifying the orthographic and phonological features of the medical vocabulary list of 481 words from the reference book, the following procedure has been adopted:

- The author transcribes all words of the TOEIC and TOEFL in alphabetical order in an Excel file
- Words were copied and pasted to the VocabProfiles of Compleat Lexical Tutor, a computer program that categorizes English words into four categories by frequencies: the 1,000 most frequently used words (K1), the second 1,000 most frequently used words (K2), the 570 most frequently used academic words (AWL), and unlisted words (OFF).
- The phonetic symbol of each word was checked using *Taishukan's Unabridged Genius English-Japanese Dictionary*. When the phonetic symbol of the word was not included in the *Genius*, the *Shogakukan Random House English-Japanese Dictionary 2nd edition* was used.
- Phonetic symbols were separated into different rows of the Excel file based on the syllables.
- The author checked whether the syllable ended with a consonant and counted the number of occurrences.
- Each phoneme of vowels was matched with corresponding graphemes.
- All graphemes of syllables were rearranged by alphabetical order
- All graphemes of syllables were divided into three categories: one, phonetic symbol that has same grapheme as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese

one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese (NNE)

An example of the vocabulary list with phonetic symbols is included in Table 4.2 and 4.3.

Table 4.2

An Example of the Vocabulary List of TOEFL Sample Test

		Word	Frequency	phonetic symbol	Number of syllables	1st syllable	2nd	3rd	4th	5th	6th
1.	K1	a	119	eɪ	1	a=eɪ					
2.	K1	able	7	ˈeɪb.l̩	2	a=eɪ	*=ə				
3.	K1	about	7	ə.ˈbaʊt	2	a=ə	ou=au				
4.	K1	above	2	ə.ˈbʌv	2	a=ə	o=ʌ				
5.	K1	accord	3	əˈk.ɔːd	2	a=ə	or=ɔ:r				
6.	K1	act	7	ækt |	1	a=æ					
7.	K1	active	12	ˈæk.tɪv	2	a=æ	i=ɪ				
8.	K1	actual	2	ˈæk.tʃuəl	2	a=æ	ua=uə				
9.	K1	add	2	æd	1	a=æ					
10.	K1	admit	1	əd.ˈmɪt	2	a=ə	i=ɪ				
11.	K1	adopt	4	ə.ˈdɒpt	2	a=ə	o=a				
12.	K1	advantage	1	əd.ˈvɑːn.tɪdʒ	3	a=æ	a=æ	a=ɪ			
13.	K1	after	4	ˈɑːf.tə	2	a=æ	er=ər				
14.	K1	again	2	ə.ˈgen	2	a=ə	ai=e				
15.	K1	age	1	eɪdʒ	1	a=eɪ					
16.	K1	agent	4	ˈeɪ.dʒənt	2	a=eɪ	e=ə				
17.	K1	ago	4	ə.ˈgəʊ	2	a=ə	o=ou				

Table 4.3

An Example of the Vocabulary List of the TOEIC Sample Test

		Word	Frequency	Phonetic symbol	Number of syllables	1st syllable	2nd	3rd	4th	5th
1.	K1	a	77	eɪ	1	a=eɪ				
2.	K1	able	3	ˈeɪbl	1	a=ei				
3.	K1	about	5	əˈbaʊt	2	a=ə	ou=au			
4.	K1	above	2	əˈbʌv	2	a=ə	o=ʌ			
5.	K1	accept	1	ækˈsɛpt	2	a=æ	e=e			
6.	K1	accord	1	əˈkɔrd	2	a=ə	or=ɔ:			
7.	K1	account	1	əˈkaʊnt	2	a=ə	ou=au			
8.	K1	act	1	ækt	1	a=æ				
9.	K1	active	5	ˈæktɪv	2	a=æ	i=ɪ			
10.	K1	actual	2	ˈækʧuəl	3	a=æ	u=u	a=ə		
11.	K1	add	5	æd	1	a=æ				
12.	K1	address	1	ˈæˌdrɛs	2	a=æ	e=e			
13.	K1	admit	1	ədˈmɪt	2	a=ə	i=ɪ			
14.	K1	advance	2	ədˈvæns	2	a=ə	a=æ			
15.	K1	after	2	ˈæftər	2	a=æ	er=ər			
16.	K1	again	1	əˈgɛn	2	a=ə	ai=eɪ			
17.	K1	ago	2	əˈgoʊ	2	a=ə	o=ou			

4.3 Results

In this section, the paper first describes the results of the analysis on the official guide to understanding test scores, after which it examines the orthographical and phonological features of the TOEFL and the TOEIC. Lastly, the section compares the results of Chapter 3’s analysis of the faculty’s curriculum with the official guide’s test scores as well as the results of Chapter 3’s analysis of the orthographical and phonological features of the medical vocabulary list of 481 words from the reference book with those on the TOEFL and the TOEIC.

4.3.1 The official guide

As described in Table 4.4, TOEFL aims to measure the test taker’s ability to read academic texts. The subcategories of the “academic reading abilities” appear to be vocabulary, grammar, inference, synthesis, identification of an expository structure of a text, and ability to retrieve a main idea from a text. With regards to vocabulary, the official guide provides advice for improvement for each level of test takers. The advice for low-level test takers appears to indicate acknowledgement of orthographic knowledge and phonological awareness. It states “Increase your vocabulary by analysing word parts; study roots, prefixes and suffixes; study word families.” (ETS, 2008) The analysis of word parts, including prefixes and suffixes, will eventually lead to increased knowledge of the systematic order of spelling.

Table 4.4

A Guide to Understanding TOEFL iBT^{R} Scores (ETS, 2008)

Level	High (22-30)	Intermediate (15-21)	Low (0-14)
Your Performance	Test takers who receive a score at the **HIGH** level typically understand academic text in English that require a wide range of reading abilities regardless of the difficulty of the texts. Test takers who score at the **HIGH** level typically: · Have a very good command of academic vocabulary and grammatical structure	Test takers who receive a score at the **INTERMEDIATE** level typically understand academic texts in English that require a wide range of reading abilities, although their understanding of certain parts of the texts is limited. Test takers who receive a score at the **INTERMEDIATE** level typically: · have a good command of common academic vocabulary, but still have some difficulty with high-level vocabulary	Test takers who receive a score at the **LOW** level typically understand some of the information presented in academic texts in English that require a wide range of reading abilities, but their understanding is limited. Test takers who receive a score at the **LOW** level typically: · have a command of basic academic vocabulary, but their understanding of less common vocabulary is inconsistent
	· Can understand and connect information, make appropriate inferences and synthesize ideas, even when the text is conceptually dense and the language is complex · Can recognize the expository organization of a text and the role that specific information serves within the larger text, even when there is conceptually dense · Can abstract major ideas from a text, even when the text is conceptually dense and contains complex language	• have a very good understanding of grammatical structure • can understand and connect information, make appropriate inferences, and synthesize information in a range of texts, but have more difficulty when the vocabulary is high level and the text is conceptually dense • can recognize the expository organization of a text and the role that specific information serves within a larger text, but have some difficulty when these are not explicit or easy to infer from the text • can abstract major ideas from a text, but have more difficulty doing so when the text is conceptually dense	• have limited ability to understand and connect information, have difficulty recognizing paraphrases of text information, and often rely on particular words and phrases rather than a complete understanding of the text • have difficulty identifying the author's purpose, except when that purpose is explicitly stated in the text or easy to infer from the text • can sometimes recognize major ideas from a text when the information is clearly presented, memorable or illustrated by examples, but have difficulty doing so when the text is more demanding

Advice for improvement	Read as much and as often as possible. Make sure to include academic texts on a variety of topics written in different genres and with different degrees of conceptual density as part of your reading. • Read major newspapers, such as *The New York Times* or *Science Time*s, and websites (National Public Radio [NPR] or the BBC). • Write summaries of texts, making sure they incorporate the organizational pattern of the originals. Continually expand your vocabulary. Continually practice using new words you encounter in your reading. This will help you remember both the meaning and correct usage of the new words.	Read as much and as often as possible. Study the organization of academic texts and overall structure of reading passages. Read an entire passage from beginning to end. • Pay attention to the relationship between the main ideas and the supporting details. • Outline the text to test your understanding of the structure of the reading passage. • Write a summary of the entire passage. • If the text is a comparison, be sure that your summary reflects that. If the text argues two points of view, be sure both points of view are reflected in your summary. Continually expand your vocabulary by developing a system for recording unfamiliar words. • Group words according to topic or meaning and study the words as a list of related words. • Study roots, prefixes and suffixes; study word families. • • Use available vocabulary resources, such as a good thesaurus or a dictionary of collocations (words commonly used together).	Read as much and as often as possible. Develop a system for recording unfamiliar words. • Group words into lists according to topic or meaning and review and study the words on a regular basis so that you remember them. • Increase your vocabulary by analyzing word parts; study roots, prefixes and suffixes; study word families. Study the organization of academic texts and overall structure of a reading passage. Read an entire passage from beginning to end. • Look at connections between sentences; look at how the end of one sentence relates to the beginning of the next sentence. • Look for the main ideas and supporting details and pay attention to the relationship between them. • Outline a text to test your understanding of the structure of a reading passage. • Begin by grouping paragraphs that address the same concept. • Write one sentence summarizing the paragraphs that discuss the same idea. • Write a summary of the entire passage.

In *TOEFL Monograph Series: TOEFL 2000 Reading Framework: A Working Paper*, ETS (2000) argues that the construct the TOEFL test measures is “a single broad construct that includes the four academic reading purposes (p. 4).” The four purposes are: reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information across multiple texts. The paper states that all the reading “requires a combination of word recognition/processing efficiency and comprehension abilities.” Although the TOEFL official guide does not explicitly include statements that directly lead to a measurement of orthographic knowledge and phonological awareness, there is a need to identify whether practice test items are measuring “word recognition/processing efficiency.”

In *TOEIC Reading Score Descriptors* and *TOEIC Can-Do Guide Executive Summary Listening & Reading* (2008*)*, ETS claims that TOEIC not only provides test takers with a score report but also a score descriptor and can-do guide available on the test’s homepage, to facilitate the test taker’s interpretation of his or her score. Each test taker receives a score report that includes information on abilities measured on the test. For example, ETS maintains that the reading section of the TOEIC measures the ability to: infer based on information in written texts, locate and understand specific information in written texts, connect information across multiple sentences in a single written text and across texts, and understand vocabulary and grammar in written texts.

Table 4.5

TOEIC Reading Score Descriptors (ETS, 2014)

Level	Strengths	Weaknesses
450	Test takers who score around 450 typically have the following strengths: • They can infer the central idea and purpose of a written text, and they can make inferences about details. • They can read for meaning. They can understand factual information, even when it is paraphrased. • They can connect information across an entire text, and they can make connections between two related texts. • They can understand a broad range of vocabulary, unusual meanings of common words, and idiomatic usage. They can also make distinctions between the meanings of closely related words. • They can understand rule-based grammatical structures. They can also understand difficult, complex, and uncommon grammatical constructions.	Test takers who score around 450 typically have weaknesses only when the information tested is particularly dense or involves difficult vocabulary.
350	Test takers who score around 350 typically have the following strengths: • They can infer the central idea and purpose of a written text, and they can make inferences about details. • They can read for meaning. They can understand factual information, even when it is paraphrased. • They can connect information across a small area within a text, even when the vocabulary and grammar of the text are difficult. • They can understand medium-level vocabulary. They can sometimes understand difficult vocabulary in context, unusual meanings of common words, and idiomatic usage. • They can understand rule-based grammatical structures. They can also understand difficult, complex, and uncommon grammatical constructions.	Test takers who score around 350 typically have the following weaknesses: • They do not connect information across a wide area within a text. • They do not consistently understand difficult vocabulary, unusual meanings of common words, or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words.

250	Test takers who score around 250 typically have the following strengths: • They can make simple inferences based on a limited amount of text. • They can locate the correct answer to a factual question when the language of the text matches the information that is required. They can sometimes answer a factual question when the answer is a simple paraphrase of the information in the text. They can sometimes connect information within one or two sentences. • They can understand easy vocabulary, and they can sometimes understand medium-level vocabulary. • They can understand common, rule-based grammatical structures. They can make correct grammatical choices, even when other features of language, such as difficult vocabulary or the need to connect information, are present.	Test takers who score around 250 typically have the following weaknesses: • They do not understand inferences that require paraphrase or connecting information. • They have a very limited ability to understand factual information expressed as a paraphrase using difficult vocabulary. They often depend on finding words and phrases in the text that match the same words and phrases in the question. • They usually do not connect information beyond two sentences. • They do not understand difficult vocabulary, unusual meanings of common words, or idiomatic usage. They usually cannot make distinctions between the meanings of closely related words. • They do not understand more-difficult, complex, or uncommon grammatical constructions.
150	Test takers who score around 150 typically have the following strengths: • They can locate the correct answer to a factual question when not very much reading is necessary and when the language of the text matches the information that is required. • They can understand easy vocabulary and common phrases. • They can understand the most-common, rule-based grammatical constructions when not very much reading is necessary.	Test takers who score around 150 typically have the following weaknesses: • They cannot make inferences about information in written texts. • They do not understand paraphrased factual information. They rely on matching words and phrases in the text to answer questions. • They are often unable to connect information even within a single sentence. • They understand only a limited range of vocabulary. • They do not understand even easy grammatical constructions when other language features, such as difficult vocabulary.

Table 4.6

Percentages of TOEIC Test Takers, by Reading Score Level, Who Indicated that They could Perform Various English-language Reading Tasks Either Easily or with Little Difficulty (ETS, 2008)

I can:	5-135	140-195	200-255	260-315	320-375	380-435	440-495	M	SD	Corr. with TOEIC reading scaled score
Read the letters of the alphabet	91	95	96	95	96	97	99	4.81	0.61	.08
Read and understand a restaurant menu	65	72	79	83	86	87	95	4.22	0.88	.23
Recognize memorized words and phrases (e.g., “Exit,” “Entrance,” and “Stop”)	63	72	78	82	87	92	97	4.16	0.84	.27
Read and understand a train or bus schedule	49	59	70	77	84	90	96	4.00	0.91	.34
Read, on storefronts, the type of store or services provided (e.g., “dry cleaning,” “book store”)	47	64	69	72	81	90	91	3.95	0.95	.31
Read and understand a simple postcard from a friend	43	58	65	75	83	90	97	3.94	0.92	.37
Read office memoranda in which the writer has used simple words or sentences	36	50	61	72	81	88	96	3.83	0.92	.39
Read and understand traffic signs	40	51	61	68	77	86	90	3.81	0.98	.33
Read s, graphs, and charts	31	40	54	64	73	83	93	3.69	0.94	.38
Read and understand directions and explanations presented in technical manuals written for beginning users	26	34	46	58	66	78	87	3.56	0.97	.40
Read and understand simple, step-by-step instructions (e.g., how to operate a copy machine)	24	34	45	55	64	79	90	3.52	0.97	.39
Find information that I need in a telephone directory	23	34	42	52	64	76	89	3.48	1.00	.39
Read and understand a letter of thanks from a client or customer	18	26	39	53	66	81	94	3.45	0.97	.47
Read entertainment-related information (e.g., tourist guides)	15	25	32	45	57	72	85	3.34	0.97	.41
Read information about products (e.g., advertisements)	14	22	29	40	52	68	88	3.27	0.98	.42
Read and understand a travel brochure	10	18	26	38	51	68	86	3.22	0.98	.44
Read and understand an agenda for a meeting	6	14	22	34	46	62	84	3.09	1.00	.48
Read and understand the main points of an article on a familiar topic in an academic or professional journal	10	17	23	30	40	53	79	3.07	0.96	.37

Read English to translate text into my own language (e.g., letters and business documents)	5	12	16	23	36	50	74	2.92	1.01	.39
Read and understand a popular novel	7	10	15	23	31	43	67	2.91	0.92	.40
Identify inconsistencies or differences in points of view in two newspaper interviews with politicians of opposing parties	7	8	13	20	30	43	69	2.82	0.97	.43
Read highly technical material in my field or area of expertise with little use of a dictionary	5	10	14	19	27	40	59	2.76	1.01	0.38
Read a newspaper editorial and understand its meaning as well as the writer's intent	6	7	10	17	25	35	57	2.71	0.95	0.41
Read and understand a proposal or contract from a client	4	7	11	17	25	42	58	2.68	1.01	0.44
Read and understand magazine articles like those found in *Time* or *Newsweek*, without using a dictionary	3	5	5	11	19	30	47	2.6	0.91	0.42

While score descriptors provide information concerning test takers' strengths and weaknesses and recommend practices for improvement, the can-do guide provides information regarding the kinds of tasks that each level of test taker can achieve with ease or difficulty. As Table 4.6 shows, most tasks illustrated are connected to real-world activities such as "read[ing] information about products" or "read[ing] and understand[ing] the main points of an article on a familiar topic in an academic or professional journal." As a whole, however, neither the score descriptor nor the can-do guide provides test items measure the orthographic or phonological awareness leading to reading fluency or the increase of medical vocabulary, both of which are primary requirements of the faculty's curriculum.

As previously described, a reading section from an actual full-length TOEFL test usually includes three or four reading passages with 12-14 questions each. The length of each passage is approximately 700 words. The types of test items included in the test are shown in Table 4.7 below.

Table 4.7

Types of Test Items Included in the Reading Section of TOEFL

	Number of questions		
	Passage 1	Passage 2	Passage 3
Factual information questions	4	4	3
Negative factual information questions	1	0	0
Inference questions	1	0	1
Rhetorical purposes questions	1	1	1
Vocabulary questions	3	4	3
Reference questions	0	1	1
Sentence simplification questions	1	1	1
Insert text questions	1	1	1
Prose summary questions	1	1	0
Fill in a questions	0	0	1
Total	13	13	12

Vocabulary questions ask readers to choose "the closest meaning" of a word or phrase. In either case, the test asks the readers to choose the meaning "as it is used in the passage." None of the vocabulary test items, however, assess the orthographic knowledge or phonological awareness of the test takers. All of the questions require the test takers to choose the closest meanings of the suggested words or phrases from four choices.

As described in the previous section, the TOEIC's reading section is composed of three parts: (1) 40 fill-in-the-blank short sentences, (2) 12 fill-in-the-blank short sentences in texts, and (3) 48 reading comprehension test items. The types of test questions in the TOEIC's reading sections are as follows:

Table 4.8

Types of Test Questions in the TOEIC's Reading Section

	Part 1	Part 2	Part 3
Grammar	30	6	0
Vocabulary	10	6	3
Reading comprehension	0	0	45
Total	40	12	48

As can be seen from the Table 4.8, 12 % of the test items appear to assess the test taker's vocabulary skills. None of the vocabulary test items, however, assess the orthographic knowledge or phonological awareness of the test takers. All of the questions require the test takers to choose the closest meanings of the suggested words or phrases from four choices.

4.3.2 Orthographic and phonological features of the TOEFL and the TOEIC

With regards to the orthographic and phonological features of the TOEFL and the TOEIC vocabulary, the TOEFL has approximately 9721 words in sample test's reading section. Of these 9721 words, 6221 (63.4%) are categorized as the 1,000 most-frequently used words (1K), 718 (7.38%) as the second 2,000 most-frequently used words (K2), 1125 (11.57%) as the 570 most frequent academic words (AWL), and 1657 (17.0%) as unlisted words (OFF). Based on the ratio of authentic English written text, which is K1 (70%), K2 (10%), AWL (10%), and OFF (10%), the number of OFF words in the TOEFL is slightly higher than in authentic text. This indicates that TOEFL reading section requires the test takers to possess a higher vocabulary knowledge level.

The TOEIC has approximately 7383 words in a sample test's reading section. Of these 7383 words, 5169 (70%) are categorized as the 1,000 most-frequently used words (1K), 567 (7.7%) as the second 2,000 most-frequently used words (K2), 718

(9.7%) as the 570 most frequent academic words (AWL), and 929 (12.6%) as unlisted words (OFF). Based on the ratio of authentic English written text, which is K1 (70%), K2 (10%), AWL (10%), and OFF (10%), the TOEIC reading section requires test takers to possess a moderate level of vocabulary knowledge. The average number of syllables in a word of the TOEIC is 1.96.

4.3.3 Syllables

The average number of syllables per word on the TOEFL is 1.94. Figure 4.1 shows the number of syllables in K1, K2, AWL, and OFF words. The majority of K1 and K2 words have one or two syllables, while most AWL and OFF words have more than two syllables.

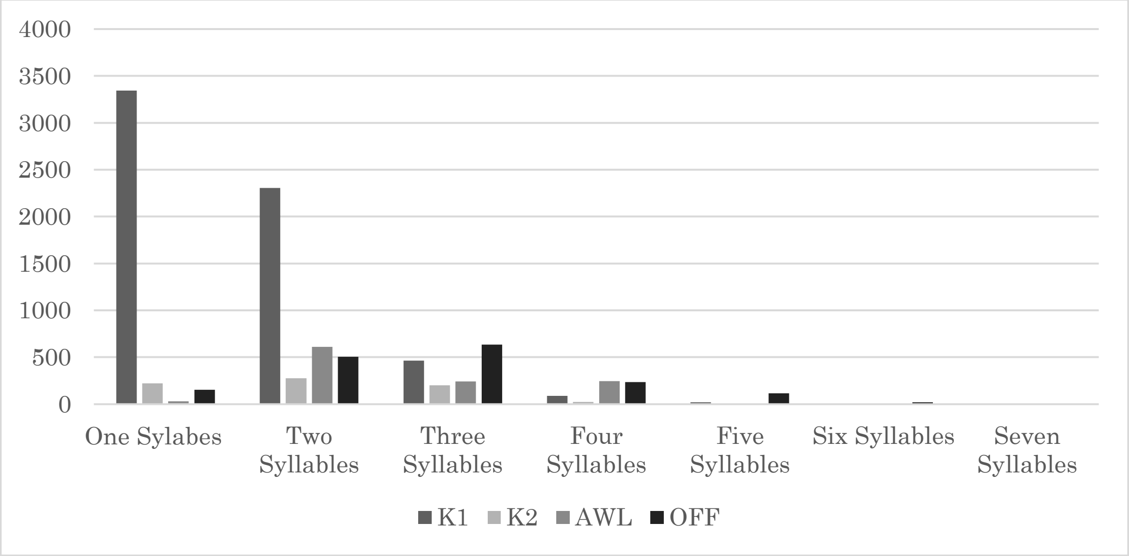


Figure 4.1. The number of syllables in K1, K2, AWL, OFF words on the TOEFL.

Figure 4.2 shows the number of syllables in K1, K2, AWL, and OFF words on the TOEIC sample test. Compared to TOEFL's syllables, the TOEIC has more two syllables words. The majority of K1 and K2 words have one or two syllables while most AWL and OFF words have more than two syllables.

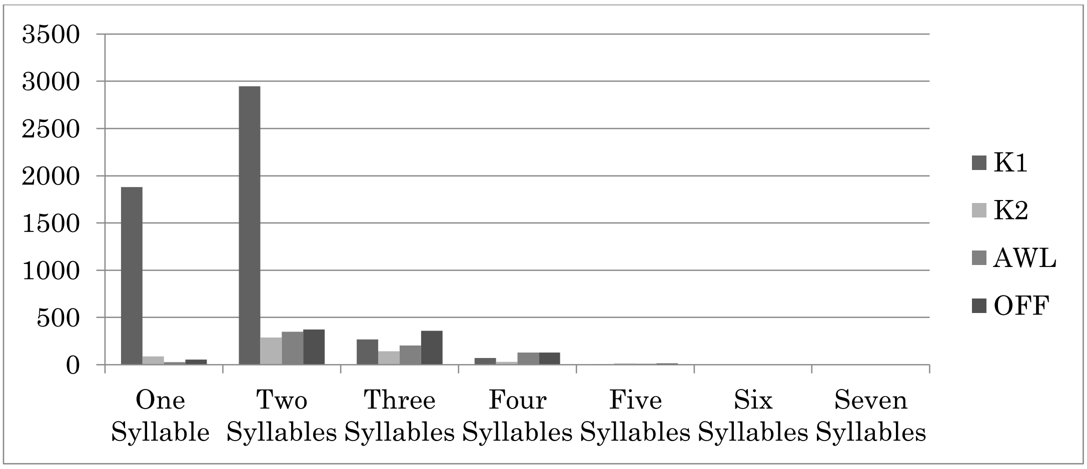


Figure 4.2. Number of syllables per word on the TOEIC.

4.3.4 Graphemes

Regarding the graphemes /a/, /e/, /i/, /o/, and /u/ on the TOEFL and the TOEIC reading section, the percentages of the grapheme /u/ exhibit similarities. The percentage of /e/ differs, however, as that of the TOEIC (47%) exceeds that of the TOEFL (31%).

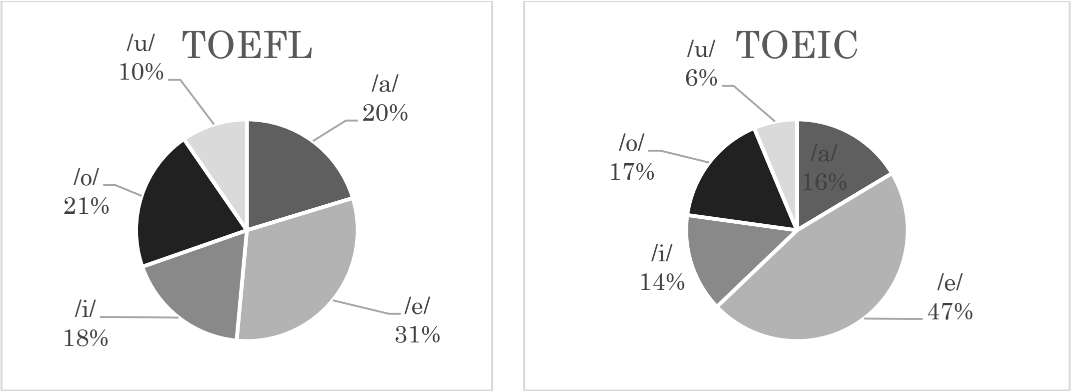


Figure 4.3 Percentage of graphemes /a/, /e/, /i/, /o/, and /u/ in the TOEFL and the TOEIC reading sections.

4.3.5 Comparison with L1 grapheme–phoneme correspondence

As described in chapter two, the author has divided the vowels of English phonemes into three categories: one, phonetic symbols that have same graphemes as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbol exists in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbol does not exist in Japanese (NNE). This categorization is used to identify the orthographic and phonological features of the two tests. For the grapheme /a/, the phonetic symbol [a] will be categorized as JE; [a:], [ai], [e], [i], and [o] are categorized as NE; and [ə], [ɔ(:)], [æ], and [ʌ] are categorized as NNE. Figure 4.4 shows the percentage of JE, NE, and NNE for the grapheme /a/ in the TOEFL and the TOEIC reading sections. The result indicates that the grapheme /a/ is rarely pronounced in the way the Japanese /a/ is pronounced. Moreover, the number of NNE phonemes exceeds not only that of JE but also NE. The reading of grapheme /a/, therefore, is deemed more challenging for Japanese learners of English.

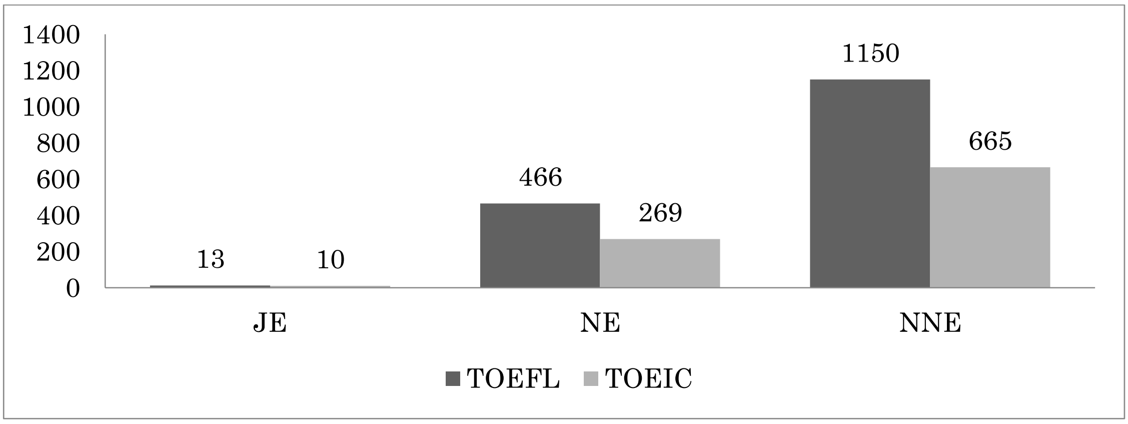


Figure 4.4. The number of JE, NE, and NNE phonemes of the grapheme /a/ in the TOEFL and the TOEIC reading sections.

Figure 4.5 shows the number of JE, NE, and NNE phonemes of the grapheme /e/ in the TOEFL and the TOEIC reading sections. With regards to the phonemes of the grapheme /e/, [e] is categorized as JE while [a] and [i:] as NE. The phonetic symbols [ɪ], [eɪ], [φ] are classified as NNE. Unlike the grapheme /a/, more than a quarter of the appearances of the grapheme /e/ on both the TOEFL and the TOEIC follow the Japanese-language's one-to-one correspondence rule, in this case, the phoneme [e]. Moreover, more than half of the graphemes are pronounced with the sound that exists in Japanese pronunciation, such as, [ei], [u], and [i:]. This result indicates that the pronunciation of the grapheme /e/ in English can be less demanding for Japanese learners of English.

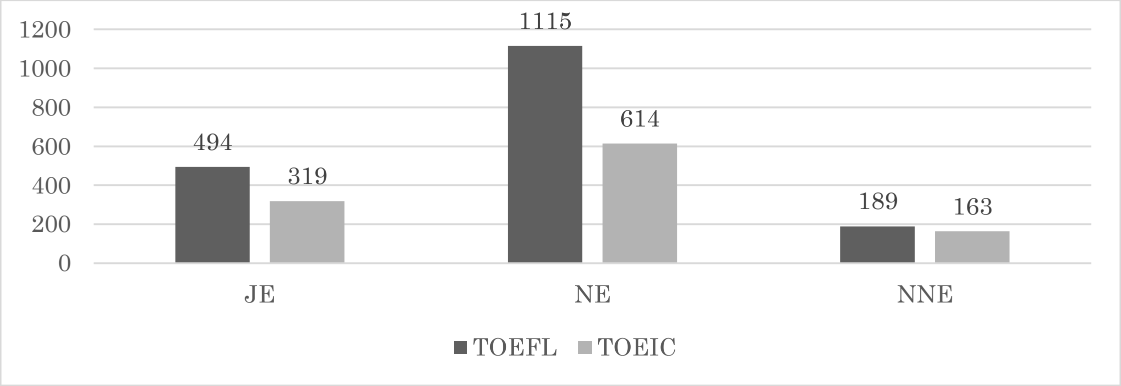


Figure 4.5. The number of JE, NE, and NNE phonemes of a grapheme /e/ in the TOEFL and the TOEIC reading sections.

With regards to the pronunciation of the grapheme /i/, [i] is categorized as JE; [i:], [ai], and [a:] are labeled as NE; and [ə], [ər], [iə], and [ju:] are sorted as NNE. The result indicates that majority of the occurrences the grapheme /i/ are pronounced in the way Japanese learners of English pronounce the grapheme /i/, in this case [i].

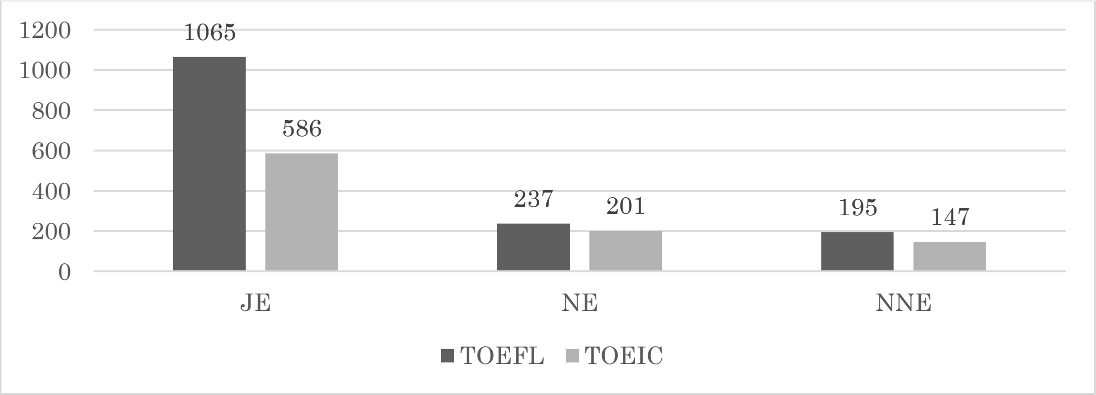


Figure 4.6. The number of JE, NE, and NNE phonemes of the grapheme /i/ in the TOEFL and the TOEIC reading sections.

Figure 4.7 shows the number of JE, NE, and NNE phonemes of /o/ in the TOEFL and TOEIC reading section. Regarding the categorization of the grapheme /o/, [ɔ] is categorized as JE; [a], [a:], [ai], [au], [i], [ou], [u], and [u:] are classified as NE, and [ə], [əu], [jəu], and [ʌ] as NNE. Reading of the grapheme /o/ will be challenging for Japanese TOEFL readers, since most occurrences of the grapheme /o/ are read differently from the Japanese pronunciation of the grapheme /o/.

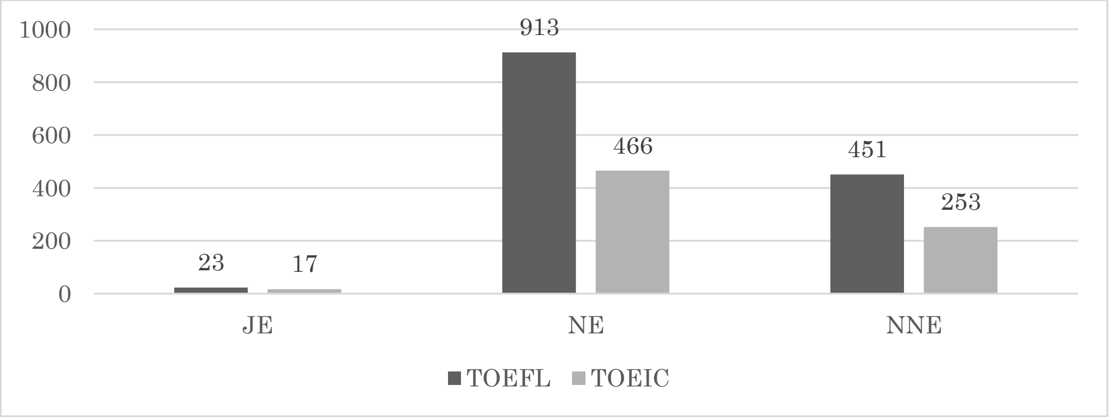


Figure 4.7. The number of JE, NE, and NNE phonemes of the grapheme /o/ in the TOEFL and the TOEIC reading sections.

The pronunciation of the grapheme /u/, although infrequently used (see Figure 4.8), will be one of the most challenging tasks for Japanese learners of English not only because the grapheme /u/ is rarely pronounced [u] in English, but also because most of its pronunciations do not exist in Japanese pronunciation. In this categorization, [u] is labeled as JE; [i], [ou], and [u:] as NE; and [ə], [jə], [ju], [ju:], [juə], [ʌ], [yu], and [yu:] as NNE.

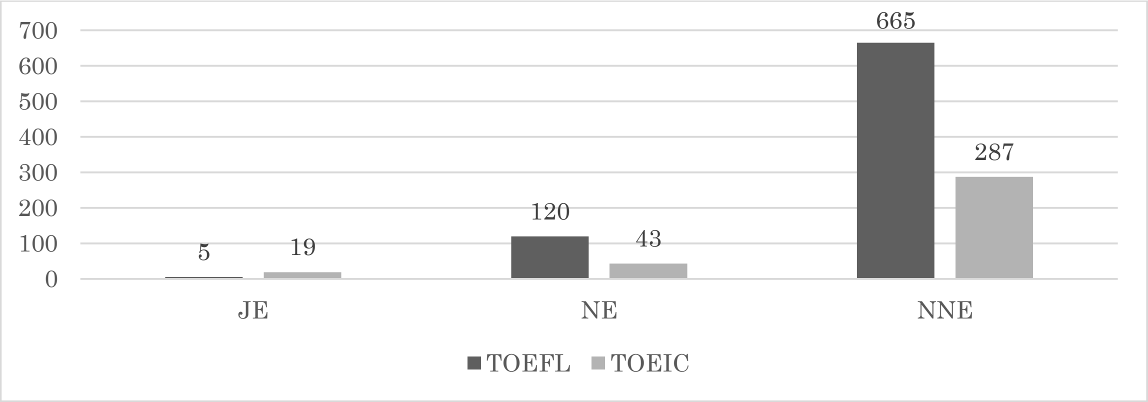
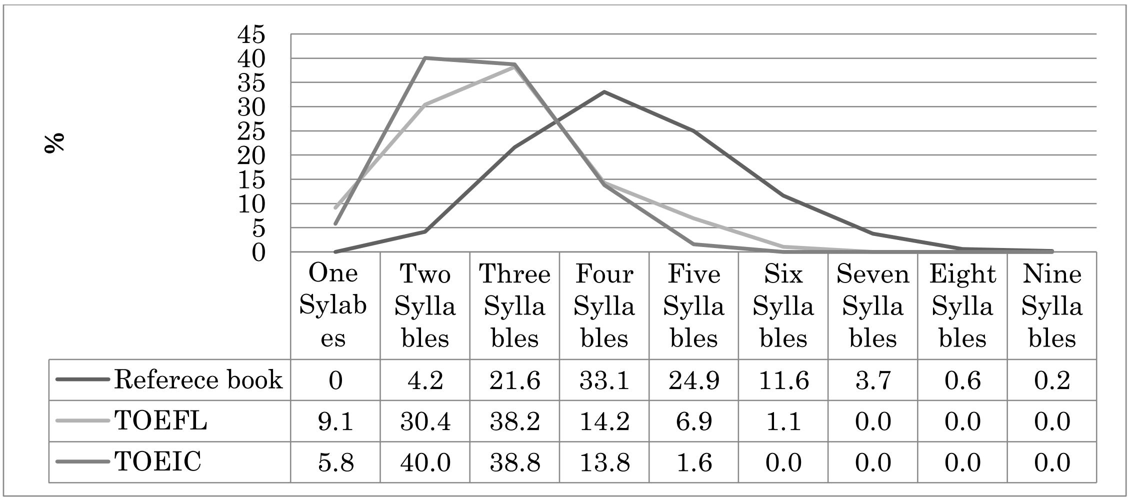


Figure 4.8. The number of JE, NE, and NNE phonemes of the grapheme /u/ in the TOEFL and TOEIC reading sections.

4.3.6 Differences between the reference book's vocabulary, the TOEFL's and the TOEIC's

In terms of the differences between the number of syllables per word in the reference books vocabulary list (4.3 syllables/ word), on the TOEFL (1.94) and on the TOEIC (1.96), Figure 4.9 shows that the average number of syllables per word was the largest in the reference book. More than 70% of words in the reference book have more than four syllables per word, while nearly 80% of words have fewer than three syllables on the TOEFL and more than 80% of words have fewer than three syllables on the TOEIC. The result indicates that the medical terms have more syllables than the vocabulary on the TOEFL and the TOEIC.



	One Sylab es	Two Sylla bles	Three Sylla bles	Four Sylla bles	Five Sylla bles	Six Sylla bles	Seven Sylla bles	Eight Sylla bles	Nine Sylla bles
— Referece book	0	4.2	21.6	33.1	24.9	11.6	3.7	0.6	0.2
— TOEFL	9.1	30.4	38.2	14.2	6.9	1.1	0.0	0.0	0.0
— TOEIC	5.8	40.0	38.8	13.8	1.6	0.0	0.0	0.0	0.0

Figure 4.9. *Number of syllables per word in the reference books' vocabulary list, on the TOEIC and on the TOEFL.*

Regarding the difference between the medical terms and those found in the two commercially produced tests of grapheme–phoneme correspondences, Figure 4.10 shows that reading the graphemes /e/, /o/, and /u/ is very challenging for Japanese learners of English. This is because the medical terms have more graphemes that are classified as NNE than the TOEFL and the TOEIC. In other words, the number of medical terms' graphemes that are not pronounced in the way that Japanese graphemes are pronounced or whose pronunciation does not exist in Japanese is larger than what is represented in the TOEFL and the TOEIC. The result indicates that the use of the TOEFL and the TOEIC as placement tests might not accurately assess the orthographic and phonological processing skills needed for students to succeed in the Faculty of Nursing.

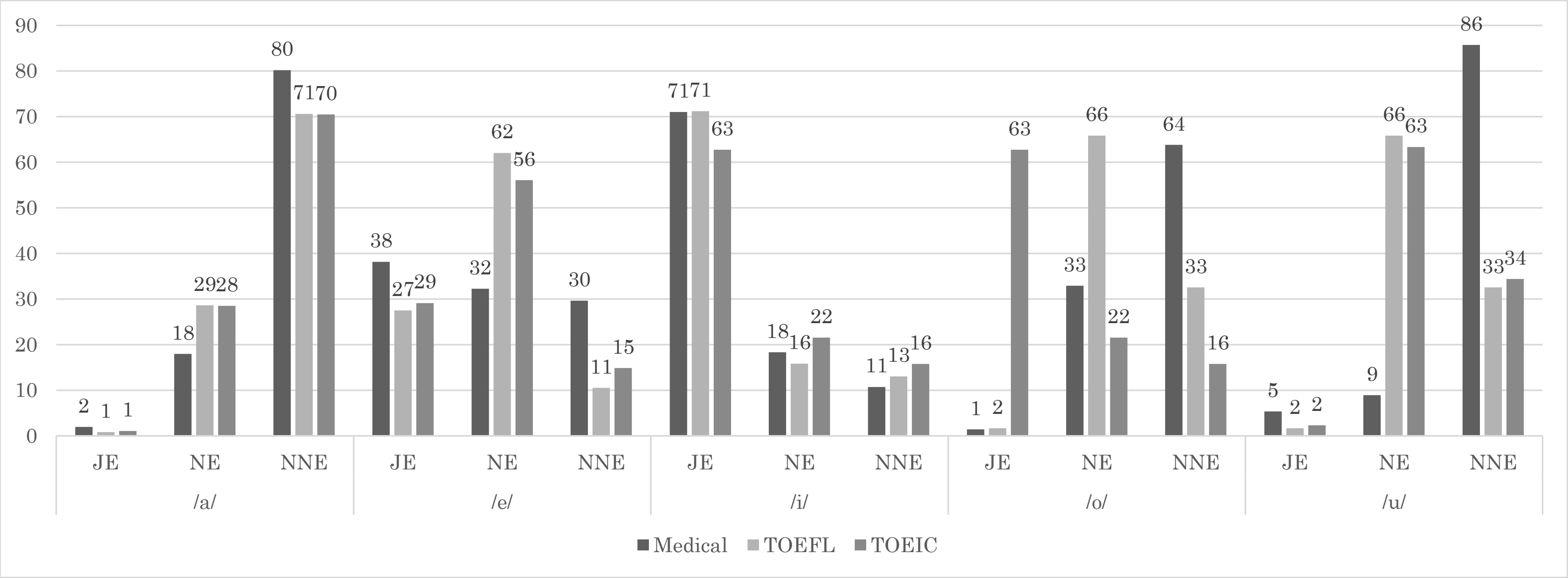


Figure 4.10. The percentage of JE, NE, and NNE phonemes of graphemes /a/, /e/, /i/, /o/, and /u/ in the English medical terms of the reference book, the TOEFL, and the TOEIC.

4.4. Summary findings and conclusion

As described above, the earlier part of this chapter was to report on the kind of test items the two commercially produced English-language proficiency tests use and the kinds of constructs they measure. As for the test items, both tests evaluated reading comprehension, grammar, and vocabulary. With regards to the construct, the TOEFL test assesses “reading for a purpose” as described and the TOEIC assesses reading comprehension, grammar and vocabulary. Neither the TOEFL test nor the TOEIC test appear to measure orthographic knowledge or phonological awareness. The number of words students have to recognise in the TOEIC test, however, is two times greater than that in the TOEFL test. In this respect, the TOEIC test might lean slightly towards measuring word recognition efficiency, which leads to reading fluency in written English texts.

As for the latter research question whether the test items reflect the Faculty of Nursing’s curriculum, the TOEIC test might be slightly more suitable for assessing students’ reading fluency in English texts. With respect to orthographic knowledge and phonological awareness, however, neither the TOEFL test nor the TOEIC test items appear to directly measure the two skills. Moreover, the orthographic and phonological features of the two tests’ terms did not match the features of medical terms regarding the number of syllables per word or the grapheme–phoneme correspondences. With this respect, the two commercially produced English-language proficiency tests might not be useful for the Faculty of Nursing’s placement purposes.

Chapter 5: Item Bank Development

The purpose of this chapter is to report on the process and results of a study examining following Research Question (RQ) 3: Does the development of an item bank for the English-language placement test constitute a valid measure for assessing students' orthographic and phonological awareness, and how well do test items reflect the content of the faculty's curriculum? This chapter first describes the process of developing test items and later reports the analysis of test items regarding their usefulness for the item bank that stores test items addressing various levels and aspects of orthographic and phonological awareness.

5.1 Test development

As described in the preceding chapters, the aim of developing the item bank is to design and store test items that assess frequently neglected areas of proficiency—namely, orthographic and phonological processing skills, in order to reflect the Faculty of Nursing's curriculum requirements. In Chapter 2, the author hypothesised that certain orthographic and phonological features are more challenging for Japanese learners of English due to the orthographic and phonological differences between L1 (Japanese) and FL (English). In Chapter 3, the author identified several orthographic and phonological features of medical English terms: their significantly more complex grapheme–phoneme correspondences than those of Japanese one-to-one relationships between graphemes–phonemes, their polysyllabic structure, their use of adapted pronunciation in loanwords, and the infrequent use of open-ended syllables, where syllables end with vowel. To validate the development of in-house test items, the author examined whether the two commercially produced English proficiency tests assess these proficiencies and found that they do not, as outlined in Chapter 4.

Based on the findings of the preceding chapters, this chapter describes the development of test items and their analysis. It first describes the development of orthographic and phonological test items in five tests (including a 2014 a Spring Term final test (14SF), a Fall 2014 final test (14FF), a 2015 placement test (15P), a Spring 2015 final test (15SF), and a Fall 2015 final test (15FF)) and analyses the use of the tests. Each test was analysed for its separation and reliability, targeting, item fit, and unidimensionality. The chapter then describes the process and the results of calibration, the construction of item measures in the internal frame of reference (Linacre, 2013).

5.1.1 Materials and participants

Table 5.1 lists tests containing orthographic and phonological items by the type and number of participants who took the test, as well as type and number of test items, including common items. Ranging in age from 18 to 30 years, all participants were either first- or second-year students in the Faculty of Nursing during the 2014–2015 academic year. Most participants were women—in fact, less than 10% were men—and most had graduated from a full–time high school, meaning that they had been taught English in school for more than six years: three years each in both junior high school and high school. To equate the tests, some items in each test were used as common items in the preceding tests: for example, five test items used in the 14SF were used as common items for equating the 14SF and 14FF. The total numbers of original participants (244) and original items (147) are therefore given on the table.

Table 5.1

Tests by Type and Number of Participants and Items

	Tests	Participants					Items		
		Original		Common		Total	Original	Common	Total
1	Spring 2014 final test (14SF)	2014 entrants (Classes A & D)	51		0	51	18	0	18
2	Fall 2014 final test (14FF)	2013 entrants (Class B)	26		0	26	23	5 (14SF)	28
3	2015 placement test (15P)	2015 entrants (Classes B & C)	60	2014 entrants (Class C)	22	82	27	0	27
4	Spring 2015 final test (15SF)	2015 entrants (Classes A & D)	56	2014 entrants (Class C)	22	78	41	18 (14SF)	59
5	Fall 2015 final test (15FF)	2014 entrants (Class B)	29		0	29	38	16 (14SF)	54
	Total		244				147		

Table 5.2 shows the number of words and types of grapheme regarding orthographic and phonological features in the original test items. Since each test item includes two words, there are 294 words in total. The letters “JE” in the table indicate that the grapheme–phoneme relationship is the same as that in Japanese. By contrast, “NE” indicates that the grapheme has a phoneme that can be pronounced by Japanese speakers but that the grapheme–phoneme relationship differs from that in Japanese. Lastly, “NNE” indicates that the target grapheme is pronounced in an English-language phoneme that does not exist in Japanese pronunciation. As described in the previous chapter, reading any grapheme that does not follow Japanese one-to-one grapheme–phoneme relationships and any phoneme that does not exist in Japanese pronunciation is the most difficult reading for Japanese learners of English. For example, reading grapheme /a/ as [æ] or [ʌ] is more difficult than reading /a/ as [ei], because the pronunciation of [ei] exists in Japanese for Japanese learners of English. In contrast, reading grapheme /a/ as [ɔ:] is less difficult than reading grapheme /a/ as [æ] or [ʌ], because phoneme [ɔ:] exists in Japanese pronunciation. Of 294 words examined, approximately half required the test takers to identify NE graphemes, while a third targeted the test taker’s awareness of NNE graphemes. Less than a fifth of the words targeted JE graphemes.

Table 5.2

Number and Types of Words in Original Test Items

		Test items						
		Number of words	JE		NE		NNE	
	Test	*n*	*n*	%	*n*	%	*n*	%
1	Spring 2014 final test (14SF)	36	5	14	21	58	10	27
2	Fall 2015 final test (14FF)	46	9	20	19	41	18	39
3	2015 placement test (15P)	54	13	24	26	48	15	28
4	Spring 2015 final test (15SF)	82	7	8	44	54	31	38
5	Fall 2015 final test (15FF)	76	17	22	30	40	29	38
Total		294	51	17	140	48	103	35

Note: The letters "JE" in the table indicate that the grapheme–phoneme relationship is the same as that in Japanese. By contrast, "NE" indicates that the grapheme has a phoneme that can be pronounced by Japanese speakers but that the grapheme–phoneme relationship differs from that in Japanese. Lastly, "NNE" indicates that the target grapheme is pronounced in an English-language phoneme that does not exist in Japanese pronunciation.

5.1.2 Analyses

The measurement computer program WINTSTEPS® Rasch version 3.81.0, was used for analysis (Linacre, 2006). Each test was analysed for its separation and reliability, targeting, item fit, and unidimensionality. This section describes the guidelines of each criterion based on the preceding statistical analysis (Bond & Fox, 2007; Boone, Staver, & Yale, 2014; Linacre, 2013).

Separation and reliability

WINSTEPS® version 3.81.0 provides two separation measures: person separation with person reliability and item separation with item reliability. In this study, the measure of person separation was used to categorise test takers based on their performance.

Fischer (2007) provides rating scale instrument quality criteria regarding person and item measurement reliability and person and item strata separation. Among these criteria, any person or item reliability coefficient greater than 0.8 and any person or item strata separation index greater than 3.0 is considered to be good. Whereas sample ability variance, test length, and sample–item targeting affect person reliability, item difficulty variance and sample size influence item reliability. WINSTEPS® person reliability is consistent with traditional test reliability, which is the ratio of true variance to observed variance equivalent to Cronbach's alpha and KR–20; whereas Cronbach's alpha and KR–20 include persons with extreme scores, WINSTEPS® reliability does not. Moreover, Cronbach's alpha approximates true variance with an analysis of variance, while KR–20 does the same with a summary of item point-biserial and WINSTEPS® with the measure of standard error (Linacre, 2013).

Linacre (2013) states in the WINSTEPS manual that any person separation index less than 2.0 with a person reliability coefficient less than 0.8 indicates a test's incapability to divide test takers by performance and thus that additional items are necessary. Item separation indicates the extent to which the test items can be separated in terms of difficulty. An item separation index less than 3.0 with a person reliability coefficient less than 0.9 suggests that the sample of persons is not large enough to distinguish items of high and low difficulty. Whereas person reliability is influenced by sample ability variance, length of test, number of categories per item, and sample–

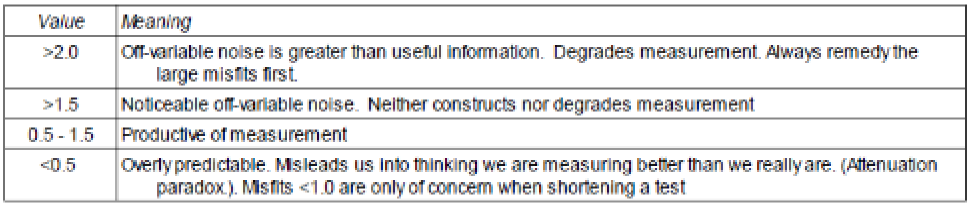
item targeting, item difficulty is influenced by item difficulty variance and person sample size. Therefore, to improve person reliability, it is necessary to have a wider range in examinee ability, numerous test items and categories, and better targeting. To improve item reliability, it is also necessary to have a wide range of difficulty in test items and a larger sample size. Since the goal of the thesis is to build an item bank with a wide variety of proficiency levels and to not divide test takers by performance, low person reliability and person separation index were not included as criteria in this process.

Test targeting

Test targeting refers to the extent to which the difficulty of test items is appropriate to the person's estimated ability level. WINSTEPS® provides a distribution of item difficulty and person ability estimates on the same continuum of measurement unit to allow comparison between them. It also provides the most probable response key map, which places the difficulty metric horizontally along the x-axis. Items are listed along the right side of the figure, with most difficult items at the top. A horizontal gap between the items indicates that no items covered that space. The map thus provides insights into which difficulty levels are lacking in the present test.

Unidimensionality

Although several tools are available for assessing unidimensionality, WINSTEPS® offers infit and outfit mean square fit statistics and principal components analysis of residuals (Linacre, 2013). Item fit statistics provide information on the extent to which the observed person's response corresponds to the expected response based on the Rasch model. WINSTEPS® provides two types of fit statistics to assess the residual difference between actual and expected responses. On the one hand, infit mean square statistic (MNSQ) is affected by unexpected responses, such as high performer mistakes on a low difficulty item. On the other, outfit MNSQ is affected by any unexpected pattern of responses proximate to a person's ability estimates. Although Fischer (2007) suggests that MNSQ values between 0.50 and 1.3 should be retained as well-fitting items, Linacre (2013) provides a table of values and their meanings, in which values between 0.5 and 1.5 are assumed to be productive measurements.



Value	*Meaning*
>2.0	Off-variable noise is greater than useful information. Degrades measurement. Always remedy the large misfits first.
>1.5	Noticeable off-variable noise. Neither constructs nor degrades measurement
0.5 - 1.5	Productive of measurement
<0.5	Overly predictable. Misleads us into thinking we are measuring better than we really are. (Attenuation paradox.). Misfits <1.0 are only of concern when shortening a test

Figure 5.1. Mean square statistic (MNSQ) values and their meanings (Linacre, 2013)

Based on the arguments above, the present research adopted Jarl, Heinemann, and Hermansson's (2012) guidelines for item fit. Items with MNSQ values greater than 1.4 with ZSTDs in excess of 2.0 were categorised as misfitting, whereas items with MNSQ values less than 0.6 with ZSTD values less than −2.0 were categorised as overfitting. Items that fit these guidelines were analysed for their content and the researcher assessed whether the value improves when the item was omitted.

Linacre (2013) suggested that the aim of the principal components analysis of Rasch residuals is "to extract the common factor that explains the most residual variance under the hypothesis that there is such a factor." When the persons and items fit the model expectation, the study examined whether the variance explained by the first contrast is less than 10% (Fisher, 2007; Hsiao, Shih, Yu, Hsieh, & Hseih, 2015).

Table 5.3 shows data from the revision process of each test's participants and items based on an analysis of the person and item fit analysis. For Spring 2014 final test (14SF), the number of person decreased to 42 from 51 by two revision. In contrast, the number of items for 14SF was constantly 18 since all items fit the Rasch model. Additionally, whereas Fall 2014 final test (14FF) took four revisions to fit the Rasch model, the Fall 2015 final test (15FF) needed only one revision.

Table 5.3

Revision Processes of Tests Regarding the Number of Participants and Items

Tests		1st analysis		1st revision		2nd revision		3rd revision		4th revision	
		Person	Item	Person	Item	Person	Item	Person	Item	Person	Item
1	Spring 2014 final test (14SF)	51	18	43	18	42	18	N/A	N/A	N/A	N/A
2	Fall 2014 final test (14FF)	26	28	24	27	23	27	22	27	21	27
3	2015 placement test (15P)	82	27	78	27	75	27	N/A	N/A	N/A	N/A
4	Spring 2015 final test (15SF)	83	59	75	59	71	59	68	59	67	59
5	Fall 2015 final test (15FF)	29	54	25	54	N/A	N/A	N/A	N/A	N/A	N/A

5.2.1 The Spring 2014 final test

As shown in Table 5.3, the Spring 2014 final test has undergone two revisions. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each revision in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals. Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.2.2 Basic feature of the Spring 2014 final test

The number of test items that assess orthographic and phonological awareness was 18 in the Spring 2014 final test. Fifty-one first-year students (2014 entrants) took a one-hour test on: reading comprehension, with ten test items; vocabulary questions, with ten test items; listening comprehension, with five test items; as well as orthographic and phonological awareness of 18 test items.

As shown in Table 5.2, ten words (27.8%) of the 36 test item words targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in Japanese pronunciation in the Spring 2014 final test. Twenty-one words (58.3%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining five words (13.9%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 18 test items, seven items (38.8%) assess students to identify two NE words, while three items assess students' ability to identify a combination of a JE word and a NE word, a NE word and a NNE word, and a NNE word and a NNE word. Of 36 words, 19 words (52.7%) were loanwords. For example, the second word of item number seven was "oven" which pronounces the grapheme /o/ as [ʌ] in English while Japanese pronounces [o:]. Of 18 test items, four test items (22.2%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the Spring 2014 final test shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.2.3 Revision process of the Spring 2014 final test

As shown in Table 5.4, the Spring 2014 final test has undergone two revisions deleting nine persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table 5.4

Misfitting Persons and Items by Revisions of the Spring 2014 Final Test

Revision	Misfitting Person MNSQ>2.0 with ZSTD>1.4 or -2.0<MNSQ with ZSTD<.6	n	Misfitting Item MNSQ>2.0 with ZSTD>1.4 or -2.0<MNSQ with ZSTD<.6	n
1st Analysis	P1, P4, P6, P10, P21, P43, P47, P51	8	N.A.	0
1st Revision	P43	1	N.A.	0
2nd Revision	N.A.	0	N.A.	0

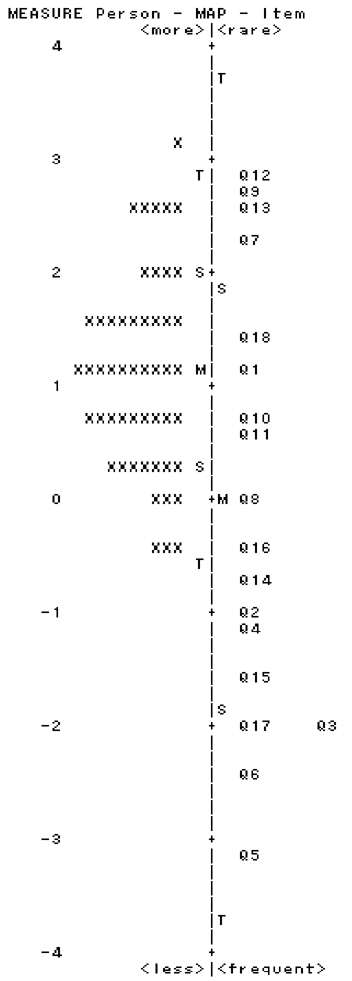
Original analysis

As shown in Table 5.3, there were 18 items and 51 persons in the original analysis. WINSTEPS^{R} version 3.81.0 provides the statistical values of the original test items.

The separation measure of the Spring 2014 final test is relatively low (0.65–0.8), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is also very low (0.35), which is due to the small number of test items. Item is high (0.93) which indicates that, if the items were given to other comparable groups of test takers, there is a high probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.58–3.62 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.2 shows that the persons mean (1.12) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which item difficulty estimates fall far below the person ability estimates for the Spring 2014 final test.

Of the 51 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 1(Infit MNSQ: 1.84, ZSTF 2.03), 4 (Infit MNSQ: 1.81, ZSTD: 2.00), 6 (Outfit MNSQ: 4.47, ZSTD: 1.84), 10 (Infit MNSQ: 0.42, ZSTD: −2.01), 21 (Infit MNSQ: 0.43, ZSTD: −2.00), 43 (Outfit MNSQ: 4.39, ZSTD: 2.77), 47 (Infit MNSQ: 0.42, ZSTD: −2.01), and 51 (Infit MNSQ: 0.43, ZSTD: −2.00)'s MNSQ values with ZSTD values, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ values greater than 1.4 with ZSTDs in excess of 2.0 or MNSQ values less than 0.6 with a ZSTD value less than −2.0 indicates possibly mismatched persons, suggesting that persons 1, 4, 6, 10, 21, 43, 47, and 5 should be deleted from the list. Of the 18 items measured, all MNSQ ranges fell between 0.6-1.4. Therefore, all 18 items remain in the 1st revision.



```
MEASURE Person - MAP - Item
            <more>|<rare>
    4             +
                  |
                  |T
                  |
                  |
                  |
               X  |
    3             +
                 T|  Q12
                  |  Q9
           XXXXX  |  Q13
                  |
                  |  Q7
                  |
    2       XXXX S+
                  |S
                  |
       XXXXXXXXX  |
                  |  Q18
                  |
      XXXXXXXXXX M|  Q1
    1             +
                  |
       XXXXXXXXX  |  Q10
                  |  Q11
                  |
         XXXXXXX S|
                  |
    0        XXX  +M Q8
                  |
                  |
             XXX  |  Q16
                 T|
                  |  Q14
                  |
   -1             +  Q2
                  |  Q4
                  |
                  |
                  |  Q15
                  |
                  |S
   -2             +  Q17    Q3
                  |
                  |
                  |  Q6
                  |
                  |
                  |
   -3             +
                  |  Q5
                  |
                  |
                  |
                  |T
                  |
   -4             +
            <less>|<frequent>
```

Figure 5.2. Item and person map for the Spring 2014 final test (original analysis*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S,* and *T* indicate the mean, 1- and 2- standard deviations respectively.

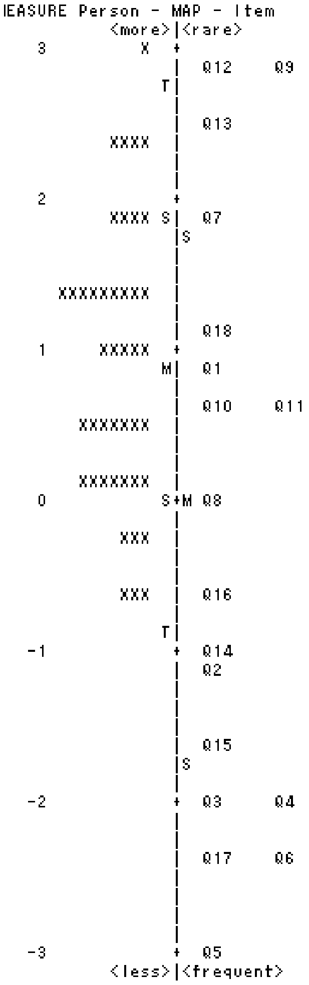
1st revision

As shown in Table 5.3, there were 18 items and 43 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items and persons.

The separation measure concerning the 1st revision of the Spring 2014 final test is relatively low (0.81–0.92), which indicates that the number of items used is rather narrow to distinguish persons. Its person reliability is low (0.43), which is due to the relatively small number of test items. The reliability of item is relatively high (0.90) which indicates that, if the items were given to other comparable groups of test takers, there is a high probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.96–2.98 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.3 shows that the persons mean (0.93) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which item difficulty estimates fall far below the person ability estimates for the Spring 2014 final test.

Of the 43 persons measured, all MNSQ ranges fell between 0.6-1.4, except for person 43's MNSQ value (2.87) with a ZSTD value (2.90), possibly indicating a mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value greater than 1.4 with a ZSTD value in excess of 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 43 should be deleted from the list. Of the 18 items measured, all MNSQ ranges fell between 0.6-1.4.



```
IEASURE Person - MAP - Item
           <more>|<rare>
    3         X  +
                 |  Q12    Q9
                T|
                 |
                 |  Q13
           XXXX  |
                 |
                 |
    2            +
           XXXX S|  Q7
                 |S
                 |
                 |
      XXXXXXXXX  |
                 |
                 |  Q18
    1     XXXXX  +
                M|  Q1
                 |
                 |  Q10    Q11
        XXXXXXX  |
                 |
                 |
        XXXXXXX  |
    0           S+M Q8
                 |
            XXX  |
                 |
                 |
            XXX  |  Q16
                 |
                T|
   -1            +  Q14
                 |  Q2
                 |
                 |
                 |
                 |  Q15
                 |S
                 |
   -2            +  Q3     Q4
                 |
                 |
                 |  Q17    Q6
                 |
                 |
                 |
                 |
   -3            +  Q5
           <less>|<frequent>
```

Figure 5.3. Item and person map for the Spring 2014 final test (1st revision*)*.

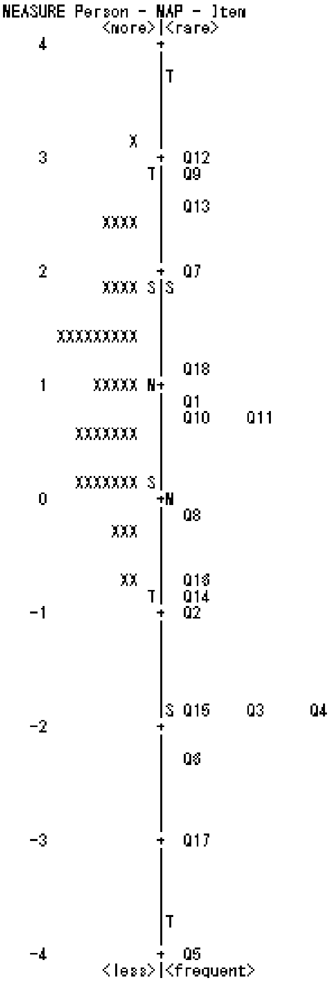
Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

2nd revision

As shown in Table 5.3, there were 18 items and 42 persons in the 2nd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items and persons of the 2nd revision.

The separation measure regarding the 2nd revision of the Spring 2014 final test is relatively low (0.77–0.89), which indicates that the number of items used is rather narrow to distinguish persons. Its person reliability is also low (0.4), which is due to the relatively small number of test items. The reliability of item is moderately high (0.89–0.90) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measures of 2.92–2.93 indicate that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.4 shows that the persons mean (0.99) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which item difficulty estimates fall far below the person ability estimates for the Spring 2014 final test.



```
MEASURE Person - MAP - Item
           <more>|<rare>
    4            +
                 |T
                 |
                 |
                 |
              X  |
    3            +  Q12
                T|  Q9
                 |
                 |  Q13
           XXXX  |
                 |
                 |
    2            +  Q7
           XXXX S|S
                 |
                 |
      XXXXXXXXX  |
                 |
                 |  Q18
    1     XXXXX M+
                 |  Q1
                 |  Q10    Q11
        XXXXXXX  |
                 |
                 |
        XXXXXXX S|
    0            +M
                 |  Q8
            XXX  |
                 |
                 |
             XX  |  Q16
                T|  Q14
   -1            +  Q2
                 |
                 |
                 |
                 |
                 |
                 |S Q15    Q3     Q4
   -2            +
                 |
                 |  Q6
                 |
                 |
                 |
                 |
   -3            +  Q17
                 |
                 |
                 |
                 |
                 |T
   -4            +  Q5
           <less>|<frequent>
```

Figure 5.4. Item and person map for the Spring 2014 final test (2nd revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 42 persons measured, all MNSQ ranges fell between 0.6-1.4. Of the 18 items measured, all MNSQ ranges fell between 0.6-1.4.

5.2.4 Unidimensionality

In this section, the unidimensionality of the final revision (2nd revision) will be described in terms of principal component analysis of residuals. Table 5.5 indicates that observed raw variance explained by measures (43.1%) mildly fits the expected raw variance explained by measure (42.7%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 31.5% of the 18 items, leaving more than half of variance (56.9%) unaccounted for by the model. This is due to the fact that the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth and fifth respectively. The largest secondary dimension (first contrast) in the Spring 2014 final test data had strength of 2.4 units (7.7%) while the variance explained by measures was larger at 13.7 units (43.1%) and the variance explained by items was larger at 10.0 units (31.5%), indicating that the secondary contrast does not create multidimensionality.

Table 5.5

Standardized Residual Variance (in Eigen-value Units) for the Spring 2014 Final Test

	Observed			Expected
Total raw variance in observations	31.7	100%		100%
Raw variance explained by measures	13.7	43.1%		42.7%
Raw variance explained by persons	3.7	11.6%		11.5%
Raw variance explained by items	10.0	31.5%		31.2%
Raw unexplained variance (total)	18.0	56.9%	100%	57.3%
Unexplained variance in 1st contrast	2.4	7.7%	13.5%	
Unexplained variance in 2nd contrast	2.1	6.6%	11.7%	
Unexplained variance in 3rd contrast	1.9	5.9%	10.3%	
Unexplained variance in 4th contrast	1.7	5.2%	9.2%	
Unexplained variance in 5th contrast	1.4	4.6%	8.0%	

Note. This figure is from WINSTEPS® version 3.81.0 output Table 23.0.

5.3.1 The Fall 2014 final test

As shown in Table 5.3, the Fall 2014 final test has undergone four revisions. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each revision in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals. Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.3.2 Basic feature of the Fall 2014 final test

The number of test items that assess orthographic and phonological awareness was 28 in the Fall 2014 final test. Twenty-six sophomore students (2013 entrants) took a one-hour test on: reading comprehension, with ten test items; vocabulary questions, with ten test items; listening comprehension, with five test items; as well as the orthographic and phonological awareness of 28 test items.

As shown in Table 5.2, 14 words (25%) of the 56 test item words targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in Japanese pronunciation in the Fall 2014 final test. Thirty-four words (60.7%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining eight words (14.3%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 28 test items, seven items (25%) assess students to identify two NE words, while six items (21.5%) assess students' ability to identify a combination of a JE word and a NE word, a NE word and a NNE word, and a NNE word and a NNE word. Of 56 words, 29 words (51.7%) were loanwords. For example, the first word of item number 16 was "waitress" which pronounces the grapheme /t/ as [t] in English while Japanese pronounces [to]. Of 28 test items, nine test items (32.1%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the Fall 2014 final test shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.3.3 Revision process of the Fall 2014 final test

As shown in Table 5.6, test has undergone four revisions deleting one test item and five persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table *5.6*

Misfitting Persons and Items by Revisions of the Fall 2014 Final Test

Revision	Misfitting Person	n	Misfitting Item	n
	MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6		MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	
1st Analysis	P2, P15	2	Q3	1
1st Revision	P10	1	N.A.	0
2nd Revision	P2	1	N.A.	0
3rd Revision	P1	1	N.A.	0
4th Revision	N.A.	0	N.A.	0

Original analysis

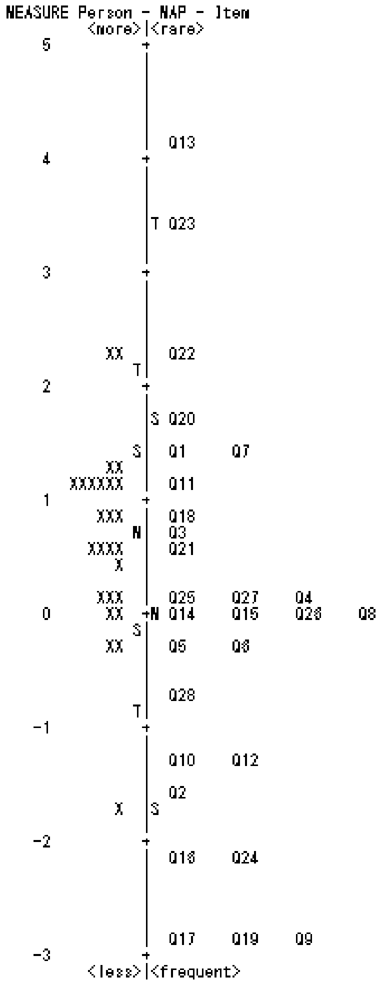
As shown in Table 5.3, there were 28 items and 26 persons in the original analysis. WINSTEPS® version 3.81.0 provides the statistical values of the original test items.

The separation measure concerning the original analysis of the Fall 2014 final test is relatively low (1.13–1.22), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.58), which is due to the relatively small number of test items. The reliability of item is moderately high (0.86–0.87) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.52–2.6 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.5 shows that the persons mean (0.66) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the Fall 2014 final test.

Of the 26 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 2 (Outfit MNSQ: 3.42, ZSTD: 2.81) and 15 (Infit MNSQ: 1.82, ZSTD: 3.49)'s MNSQ value with a ZSTD value, possibly indicating mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 2 and 15 should be deleted from the list.

Of the 28 items measured, all MNSQ ranges fell between 0.6-1.4, except for item 3's outfit MNSQ value (1.85) with an outfit ZSTD value (4.05), possibly indicating a mismatched item. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched item, suggesting that item 3 should be deleted from the list.



```
MEASURE Person - MAP - Item
         <more>|<rare>
    5          +
               |
               |
               |
               |
               |
               |  Q13
    4          +
               |
               |
               |
               |T Q23
               |
               |
    3          +
               |
               |
               |
               |
           XX  |  Q22
              T|
    2          +
               |
               |S Q20
               |
              S|  Q1     Q7
           XX  |
       XXXXXX  |  Q11
    1          +
          XXX  |  Q18
              M|  Q3
         XXXX  |  Q21
            X  |
               |
          XXX  |  Q25    Q27    Q4
    0      XX +M  Q14    Q15    Q26    Q8
              S|
           XX  |  Q5     Q6
               |
               |
               |  Q28
              T|
   -1          +
               |
               |  Q10    Q12
               |
               |  Q2
            X  |S
               |
   -2          +
               |  Q16    Q24
               |
               |
               |
               |
               |  Q17    Q19    Q9
   -3          +
         <less>|<frequent>
```

Figure 5.5. Item and person map for the Fall 2014 final test (original analysis*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

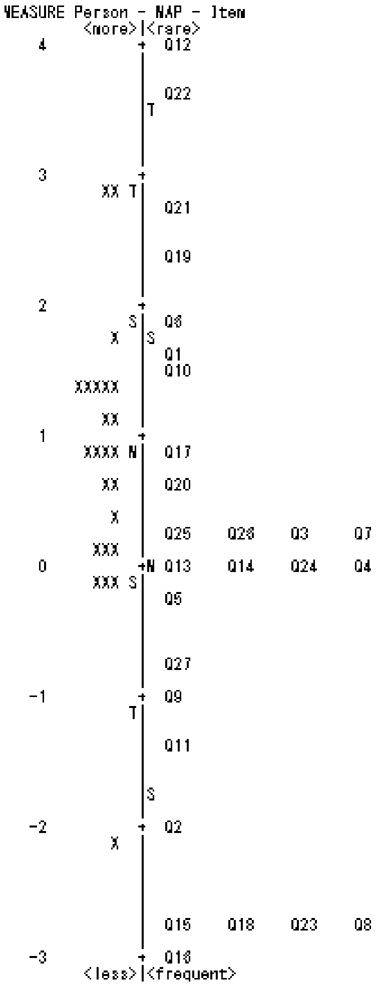
1st *revision*

As shown in Table 5.3, there were 27 items and 24 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 1st revision.

The separation measure concerning the 1st revision of the Fall 2014 final test is relatively low (1.48–1.58), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.69), which is due to the relatively small number of test items. Item is relatively high (0.85–0.86) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.38–2.47 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.6 shows that the persons mean (0.84) is located above the items mean that is set to 0.00 by default, indicating that, on average, the items are relatively easy for the persons. There were, however, some item difficulty estimates fall far below the person ability estimates for the Fall 2014 final test.

Of the 24 persons measured, all MNSQ ranges fell between 0.6-1.4, except for person 10’s outfit MNSQ value (2.50) with an outfit ZSTD value (2.06), possibly indicating a mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 10 should be deleted from the list. Of the 27 items measured, all MNSQ ranges fell between 0.6-1.4.



```
MEASURE Person - MAP - Item
          <more>|<rare>
    4          +  Q12
               |
               |
               |  Q22
               |T
               |
               |
               |
    3          +
          XX T |
               |  Q21
               |
               |
               |  Q19
               |
               |
    2          +
             S |  Q6
           X   |S
               |  Q1
               |  Q10
       XXXXX   |
               |
          XX   |
    1          +
        XXXX M |  Q17
               |
          XX   |  Q20
               |
           X   |
               |  Q25    Q26    Q3     Q7
         XXX   |
    0          +M Q13    Q14    Q24    Q4
         XXX S |
               |  Q5
               |
               |
               |
               |  Q27
               |
   -1          +  Q9
             T |
               |
               |  Q11
               |
               |
               |S
               |
   -2          +  Q2
           X   |
               |
               |
               |
               |
               |  Q15    Q18    Q23    Q8
               |
   -3          +  Q16
          <less>|<frequent>
```

Figure 5.6. Item and person map for the Fall 2014 final test (1st revision*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M, S,* and *T* indicate the mean, 1- and 2- standard deviations respectively.

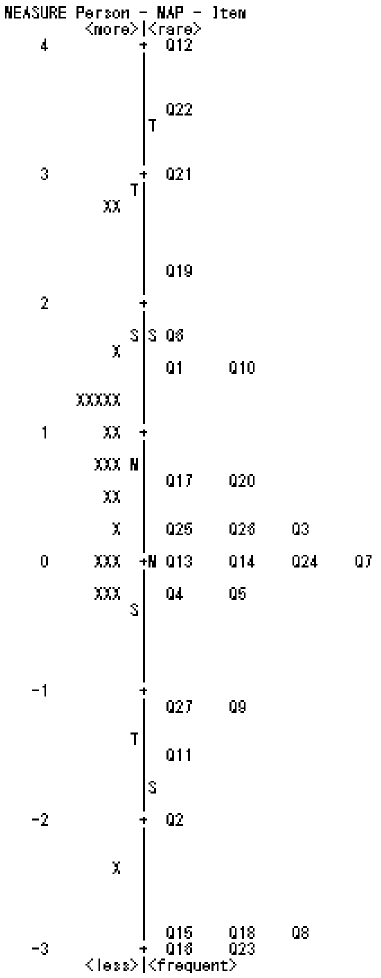
2nd revision

As shown in Table 5.3, there were 27 items and 23 persons in the 2nd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 2nd revision.

The separation measure concerning the 2nd revision of the Fall 2014 final test is relatively low (1.55–1.66), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.71), which is due to the relatively small number of test items and a narrow range of person measure. The reliability of item is moderately high (0.84–0.85) which indicates that, if the items were given to other comparable groups of test takers, there is a moderate probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.32–2.39 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.7 shows that the persons mean (0.71) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are several item difficulty estimates fall below the person ability estimates for the Fall 2014 final test.

Of the 23 persons measured, all MNSQ ranges fell between 0.6-1.4, except for person 2’s outfit MNSQ value (2.47) with outfit ZSTD value (2.10), possibly indicating a mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 2 should be deleted from the list. Of the 27 items measured, all MNSQ ranges fell between 0.6-1.4.



```
MEASURE Person - MAP - Item
           <more>|<rare>
    4            +  Q12
                 |
                 |
                 |
                 |  Q22
                 |T
                 |
                 |
    3            +  Q21
                T|
             XX  |
                 |
                 |
                 |
                 |  Q19
                 |
    2            +
                 |
                S|S Q6
              X  |
                 |  Q1      Q10
                 |
          XXXXX  |
                 |
    1        XX  +
                 |
            XXX M|
                 |  Q17     Q20
             XX  |
                 |
              X  |  Q25     Q26     Q3
                 |
    0       XXX  +M Q13     Q14     Q24     Q7
                 |
            XXX  |  Q4      Q5
                S|
                 |
                 |
                 |
                 |
   -1            +
                 |  Q27     Q9
                 |
                T|
                 |  Q11
                 |
                 |S
                 |
   -2            +  Q2
                 |
                 |
              X  |
                 |
                 |
                 |
                 |  Q15     Q18     Q8
   -3            +  Q16     Q23
           <less>|<frequent>
```

Figure 5.7. Item and person map for the Fall 2014 final test (2nd revision*).*

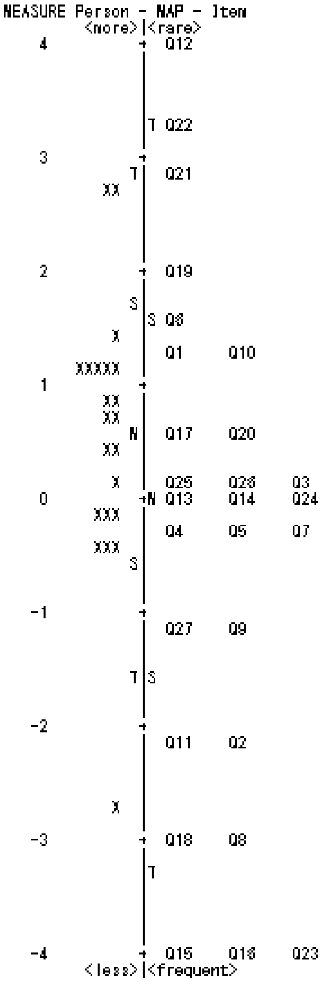
Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

3rd revision

As shown in Table 5.3, there were 27 items and 22 persons in the 3rd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 3rd revision.

The separation measure concerning the 3rd revision of the Fall 2014 final test is relatively low (1.63–1.74), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.72), which is due to the relatively small number of test items). The reliability of item is moderately high (0.84) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.25–2.31 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.8 shows that the persons mean (0.58) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are some items which item difficulty estimates fall below the person ability estimates for the 3rd revision of the Fall 2014 final test.



```
MEASURE Person - MAP - Item
         <more>|<rare>
    4          +  Q12
               |
               |
               |
               |T Q22
               |
    3          +
              T|  Q21
           XX  |
               |
               |
               |
    2          +  Q19
               |
              S|
               |S Q6
            X  |
               |  Q1     Q10
        XXXXX  |
    1          +
           XX  |
           XX  |
              M|  Q17    Q20
           XX  |
               |
            X  |  Q25    Q26    Q3
    0          +M Q13    Q14    Q24
          XXX  |
               |  Q4     Q5     Q7
          XXX  |
              S|
               |
               |
   -1          +
               |  Q27    Q9
               |
               |
              T|S
               |
               |
   -2          +
               |  Q11    Q2
               |
               |
               |
            X  |
               |
   -3          +  Q18    Q8
               |
               |T
               |
               |
               |
   -4          +  Q15    Q16    Q23
         <less>|<frequent>
```

Figure 5.8. Item and person map for the Fall 2014 final test (3rd revision*).*

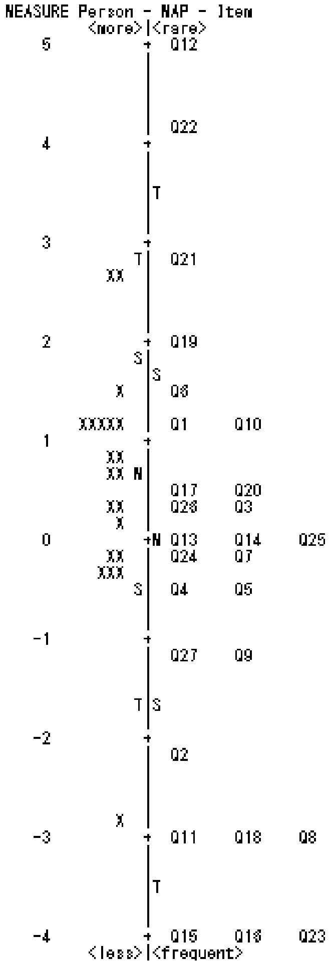
Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 22 persons measured, all MNSQ ranges of person fit fell between 0.6–1.4, except for person 1's outfit MNSQ value (2.19) with an outfit ZSTD value (2.12) possibly indicating a mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 1 should be deleted from the list. Of the 27 items measured, all MNSQ ranges of item fit fell between 0.6 and 1.4, suggesting a reasonable fit of the data to the model.

4th revision

As shown in Table 5.3, there were 27 items and 21 persons in the 4th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items 4th revision.

The separation measure concerning the 4th revision of the Fall 2014 final test is relatively low (1.68–1.79), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.73), which is due to the relatively small number of test items. The reliability of item is moderately high (0.83–0.84) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.24–2.31 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).



```
MEASURE Person - MAP - Item
          <more>|<rare>
    5           +  Q12
                |
                |
                |
                |  Q22
    4           +
                |
                |T
                |
                |
    3           +
               T|  Q21
          XX    |
                |
                |
    2           +  Q19
               S|
                |S
           X    |  Q6
                |
       XXXXX    |  Q1      Q10
    1           +
          XX    |
          XX  M |
                |  Q17     Q20
          XX    |  Q26     Q3
           X    |
    0           +M Q13     Q14     Q25
          XX    |  Q24     Q7
         XXX    |
               S|  Q4      Q5
                |
                |
   -1           +
                |  Q27     Q9
                |
                |
               T|S
                |
   -2           +
                |  Q2
                |
                |
                |
           X    |
   -3           +  Q11     Q18     Q8
                |
                |
                |T
                |
                |
   -4           +  Q15     Q16     Q23
          <less>|<frequent>
```

Figure 5.9. Item and person map for the Fall 2014 final test (4th revision*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Figure 5.9 shows that the persons mean (0.61) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the Fall 2014 final test. Of the 21 persons measured, all MNSQ ranges of person fit fell between 0.6–1.4 and of the 27 items measured, all MNSQ ranges of item fit fell between 0.6–1.4.

5.3.4 Unidimensionality

In this section, the unidimensionality of the final (4th) revision is described in terms of principal component analysis of residuals. Table 5.7 indicates that observed raw variance explained by measures (38.2%) mildly fits the expected raw variance explained by measure (38.5%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 25% of the 27 items, leaving more than half of variance (61.8%) unaccounted for. This is because the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth, and fifth, respectively. The largest secondary dimension (first contrast) in the Fall 2014 final test data had a strength of 4.0 units (10.8%), while the variance explained by measures was larger at 14.2 units (38.2%). The variance explained by items was larger at 9.3 units (25.0%), suggesting the possibility of multidimensionality.

Table 5.7

Standardized Residual Variance (in Eigen-value Units) for the Fall 2014 Final Test

	Observed			Expected
Total raw variance in observations	37.2	100%		100%
Raw variance explained by measures	14.2	38.2%		38.5%
Raw variance explained by persons	4.9	13.2%		13.3%
Raw variance explained by items	9.3	25.0%		25.2%
Raw unexplained variance (total)	23.0	61.8%	100%	61.5%
Unexplained variance in 1st contrast	4.0	10.8%	17.4%	
Unexplained variance in 2nd contrast	3.4	9.0%	14.6%	
Unexplained variance in 3rd contrast	2.6	7.0%	11.4%	
Unexplained variance in 4th contrast	2.2	5.9%	9.5%	
Unexplained variance in 5th contrast	1.8	4.8%	7.8%	

Note. This figure is from WINSTEPS^{R} version 3.81.0 output Table 23.0.

The close examination of the content of items reveals that all 27 items are targeted to assess the test takers' ability to identify grapheme-phoneme relationships of English words. As Linacre (2013) points out that while the Fall 2014 final test's 1st contrast had a strength of 4.0 units, the instrument measures students' orthographic and phonological processing skills in general and is therefore unidimensional for the purpose.

5.4.1 The 2015 placement test

As shown in Table 5.3, the 2015 placement test has undergone two revisions. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each revision in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals. Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.4.2 Basic feature of the 2015 placement test

The number of test items that assess orthographic and phonological awareness was 27 in the 2015 placement test. Sixty 2015 entrants and twenty-two 2014 entrants took a one-hour test on: reading comprehension, with ten test items; vocabulary questions, with ten test items; listening comprehension, with five test items; as well as the orthographic and phonological awareness of 27 test items. As shown in Table 5.2, 16 words (29.6%) of the 54 test item words targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in Japanese pronunciation in the 2015 placement test. Twenty-six words (48.2%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining 12 words (22.2%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 27 test items, 11 items (40.7%) assess students to identify two NE words, while five items (18.5%) assess students' ability to identify a combination of a JE word and a JE word and a NNE word and a NNE word. Of 54 words, 26 words (48.1%) were loanwords. For example, the second word of item number 25 was "slow" which pronounces the grapheme /s/ as [s] in English while Japanese pronounces [su]. Of 27 test items, four test items (14.8%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the 2015 placement test shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.4.3 Revision process of the 2015 placement test

As shown in Table 5.8, the 2015 placement test has undergone two revisions deleting six persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table 5.8

Misfitting Persons and Items by Revisions of the 2015 Placement Test

Revision	Misfitting Person	n	Misfitting Item	n
	MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6		MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	
1st Analysis	P72, P75, P76, P80	4	N.A.	1
1st Revision	P7, P72	2	N.A.	0
2nd Revision	N.A.	0	N.A.	0

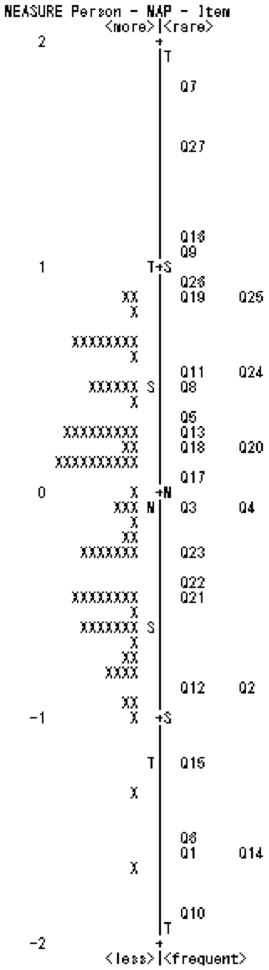
Original analysis

As shown in Table 5.3, there were 27 items and 82 persons in the original analysis. WINSTEPS® version 3.81.0 provides the statistical values of the original test items.

The separation measure concerning the original analysis of the 2015 placement test is relatively low (0.65–0.73), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.36), which is due to the relatively small number of test items. The reliability of item is relatively high (0.93) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.65–3.69 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.10 shows that the persons mean (−0.08) is located just below the items mean that is set to 0.00 by default, indicating that, on average, items are just at the right level for persons. There are, however, some items that item difficulty estimates fall far below the person ability estimates for the 2015 placement test.

Of the 82 persons measured, all MNSQ ranges of person fit fell between 0.6–1.4, except for persons 72 (Infit MNSQ: 1.46, ZSTD: 2.51), 75 (Outfit MNSQ: 1.57, ZSTD: 2.12), 76 (Outfit MNSQ: 1.72, ZSTD: 2.31), and 80 (Outfit MNSQ: 2.13, ZSTD: 3.89)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.6 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 72, 75, 76, and 80 should be deleted from the list. Regarding item fit of the 27 items measured, all MNSQ ranges fell between 0.6–1.4.



```
MEASURE Person - MAP - Item
            <more>|<rare>
    2             +
                  |T
                  |
                  |  Q7
                  |
                  |
                  |
                  |  Q27
                  |
                  |
                  |
                  |
                  |
                  |  Q16
                  |  Q9
    1            T+S
                  |  Q26
              XX  |  Q19    Q25
               X  |
                  |
        XXXXXXXX  |
               X  |
                  |  Q11    Q24
          XXXXXX S|  Q8
               X  |
                  |  Q5
       XXXXXXXXX  |  Q13
              XX  |  Q18    Q20
      XXXXXXXXXX  |
                  |  Q17
    0          X  +M
             XXX M|  Q3     Q4
               X  |
              XX  |
         XXXXXXX  |  Q23
                  |
                  |  Q22
        XXXXXXXX  |  Q21
               X  |
         XXXXXXX S|
               X  |
              XX  |
            XXXX  |
                  |  Q12    Q2
              XX  |
   -1          X  +S
                  |
                  |
                 T|  Q15
                  |
               X  |
                  |
                  |
                  |  Q6
                  |  Q1     Q14
               X  |
                  |
                  |
                  |  Q10
                  |T
   -2             +
            <less>|<frequent>
```

Figure 5.10. Item and person map for the 2015 placement test (original analysis*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

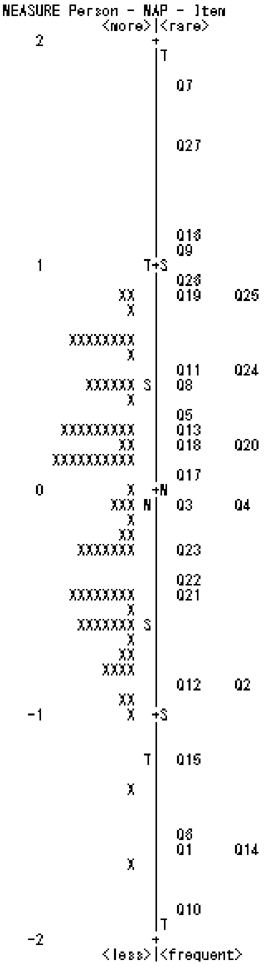
1st revision

As shown in Table 5.3, there were 27 items and 78 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 1st revision.

The separation measure concerning the 1st revision of the 2015 placement test is relatively low (0.67–0.75), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is very low (0.37), which is due to the relatively small number of test items and a narrow ability range of persons measured. The reliability of item is relatively high (0.94) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.81–3.85 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.11 shows that the persons mean (−0.1) is located just below the items mean that is set to 0.00 by default, indicating that, on average, items are approximately appropriate levels of difficulty. Some items difficulty estimates, however, fall far below the person ability estimates for the 2015 placement test.

With regards to the person fit of the 78 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 7 (Outfit MNSQ: 1.60, ZSTD: 2.24) and 72 (Infit MNSQ: 1.47, ZSTD: 2.43)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 7 and 72 should be deleted from the list. All MNSQ ranges of item fit fell between 0.6-1.4 of the 27 items measured.



```
MEASURE   Person - MAP - Item
               <more>|<rare>
    2                +
                     |T
                     |
                     |  Q7
                     |
                     |
                     |
                     |  Q27
                     |
                     |
                     |
                     |
                     |
                     |  Q16
                     |  Q9
    1               T+S
                     |  Q26
                 XX  |  Q19     Q25
                  X  |
                     |
           XXXXXXXX  |
                  X  |
                     |  Q11     Q24
             XXXXXX S|  Q8
                  X  |
                     |  Q5
          XXXXXXXXX  |  Q13
                 XX  |  Q18     Q20
         XXXXXXXXXX  |
                     |  Q17
    0             X  +M
                XXX M|  Q3      Q4
                  X  |
                 XX  |
            XXXXXXX  |  Q23
                     |
                     |  Q22
           XXXXXXXX  |  Q21
                  X  |
            XXXXXXX S|
                  X  |
                 XX  |
               XXXX  |
                     |  Q12     Q2
                 XX  |
   -1             X  +S
                     |
                     |
                    T|  Q15
                     |
                  X  |
                     |
                     |
                     |  Q6
                     |  Q1      Q14
                  X  |
                     |
                     |
                     |  Q10
                     |T
   -2                +
               <less>|<frequent>
```

Figure 5.11. Item and person map for the 2015 placement test (1st revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

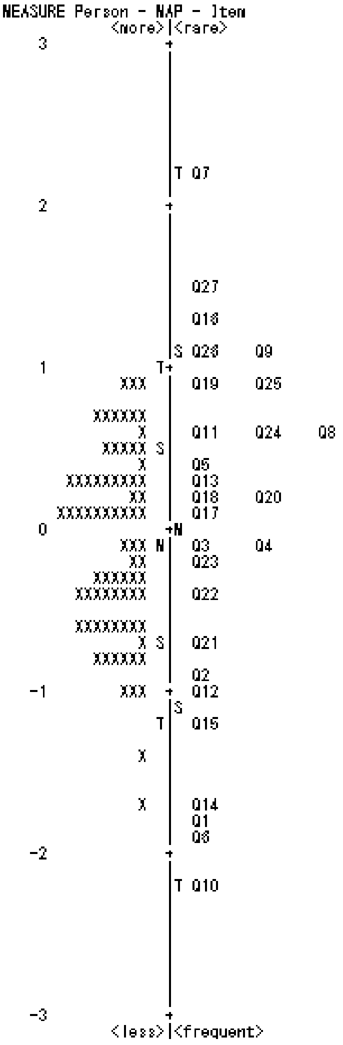
2nd revision

As shown in Table 5.3, there were 27 items and 76 persons in the 2nd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 2nd revision.

The separation measure concerning the 2nd revision of the 2015 placement test is relatively low (0.71–0.78), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is very low (0.39), which is due to the small number of test items. The reliability of item is relatively high (0.94) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.84–3.88 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.12 shows that the persons mean (−0.1) is located just below the items mean that is set to 0.00 by default, indicating that, on average, item difficulty fits that of person ability.

With regards to person fit, of the 76 persons measured, all MNSQ ranges fell between 0.6-1.4. Of the 27 items measured, all MNSQ ranges of item fit fell between 0.4−1.6.



```
MEASURE Person - MAP - Item
             <more>|<rare>
    3              +
                   |
                   |
                   |
                   |
                   |
                   |
                   |
                   |T Q7
                   |
    2              +
                   |
                   |
                   |
                   |
                   |  Q27
                   |
                   |  Q16
                   |
                   |S Q26    Q9
    1             T+
              XXX  |  Q19    Q25
                   |
           XXXXXX  |
                X  |  Q11    Q24    Q8
            XXXXX S|
                X  |  Q5
        XXXXXXXXX  |  Q13
               XX  |  Q18    Q20
       XXXXXXXXXX  |  Q17
    0              +M
              XXX M|  Q3     Q4
               XX  |  Q23
           XXXXXX  |
         XXXXXXXX  |  Q22
                   |
         XXXXXXXX  |
                X S|  Q21
           XXXXXX  |
                   |  Q2
   -1         XXX  +  Q12
                   |S
                  T|  Q15
                   |
                X  |
                   |
                   |
                X  |  Q14
                   |  Q1
                   |  Q6
   -2              +
                   |
                   |T Q10
                   |
                   |
                   |
                   |
                   |
                   |
                   |
   -3              +
             <less>|<frequent>
```

Figure 5.12. Item and person map for the 2015 placement test (2nd revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

5.4.4 Unidimensionality

In this section, the unidimensionality of the final revision (2nd revision) will be described in terms of principal component analysis of residuals. Table 5.9 indicates that observed raw variance explained by measures (24.2%) mildly fits the expected raw variance explained by measure (24.1%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 19.6 % of the 27 items, leaving more than half of variance (75.8%) unaccounted for by the model. This is due to the fact that the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth and fifth respectively. The largest secondary dimension (first contrast) in the 2015 placement test data had strength of 2.7 units (7.7%) while the variance explained by measures was larger at 8.6 units (24.2%), indicating that the secondary contrast does not create multidimensionality.

Table 5.9

Standardized Residual Variance (in Eigen-value Units) for the 2015 Placement Test

	Observed			Expected
Total raw variance in observations	35.6	100%		100%
Raw variance explained by measures	8.6	24.2%		24.1%
Raw variance explained by persons	1.6	4.6%		4.6%
Raw variance explained by items	7.0	19.6%		19.6%
Raw unexplained variance (total)	27.0	75.8%	100%	75.9%
Unexplained variance in 1st contrast	2.7	7.7%	10.1%	
Unexplained variance in 2nd contrast	2.2	6.0%	8.0%	
Unexplained variance in 3rd contrast	1.9	5.3%	7.0%	
Unexplained variance in 4th contrast	1.8	4.9%	6.5%	
Unexplained variance in 5th contrast	1.6	4.5%	6.0%	

Note. This figure is from WINSTEPS^{R} version 3.81.0 output Table 23.0.

5.5.1 The Spring 2015 final test

As shown in Table 5.3, the Spring 2015 final test has undergone four revisions. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each revision in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals.

Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.5.2 Basic feature of the Spring 2015 final test

The number of test items that assess orthographic and phonological awareness was 59 in the Spring 2015 final test. Fifty-six first-year students (2015 entrants) and twenty-two sophomore students (2014 entrants) took one-hour test on: reading comprehension, with ten test items; vocabulary questions, with ten test items; listening comprehension, with five test items; as well as the orthographic and phonological awareness of 59 test items.

As shown in Table 5.2, 43 words (36.4%) of the 118 test item words targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in Japanese pronunciation in the Spring 2015 final test. Sixty-five words (60.1%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining ten words (8.4%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 59 test items, 24 items (40.7%) assess students to identify two NE words, while 13 items (22.2%) assess students' ability to identify a combination of a NE word and a NNE word. Of 118 words, 45 words (38.1%) were loanwords. For example, the first word of item number 39 was "image" which pronounces the grapheme /a/ as [ə/ɪ] in English while Japanese pronounces [e:]. Of 59 test items, four test items (6.7%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the Spring 2015 final test shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.5.3 Revision process of the Spring 2015 final test

As shown in Table 5.10, the Spring 2015 final test has undergone four revisions deleting 16 persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table 5.10

Misfitting Persons and Items by Revisions of the Spring 2015 Final Test

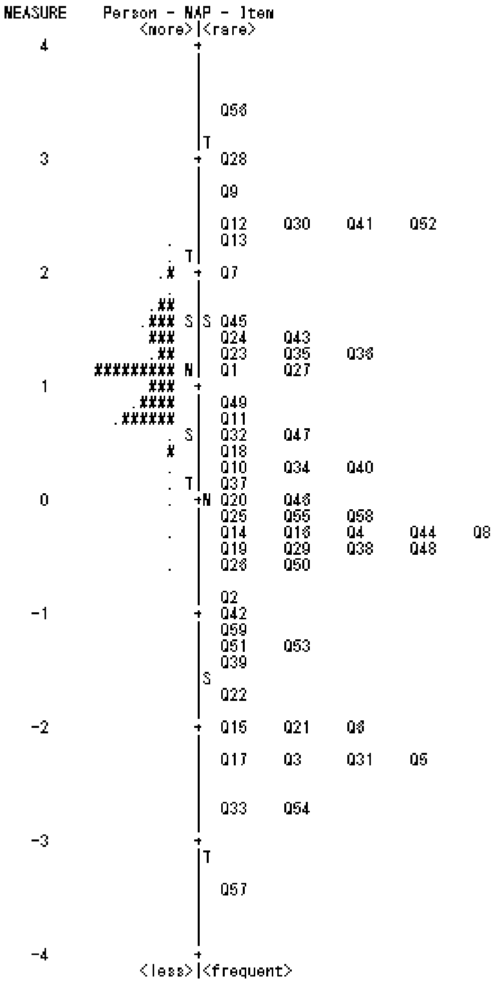
Revision	Misfitting Person	n	Misfitting Item	n
	MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6		MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	
1st Analysis	P4, P16, P17, P27, P48, P57, P68, P80	8	N.A.	0
1st Revision	P29, P31, P62, P64	4	N.A.	0
2nd Revision	P49, P53, P58	3	N.A.	0
3rd Revision	P56	1	N.A.	0
4th Revision	N.A.	0	N.A.	0

Original analysis

As shown in Table 5.3, there were 59 items and 83 persons in the original analysis. WINSTEPS® version 3.81.0 provides the statistical values of the original test items.

The separation measure concerning the original analysis of the Spring 2015 final test is relatively low (1.00–1.09), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.55), which is due to the relatively small number of test items. The reliability of item is relatively high (0.94–0.95) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measures of 4.13–4.14 indicates that the items can be separated into more than four strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.13 shows that the persons mean (1.1) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which items difficulty estimates fall far below the person ability estimates for the Spring 2015 final test.



```
MEASURE    Person - MAP - Item
               <more>|<rare>
    4                +
                     |
                     |
                     |
                     |  Q56
                     |
                     |T
    3                +  Q28
                     |
                     |  Q9
                     |
                     |  Q12   Q30   Q41   Q52
                 .   |  Q13
                 . T |
    2           .#   +  Q7
                 .   |
               .##   |
              .### S |S Q45
               ###   |  Q24   Q43
               .##   |  Q23   Q35   Q36
         ######### M |  Q1    Q27
    1          ###   +
             .####   |  Q49
           .######   |  Q11
                 . S |  Q32   Q47
                 #   |  Q18
                 .   |  Q10   Q34   Q40
                 . T |  Q37
    0            .   +M Q20   Q46
                     |  Q25   Q55   Q58
                 .   |  Q14   Q16   Q4    Q44   Q8
                     |  Q19   Q29   Q38   Q48
                 .   |  Q26   Q50
                     |
                     |  Q2
   -1                +  Q42
                     |  Q59
                     |  Q51   Q53
                     |  Q39
                     |S
                     |  Q22
                     |
   -2                +  Q15   Q21   Q6
                     |
                     |  Q17   Q3    Q31   Q5
                     |
                     |
                     |  Q33   Q54
                     |
   -3                +
                     |T
                     |
                     |  Q57
                     |
                     |
                     |
   -4                +
               <less>|<frequent>
```

Figure 5.13. Item and person map for the Spring 2015 final test (original analysis*)*.

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

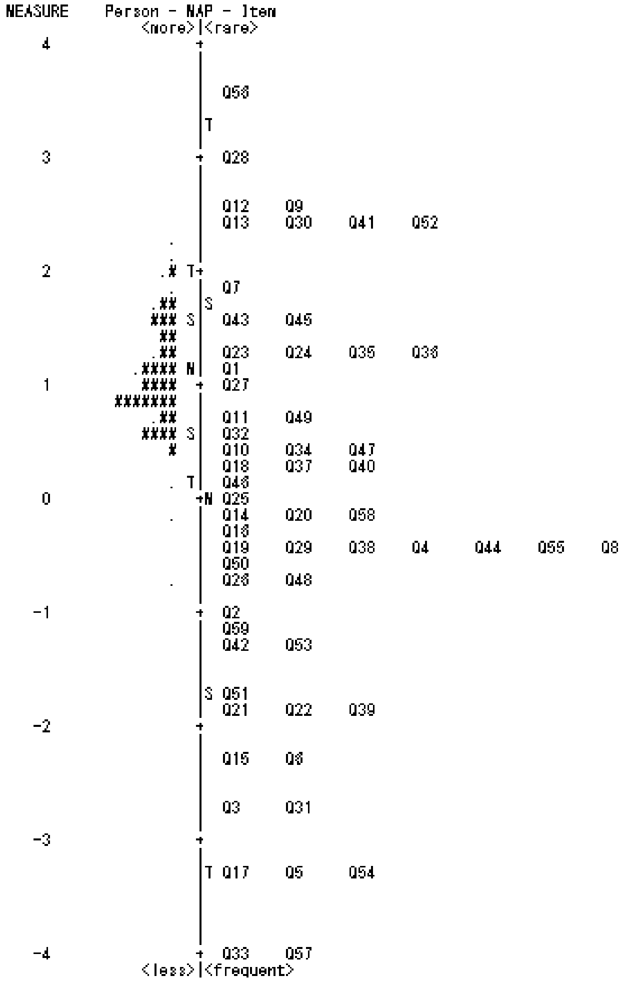
Regarding person fit, of the 83 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 4 (Infit MNSQ: 0.57, ZSTD: −2.92), 16 (Outfit MNSQ: 1.91, ZSTD: 2.18), 17 (Infit MNSQ: 0.57, ZSTD: −2.71), 27 (Outfit MNSQ: 2.14, ZSTD: 2.10), 48 (Outfit MNSQ: 2.16, ZSTD: 2.00), 57 (Infit MNSQ: 1.82, ZSTD: 4.73), 68 (Outfit MNSQ: 1.98, ZSTD: 2.94), and 80 (Infit MNSQ: 1.62, ZSTD: 3.56)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value greater than 1.4 with ZSTDs in excess of 2.0 or MNSQ values less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that the persons 4, 16, 17, 27, 48, 57, 68, and 80 should be deleted from the list. Regarding item fit, all MNSQ ranges fell between 0.6-1.4 of the 59 items measured.

1st revision

As shown in Table 5.3, there were 59 items and 75 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 1st revision.

The separation measure concerning the 1st revision of the Spring 2015 final test is relatively low (0.95–1.02), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.51), which is due to the relatively small number of test items. The reliability of item is relatively high (0.92) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.32–3.33 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.14 shows that the persons mean (1.08) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which items difficulty estimates fall far below the person ability estimates for the Spring 2015 final test.



```
MEASURE    Person - MAP - Item
                <more>|<rare>
    4                 +
                      |
                      |
                      |  Q56
                      |
                      |T
                      |
    3                 +  Q28
                      |
                      |
                      |  Q12     Q9
                      |  Q13     Q30     Q41     Q52
                .     |
                .     |
    2          .#   T+
                .     |  Q7
              .##     |S
              ###   S|  Q43     Q45
               ##     |
              .##     |  Q23     Q24     Q35     Q36
           .####   M|  Q1
    1        ####     +  Q27
          #######     |
              .##     |  Q11     Q49
             ####   S|  Q32
                #     |  Q10     Q34     Q47
                      |  Q18     Q37     Q40
                .   T|  Q46
    0                 +M Q25
                .     |  Q14     Q20     Q58
                      |  Q16
                      |  Q19     Q29     Q38     Q4      Q44     Q55     Q8
                      |  Q50
                .     |  Q26     Q48
                      |
   -1                 +  Q2
                      |  Q59
                      |  Q42     Q53
                      |
                      |
                      |S Q51
                      |  Q21     Q22     Q39
   -2                 +
                      |
                      |  Q15     Q6
                      |
                      |
                      |  Q3      Q31
                      |
   -3                 +
                      |
                      |T Q17     Q5      Q54
                      |
                      |
                      |
                      |
   -4                 +  Q33     Q57
                <less>|<frequent>
```

Figure 5.14. Item and person map for the Spring 2015 final test (1st revision*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 75 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 29 (Outfit MNSQ: 1.80, ZSTD: 2.06), 31 (Outfit MNSQ: 3.25, ZSTD: 2.96), 62 (Outfit MNSQ: 1.90, ZSTD: 2.17), and 64 (Outfit MNSQ: 1.85, ZSTD: 2.22)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 29, 31, 62, and 64 should be deleted from the list. Of the 59 items measured, all MNSQ ranges fell between 0.6-1.4.

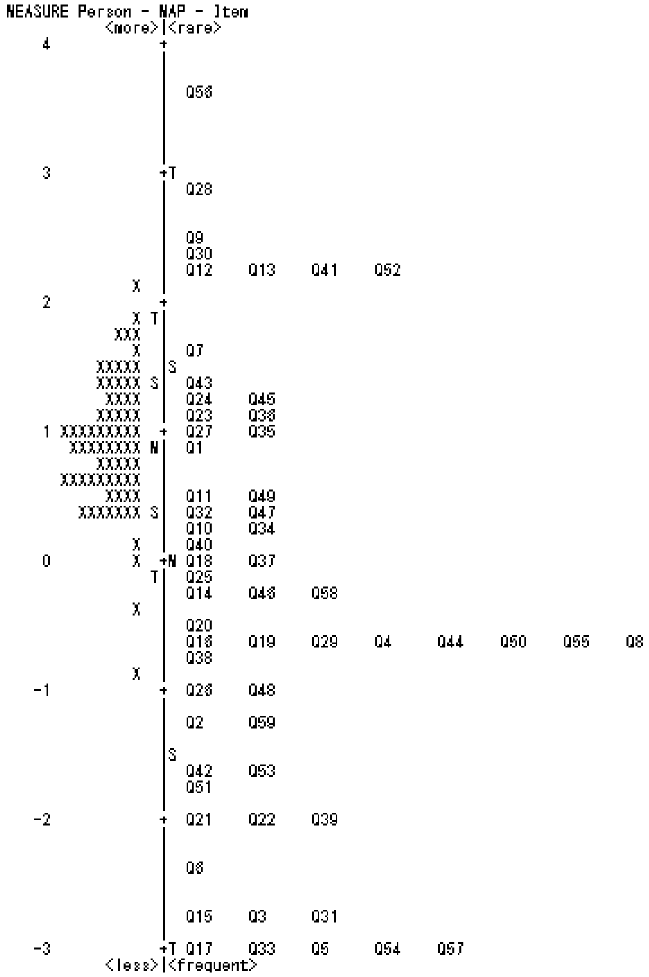
2nd revision

As shown in Table 5.3, there were 59 items and 71 persons in the 2nd revision.

WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 2nd revision.

The separation measure concerning the 2nd revision of the Spring 2015 final test is relatively low (0.96–1.04), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.54), which is due to the relatively small number of test items. The reliability of item is moderately high (0.89–0.90) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.92–2.93 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.15 shows that the persons mean (0.9) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which items difficulty estimates fall far below the person ability estimates for the Spring 2015 final test.



```
MEASURE Person - MAP - Item
           <more>|<rare>
    4            +
                 |
                 |
                 |  Q56
                 |
                 |
                 |
                 |
    3            +T
                 |  Q28
                 |
                 |
                 |  Q9
                 |  Q30
                 |  Q12    Q13    Q41    Q52
              X  |
    2            +
              X T|
            XXX  |
              X  |  Q7
          XXXXX  |S
          XXXXX S|  Q43
           XXXX  |  Q24    Q45
          XXXXX  |  Q23    Q36
    1 XXXXXXXXX  +  Q27    Q35
       XXXXXXXX M|  Q1
          XXXXX  |
      XXXXXXXXX  |
           XXXX  |  Q11    Q49
        XXXXXXX S|  Q32    Q47
                 |  Q10    Q34
              X  |  Q40
    0         X  +M Q18    Q37
                T|  Q25
                 |  Q14    Q46    Q58
              X  |
                 |  Q20
                 |  Q16    Q19    Q29    Q4     Q44    Q50    Q55    Q8
                 |  Q38
              X  |
   -1            +  Q26    Q48
                 |
                 |  Q2     Q59
                 |
                 |S
                 |  Q42    Q53
                 |  Q51
                 |
   -2            +  Q21    Q22    Q39
                 |
                 |
                 |  Q6
                 |
                 |
                 |  Q15    Q3     Q31
                 |
   -3            +T Q17    Q33    Q5     Q54    Q57
           <less>|<frequent>
```

Figure 5.15. Item and person map for the Spring 2015 final test (2nd revision*).*

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 71 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 49 (Outfit MNSQ: 1.75, ZSTD: 2.11), 53 (Outfit MNSQ: 1.96, ZSTD: 2.45), and 58 (Outfit MNSQ: 1.70, ZSTD: 2.15)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 49, 53, and 58 should be deleted from the list. Of the 59 items measured, all MNSQ ranges fell between 0.6-1.4.

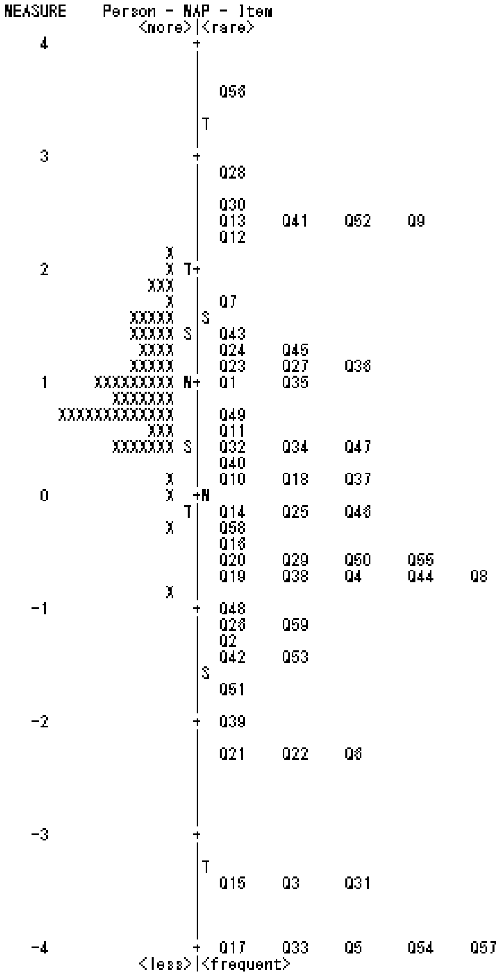
3rd revision

As shown in Table 5.3, there were 59 items and 68 persons in the 3rd revision.

WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 3rd revision.

The separation measure concerning the 3rd revision of the Spring 2015 final test is relatively low (1.01–1.09), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.55), which is due to the relatively small number of test items. The reliability of item is moderately high (0.89) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.88–2.89 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.16 shows that the persons mean (0.95) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which items difficulty estimates fall far below the person ability estimates for the Spring 2015 final test.



```
MEASURE    Person - MAP - Item
                <more>|<rare>
    4                 +
                      |
                      |  Q56
                      |
                      |T
                      |
    3                 +
                      |  Q28
                      |
                      |  Q30
                      |  Q13    Q41    Q52    Q9
                      |  Q12
                  X   |
    2             X  T+
                XXX   |
                  X   |  Q7
              XXXXX   |S
              XXXXX  S|  Q43
               XXXX   |  Q24    Q45
              XXXXX   |  Q23    Q27    Q36
    1     XXXXXXXXX  M+  Q1     Q35
            XXXXXXX   |
      XXXXXXXXXXXXX   |  Q49
                XXX   |  Q11
            XXXXXXX  S|  Q32    Q34    Q47
                      |  Q40
                  X   |  Q10    Q18    Q37
    0             X   +M
                     T|  Q14    Q25    Q46
                  X   |  Q58
                      |  Q16
                      |  Q20    Q29    Q50    Q55
                      |  Q19    Q38    Q4     Q44    Q8
                  X   |
   -1                 +  Q48
                      |  Q26    Q59
                      |  Q2
                      |  Q42    Q53
                      |S
                      |  Q51
                      |
   -2                 +  Q39
                      |
                      |  Q21    Q22    Q6
                      |
                      |
                      |
   -3                 +
                      |
                      |T
                      |  Q15    Q3     Q31
                      |
                      |
   -4                 +  Q17    Q33    Q5     Q54    Q57
                <less>|<frequent>
```

Figure 5.16. Item and person map for the Spring 2015 final test (3rd revision*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

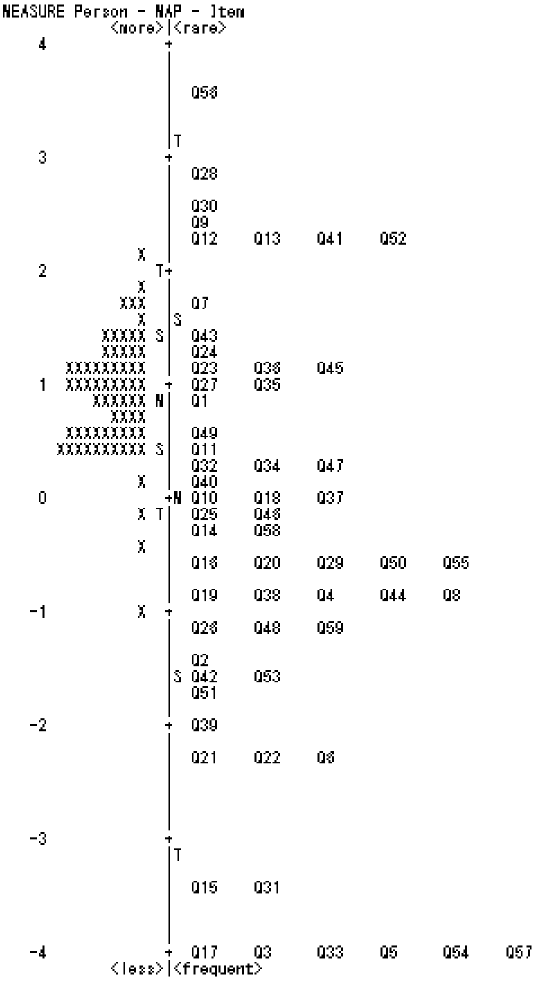
Of the 68 persons measured, all MNSQ ranges fell between 0.6-1.4, except for person 56's outfit MNSQ value (2.23) with a ZSTD value, possibly indicating mismatched person. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 56 should be deleted from the list. Of the 59 items measured, all MNSQ ranges fell between 0.6-1.4.

4th revision

As shown in Table 5.3, there were 59 items and 67 persons in the 4th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 4th revision.

The separation measure concerning the 4th revision of the Spring 2015 final test is relatively low (1.03–1.01), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is relatively low (0.55), which is due to the relatively small number of test items. The reliability of item is moderately high (0.89) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.8 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.17 shows that the persons mean (0.89) is located above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are items which items difficulty estimates fall far below the person ability estimates for the Spring 2015 final test.



```
MEASURE Person - MAP - Item
            <more>|<rare>
    4             +
                  |
                  |
                  |  Q56
                  |
                  |
                  |T
    3             +
                  |  Q28
                  |
                  |  Q30
                  |  Q9
                  |  Q12    Q13    Q41    Q52
               X  |
    2            T+
               X  |
             XXX  |  Q7
               X  |S
           XXXXX S|  Q43
           XXXXX  |  Q24
       XXXXXXXXX  |  Q23    Q36    Q45
    1  XXXXXXXXX  +  Q27    Q35
          XXXXXX M|  Q1
            XXXX  |
       XXXXXXXXX  |  Q49
      XXXXXXXXXX S|  Q11
                  |  Q32    Q34    Q47
               X  |  Q40
    0             +M Q10    Q18    Q37
               X T|  Q25    Q46
                  |  Q14    Q58
               X  |
                  |  Q16    Q20    Q29    Q50    Q55
                  |
                  |  Q19    Q38    Q4     Q44    Q8
   -1          X  +
                  |  Q26    Q48    Q59
                  |
                  |  Q2
                  |S Q42    Q53
                  |  Q51
                  |
   -2             +  Q39
                  |
                  |  Q21    Q22    Q6
                  |
                  |
                  |
                  |
   -3             +
                  |T
                  |
                  |  Q15    Q31
                  |
                  |
                  |
   -4             +  Q17    Q3     Q33    Q5     Q54    Q57
            <less>|<frequent>
```

Figure 5.17. Item and person map for the Spring 2015 final test (4th revision*).*

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 67 persons measured, all MNSQ ranges fell between 0.6-1.4. Of the 59 items measured, all MNSQ ranges fell between 0.6-1.4.

5.5.4 Unidimensionality

In this section, the unidimensionality of the final revision (4th revision) will be described in terms of principal component analysis of residuals. Table 5.11 indicates that observed raw variance explained by measures (33.3%) mildly fits the expected raw variance explained by measure (33.4%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 27.5 % of the 59 items, leaving more than half of variance (66.7%) unaccounted for by the model. This is due to the fact that the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth and fifth respectively. The largest secondary dimension (first contrast) in the Spring 2015 final test data had strength of 3.5 units (4.4%) while the variance explained by measures was larger at 26.4 units (33.3%), indicating that the secondary contrast does not create multidimensionality.

Table 5.11

Standardized Residual Variance (in Eigen-value Units) for the Spring 2015 Final Test

	Observed			Expected
Total raw variance in observations	79.4	100%		100%
Raw variance explained by measures	26.4	33.3%		33.4%
Raw variance explained by persons	4.6	5.8%		5.8%
Raw variance explained by items	21.8	27.5%		27.5%
Raw unexplained variance (total)	53.8	66.7%	100%	66.6%
Unexplained variance in 1st contrast	3.5	4.4%	6.6%	
Unexplained variance in 2nd contrast	3.1	4.0%	5.9%	
Unexplained variance in 3rd contrast	2.9	3.7%	5.5%	
Unexplained variance in 4th contrast	2.5	3.2%	4.7%	

Note. This figure is from WINSTEPS^{R} version 3.81.0 output Table 23.0.

5.6.1 The Fall 2015 final test

As shown in Table 5.3, the Fall 2015 final test has undergone one revision. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each analysis in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals. Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.6.2 Basic feature of the Fall 2015 final test

The number of test items that assess orthographic and phonological awareness was 54 in the Fall 2015 final test. Twenty-nine sophomore students (2014 entrants) took a one-hour test on: reading comprehension, with ten test items; vocabulary questions, with ten test items; listening comprehension, with five test items; as well as the orthographic and phonological awareness of 54 test items.

As shown in Table 5.2, 39 words (36.2%) of the 108 test item words targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in Japanese pronunciation in the Fall 2015 final test. Forty-eight words (44.4%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining 21 words (19.4%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 54 test items, 14 items (25.9%) assess students to identify a combination of a NE word and a NNE word, while 13 items (24.1%) assess students' ability to identify two NE words. Of 108 words, 41 words (37.9%) were loanwords. For example, the first word of item number 12 was "angel" which pronounces the grapheme /a/ as [eɪ] in English while Japanese pronounces [e]. Of 54 test items, 11 test items (20.3%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the Fall 2015 final test shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.6.3 Revision process of the Fall 2015 final test

As shown in Table 5.12, the Fall 2015 final test has undergone one revision deleting four persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table 5.12

Misfitting Persons and Items by Revisions of the Fall 2015 Final Test

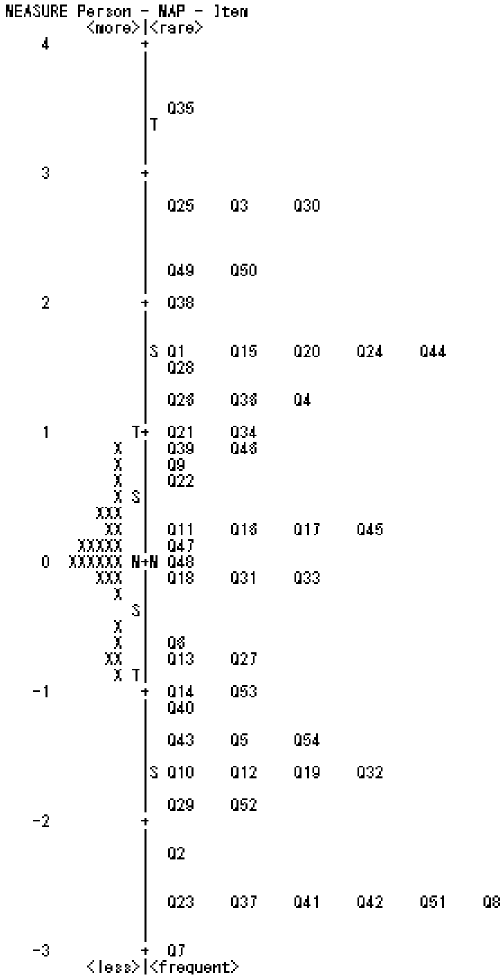
Revision	Misfitting Person	n	Misfitting Item	n
	MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6		MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	
1st Analysis	P9, P10, P19, P23	4	N.A.	0.
1st Revision	N.A.	0	N.A.	0

Original analysis

As shown in Table 5.3, there were 54 items and 29 persons in the original analysis. WINSTEPS® version 3.81.0 provides the statistical values of the original test items.

The separation measure concerning the original analysis of the Fall 2015 final test is relatively low (0.77–0.85), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.41), which is due to a relatively narrow range of person measure compared to that of item measure. The reliability of item is moderately high (0.89) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.83–2.87 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.18 shows that the persons mean (0.04) is located slightly above the items mean that is set to 0.00 by default, indicating that, on average, items are approximately at the right level for persons. Some item difficulty estimates, however, fall far below the person ability estimates for the Fall 2015 final test.



```
MEASURE Person - MAP - Item
           <more>|<rare>
     4           +
                 |
                 |
                 |
                 |  Q35
                 |T
                 |
                 |
     3           +
                 |
                 |  Q25     Q3      Q30
                 |
                 |
                 |
                 |  Q49     Q50
                 |
     2           +  Q38
                 |
                 |
                 |S Q1      Q15     Q20     Q24     Q44
                 |  Q28
                 |
                 |  Q26     Q36     Q4
                 |
     1          T+  Q21     Q34
              X  |  Q39     Q46
              X  |  Q9
              X  |  Q22
              X S|
            XXX  |
             XX  |  Q11     Q16     Q17     Q45
          XXXXX  |  Q47
     0   XXXXXX M+M Q48
            XXX  |  Q18     Q31     Q33
              X  |
                S|
              X  |
              X  |  Q6
             XX  |  Q13     Q27
              X T|
    -1           +  Q14     Q53
                 |  Q40
                 |
                 |  Q43     Q5      Q54
                 |
                 |S Q10     Q12     Q19     Q32
                 |
                 |  Q29     Q52
    -2           +
                 |
                 |  Q2
                 |
                 |
                 |  Q23     Q37     Q41     Q42     Q51     Q8
                 |
                 |
    -3           +  Q7
           <less>|<frequent>
```

Figure 5.18. Item and person map for the Fall 2015 final test (original analysis*)*.

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

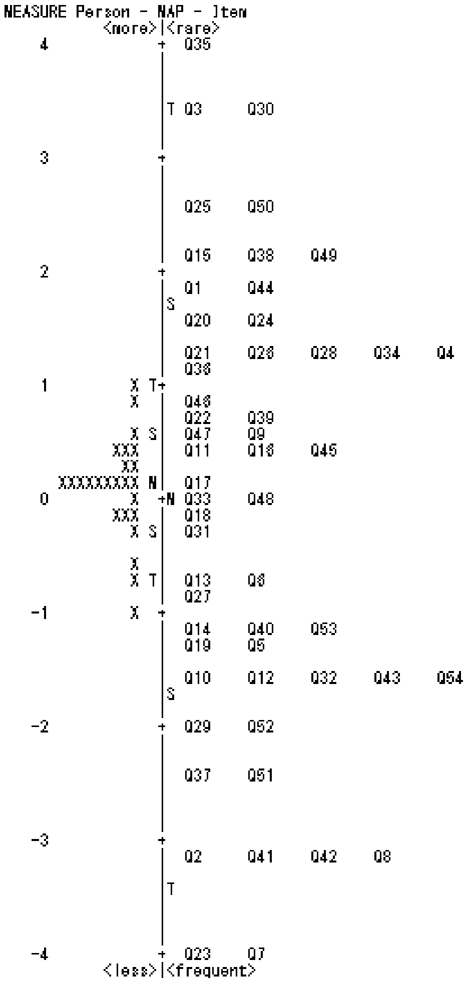
Of the 29 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 9 (Infit MNSQ: 1.52, ZSTD: 2.51), 10 (Infit MNSQ: 1.41, ZSTD: 2.10), 19 (Infit MNSQ: 0.55, ZSTD: −3.04), and 23 (Outfit MNSQ: 2.33, ZSTD: 2.86)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 9, 10, 19, and 23 should be deleted from the list. Of the 54 items measured, all MNSQ ranges fell between 0.6-1.4.

1st revision

As shown in Table 5.3, there were 54 items and 25 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 1st revision.

The separation measure concerning the 1st revision of the Fall 2015 final test is low (0.61–0.68), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is very low (0.3), which is due to the relatively small number of test items. The reliability of item is mdoerately high (0.86–0.87) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.53–2.56 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.19 shows that the persons mean (0.12) is located just above the items mean that is set to 0.00 by default, indicating that, on average, items are at an appropriate level of difficulty for persons. Some item difficulty estimates, however, fall far below the person ability estimates for the Fall 2015 final test.



```
MEASURE Person - MAP - Item
           <more>|<rare>
   4             +  Q35
                 |
                 |
                 |
                 |T Q3      Q30
                 |
                 |
   3             +
                 |
                 |
                 |  Q25     Q50
                 |
                 |
                 |  Q15     Q38     Q49
   2             +
                 |  Q1      Q44
                 |S
                 |  Q20     Q24
                 |
                 |  Q21     Q26     Q28     Q34     Q4
                 |  Q36
   1          X T+
              X  |  Q46
                 |  Q22     Q39
              X S|  Q47     Q9
            XXX  |  Q11     Q16     Q45
             XX  |
      XXXXXXXXX M|  Q17
   0          X  +M Q33     Q48
            XXX  |  Q18
              X S|  Q31
                 |
              X  |
              X T|  Q13     Q6
                 |  Q27
  -1          X  +
                 |  Q14     Q40     Q53
                 |  Q19     Q5
                 |
                 |  Q10     Q12     Q32     Q43     Q54
                 |S
  -2             +  Q29     Q52
                 |
                 |
                 |  Q37     Q51
                 |
                 |
                 |
                 |
  -3             +
                 |  Q2      Q41     Q42     Q8
                 |
                 |T
                 |
                 |
                 |
  -4             +  Q23     Q7
           <less>|<frequent>
```

Figure 5.19. Item and person map for the Fall 2015 final test (1st revision*)*.

Note. This figure is WINSTEPS[R] version 3.81.0 output Table 1.2. Crosses on left side represent examinees, while letter and number combinations on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2- standard deviations respectively.

Of the 25 persons measured, all MNSQ ranges fell between 0.6-1.4. Of the 54 items measured, all MNSQ ranges fell between 0.6-1.4.

5.6.4 Unidimensionality

In this section, the unidimensionality of the final revision (2nd revision) will be described in terms of principal component analysis of residuals. Table 5.13 indicates that observed raw variance explained by measures (38.8%) fits the expected raw variance explained by measure (38.8%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 35 % of the 54 items, leaving more than half of variance (61.2%) unaccounted for by the model. This is due to the fact that the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth and fifth respectively. The largest secondary dimension (first contrast) in the Fall 2015 final test data had strength of 5.6 units (6.8%) while the variance explained by measures was larger at 32.3 units (38.8%), indicating that the secondary contrast does not create multidimensionality.

Table 5.13

Standardized Residual Variance (in Eigen-value Units) for the Fall 2015 Final Test

	Observed			Expected
Total raw variance in observations	83.3	100%		100%
Raw variance explained by measures	32.3	38.8%		38.8%
Raw variance explained by persons	3.2	3.8%		3.8%
Raw variance explained by items	29.1	35.0%		34.9%
Raw unexplained variance (total)	51.0	61.2%	100%	61.2%
Unexplained variance in 1st contrast	5.6	6.8%	11.0%	
Unexplained variance in 2nd contrast	4.7	5.6%	9.2%	
Unexplained variance in 3rd contrast	4.4	5.3%	8.7%	
Unexplained variance in 4th contrast	4.2	5.0%	8.2%	
Unexplained variance in 5th contrast	4.0	4.8%	7.8%	

Note. This figure is from WINSTEPS^{R} version 3.81.0 output Table 23.0.

5.7.1 Equating test items

To organize all of the test items into the same frame of reference, two test were deleted from the list: the Fall 2014 final test (14FF) and the 2015 placement test (15P). After the Rasch analysis, 14FF had fewer than five common items, which aligns with Linacre's (2013) recommendation that more than five items be used as common items. The common persons between 15P and the Spring 2015 final test (15SF) also numbered fewer than five after analysis.

To equate the Spring 2014 final test (14SF), 15SF, and the Fall 2015 final test (15FF), a common item equating with the Rasch measurement framework was performed using WINSTEPS' 3.92.1

5.7.2 Materials and procedure

As Table 5.1 illustrates, the 14SF, 15SF, and 15FF contained 16 common items. Since there were three tests, concurrent equating, in which tests are analysed together as one dataset, was used to calibrate the tests. Firstly, Differential Test Functioning

(DTF) anaylsis was performed to investigate whether the test items for two groups of test takers functioned similarly. Secondly, outlier items were removed until all items functioned similarly for all groups. Lastly, all items, including common and original ones, were combined to align all measures in the same frame of reference.

5.7.3 First analysis

Figure 5.20 displays a scatterplot for the 14SF and 15SF item difficulties. The correlation coefficient between the 14SF and 15SF item difficulties was 0.954, and the disattenuated correlation was 1. Item 3 and 13 seemed to function differently and were therefore removed to improve the correlation coefficient between the 14SF and 15SF item difficulties.

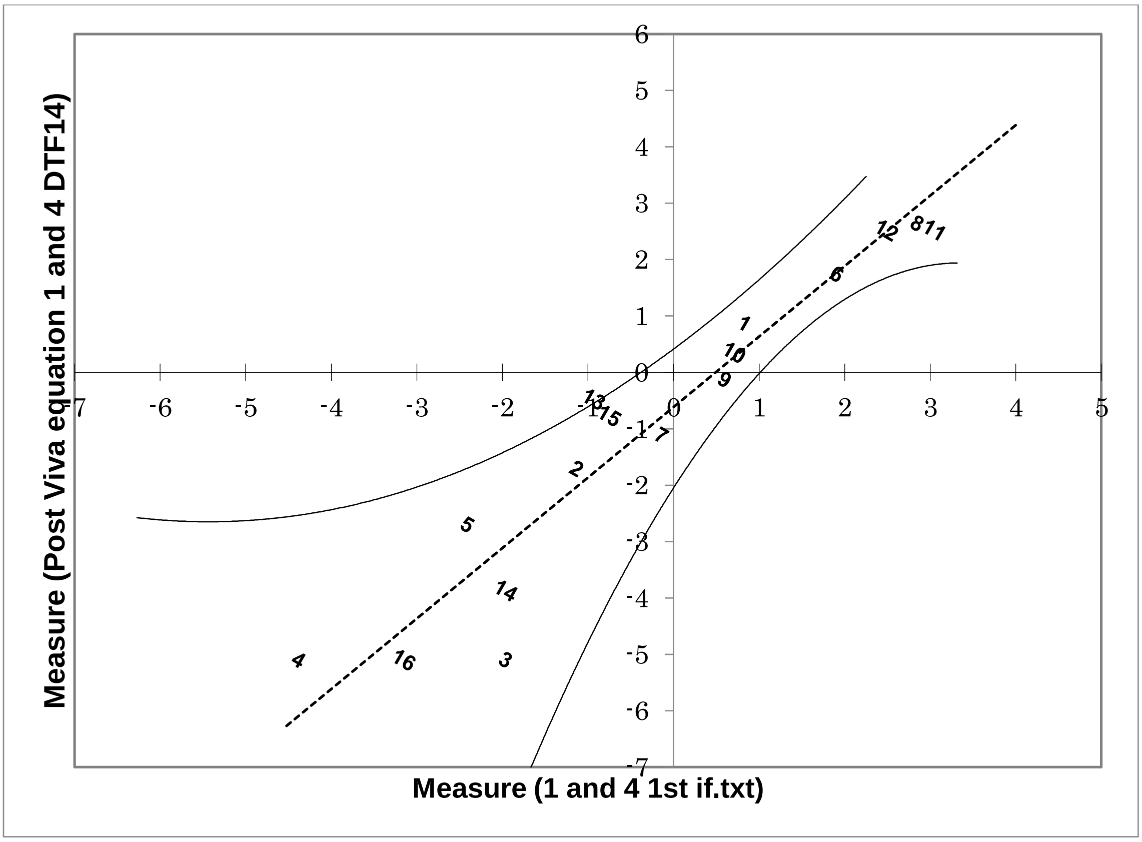


Figure 5.20. Differential Test Functioning for 14SF and 15SF (1st)

Figure 5.21 displays a scatterplot for the 14SF and 15FF item difficulties. The correlation coefficient between the 14SF and 15FF item difficulties was 0.90, and the disattenuated correlation was 1. Item 3, 8, and 13 seemed to function differently and were therefore removed to improve the correlation coefficient between the 14SF and

15FF item difficulties.

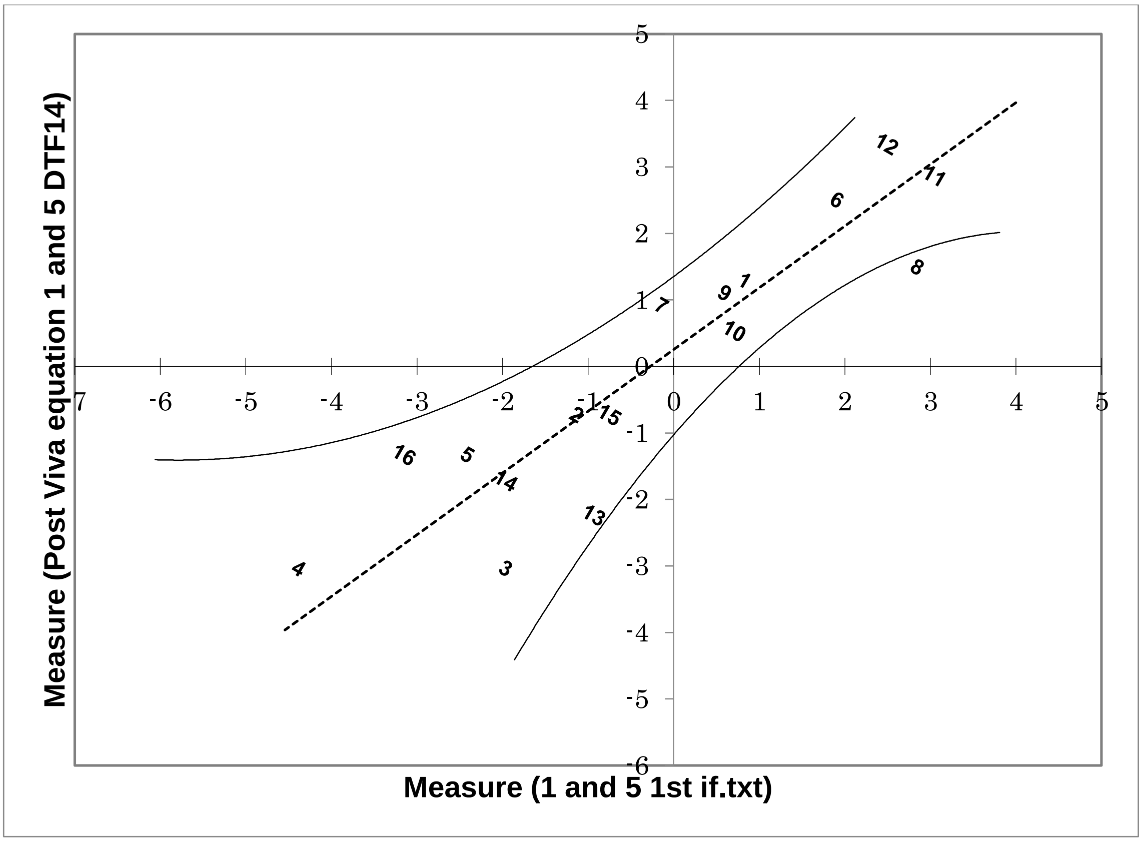


Figure 5.21. Differential Test Functioning for 14SF and 15FF (1st)

Figure 5.22 displays a scatterplot for the 15SF and 15FF item difficulties. The correlation coefficient between the 15SF and 15FF item difficulties was 0.88, and the disattenuated correlation was 1. Item 7, 8, 13, and 16 seemed to function differently and were therefore removed to improve the correlation coefficient between the 15SF and 15FF item difficulties.

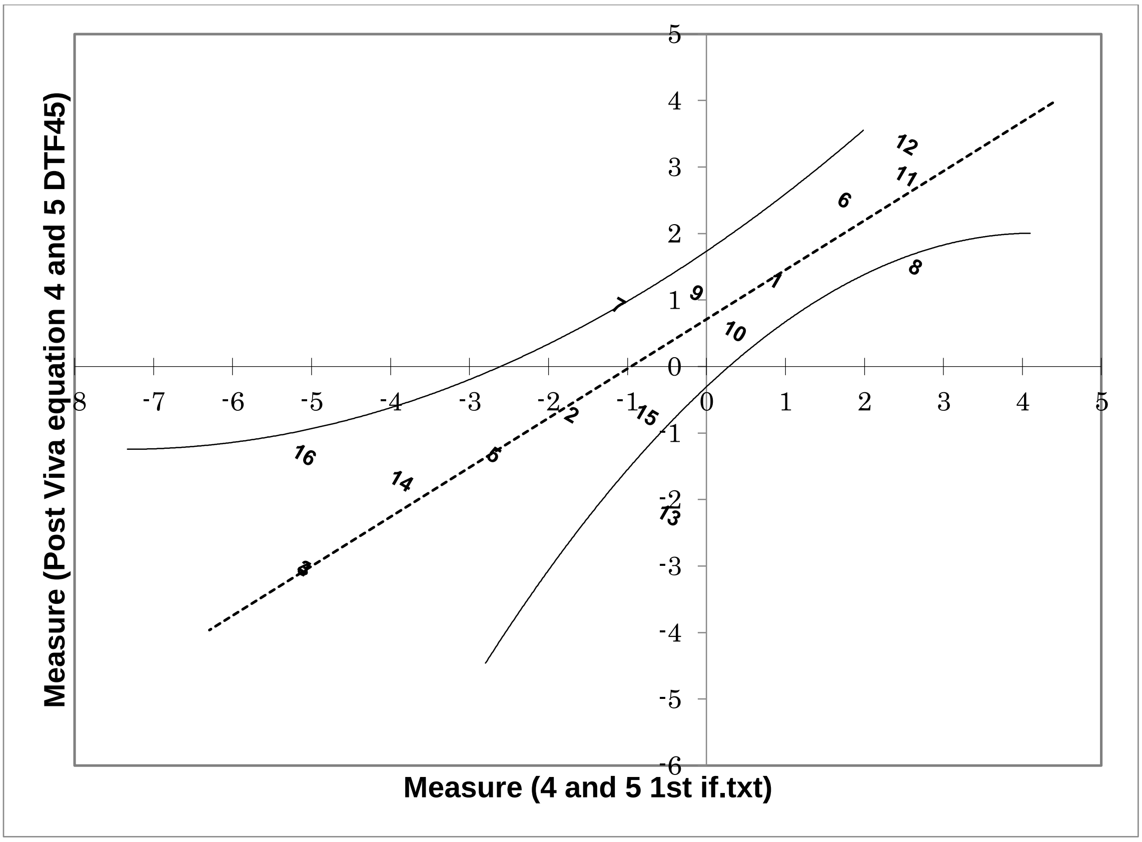


Figure 5.22. Differential Test Functioning for 15SF and 15FF (1st)

5.7.4 Second analysis

Figure 5.23 displays a scatterplot for the 14SF and 15SF item difficulties. The correlation coefficient between the 14SF and 15SF item difficulties was 0.98, and the disattenuated correlation was 1. The similarity of the measures of ability in 14SF and 15SF indicates that all test items function the same way for 14SF and 15SF.

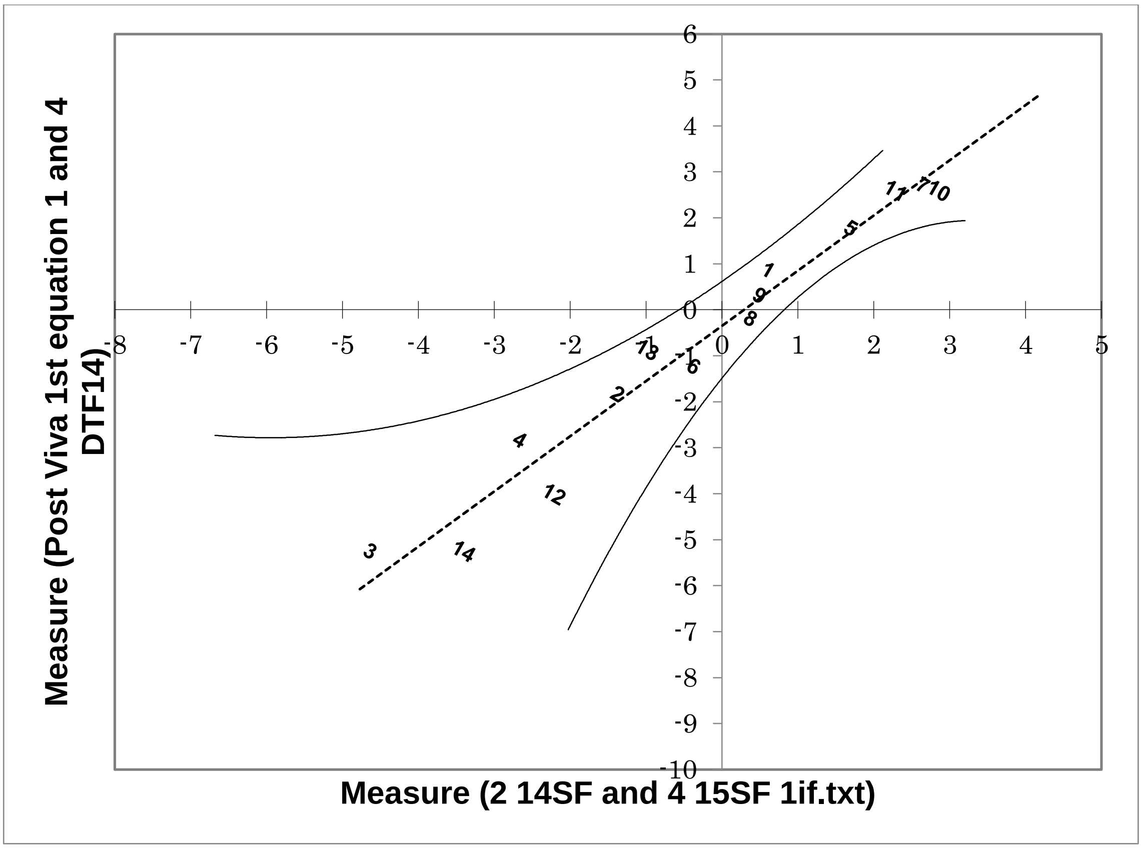


Figure 5.23. Differential Test Functioning for 14SF and 15SF (2nd)

Figure 5.24 displays a scatterplot for the 14SF and 15FF item difficulties. The correlation coefficient between the 14SF and 15FF item difficulties was 0.97, and the disattenuated correlation was 1. The similarity of the measures of ability in 14SF and 15FF indicates that all test items function the same way for 14SF and 15FF.

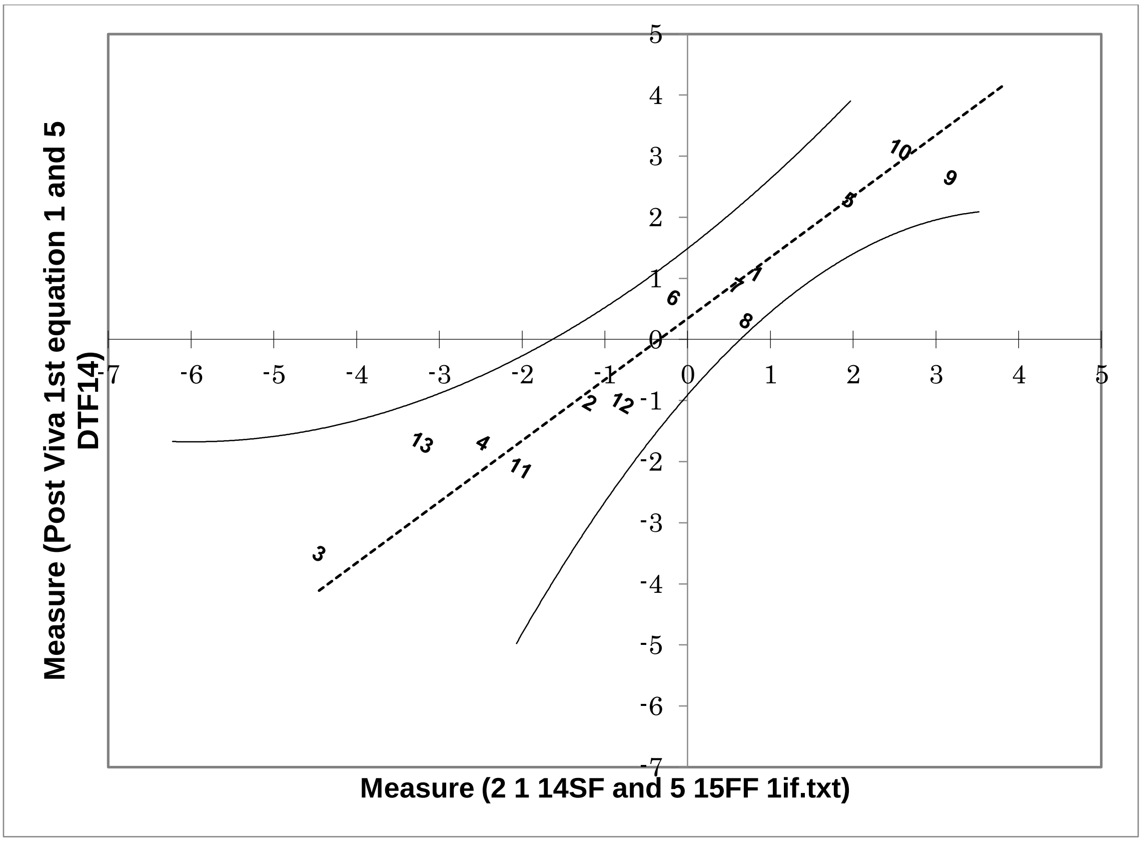


Figure 5.24. Differential Test Functioning for A and B (2nd)

Figure 5.25 displays a scatterplot for the 15SF and 15FF item difficulties. The correlation coefficient between the 15SF and 15FF item difficulties was 0.98, and the disattenuated correlation was 1. The similarity of the measures of ability in 15SF and 15FF indicates that all test items function the same way for 15SF and 15FF.

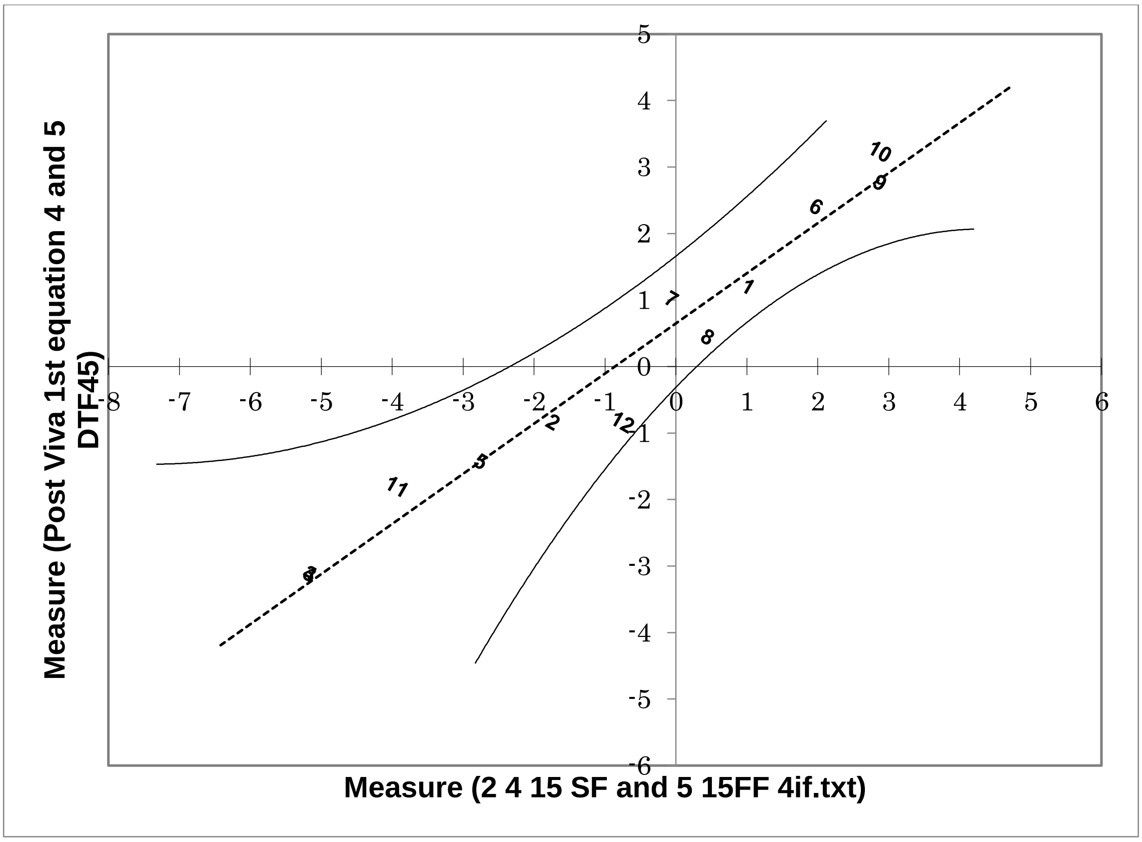


Figure 5.25. Differential Test Functioning for 15SF and 15FF (2nd)

The eleven items (1, 2, 4, 5, 6, 9, 10, 11, 12, 14, and 15) were deemed satisfactory as anchors for concurrent equating.

5.8.1 The equated tests

As shown in Table 5.14, the equated tests has undergone eight revisions. The section will first describe the basic features of the test including test takers, the type of test, and orthographic and phonological features of test items. The section will then describe each analysis in terms of reliability and test separation, test targeting, and item and person fit. In the final revision, the paper will describe the unidimensionality of the test in terms of the principle component analysis of residuals. Moreover, the relationships between the item difficulty measures and orthographic and phonological features are discussed.

5.8.2 Basic feature of the equated tests

The number of test items that assess orthographic and phonological awareness was 90 in the equated tests. Of the 180 test item words, 68 words (38%) targeted to assess test takers' ability to identify a grapheme with a phoneme that does not exist in

Japanese pronunciation in the equated tests. Eighty-eight words (49%) required test takers to identify a grapheme pronounced in a phoneme that exists in Japanese but that does not follow a one-to-one grapheme–phoneme relationship in Japanese. The remaining 24 words (13%) targeted the test takers' ability to identify a grapheme with a phoneme that follows a Japanese one-to-one grapheme–phoneme relationship. Of 90 test items, 22 items (24.4%) assess students to identify a combination of two NNE words, while 21 items (23.3%) assess students' ability to identify two NE words. Of 180 words, 63 words (35%) were loanwords. Of 90 test items, nine test items (10%) aimed to assess English consonants while others assess English vowels. Results of the Rasch analysis of the equated tests shown below reveal the statistical characteristics of the test and the problems needed to be resolved in its revision.

5.8.3 Revision process of the equated tests

As shown in Table 5.14, the equated tests has undergone eight revision deleting 44 persons from the original analysis. This section shows the features of each analysis with regards to reliability and test separation, test targeting, and item and person fit.

Table 5.14

Misfitting Persons and Items by Revisions of the Equated Tests

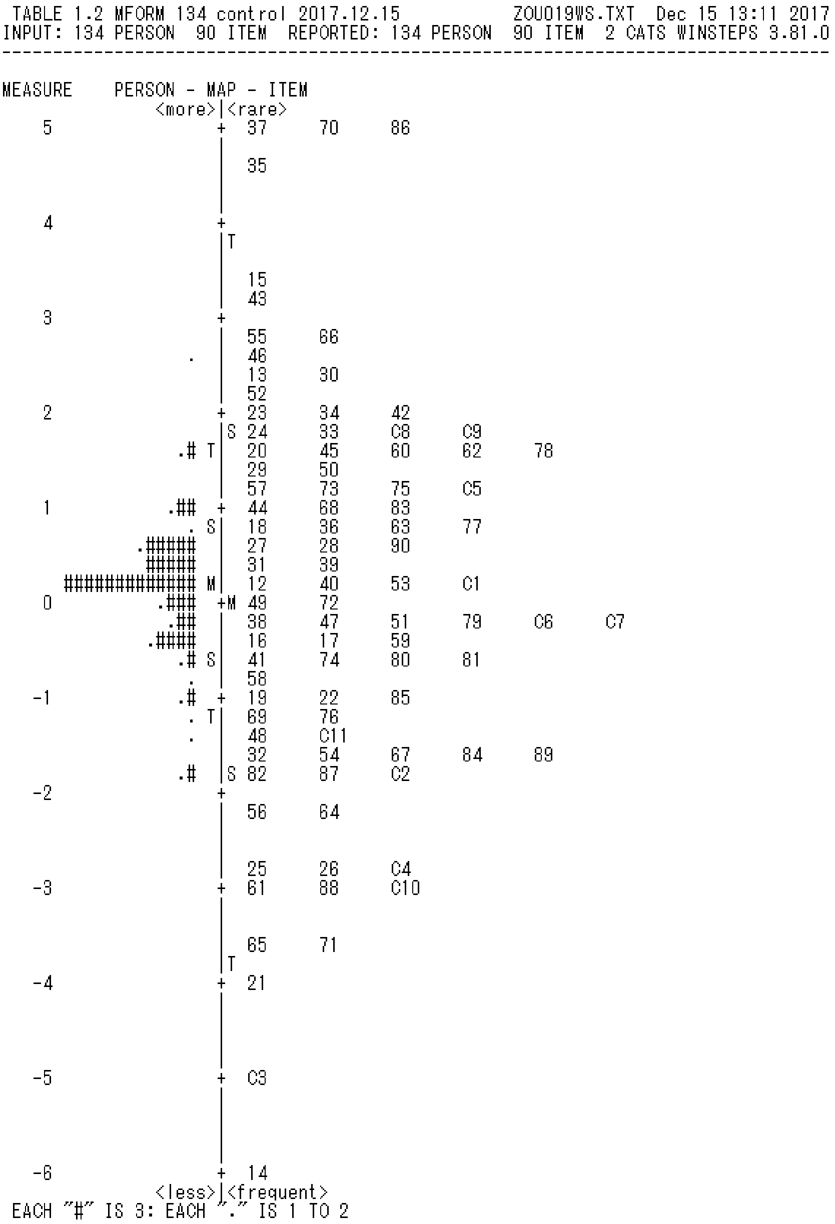
Revision	Misfitting Person MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	n	Misfitting Item MNSQ>2.0 with ZSTD>1.4 or −2.0<MNSQ with ZSTD<0.6	n
1st Analysis	P12, P64, P68, P75, P81, P93, P96, P100, P111, p121	10	N.A.	0.
1st Revision	P25, P59, P67, P76, P87, P108, P127, P132	8	N.A.	0
2nd Revision	P42, P46, P69, P94, P109, P114, P117	7	N.A.	0.
3rd Revision	P22, P63, P95, P123, P134	5	N.A.	0
4th Revision	P30, P41, P48	3	N.A.	0
5th revision	P4, P15, P37, P74, P78, P82, P119, P126	8	N.A.	0
6th revision	P27, P124	2	N.A.	0
7th revision	P112	1	N.A.	0
8th revision	N.A.	0	N.A.	0

Original analysis

As shown in Table 5.14, there were 90 items and 134 persons in the original analysis. WINSTEPS® version 3.81.0 provides the statistical values of the original test items.

The separation measure concerning the original analysis of the equated tests is relatively low (0.67–0.80), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.31), which is due to the relatively small number of test items. The reliability of item is relatively high (0.91) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 3.21–3.24 indicates that the items can be separated into more than three strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.26 shows that the persons mean (0.14) is located slightly above the items mean that is set to 0.00 by default, indicating that, on average, items are approximately at the right level for persons. Some item difficulty estimates, however, fall far below the person ability estimates for the equated tests.



```
 TABLE 1.2 MFORM 134 control 2017.12.15           ZOU019WS.TXT  Dec 15 13:11 2017
INPUT: 134 PERSON  90 ITEM  REPORTED: 134 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5                +  37     70     86
                     |
                     |  35
                     |
                     |
    4                +
                     |T
                     |
                     |  15
                     |  43
    3                +
                     |  55     66
                  .  |  46
                     |  13     30
                     |  52
    2                +  23     34     42
                     |S 24     33     C8     C9
                 .# T|  20     45     60     62     78
                     |  29     50
                     |  57     73     75     C5
    1           .##  +  44     68     83
                  . S|  18     36     63     77
             .#####  |  27     28     90
              #####  |  31     39
      ############# M|  12     40     53     C1
    0          .###  +M 49     72
                .##  |  38     47     51     79     C6     C7
              .####  |  16     17     59
                 .# S|  41     74     80     81
                  .  |  58
   -1            .#  +  19     22     85
                  . T|  69     76
                  .  |  48     C11
                     |  32     54     67     84     89
                 .#  |S 82     87     C2
   -2                +
                     |  56     64
                     |
                     |
                     |  25     26     C4
   -3                +  61     88     C10
                     |
                     |
                     |  65     71
                     |T
   -4                +  21
                     |
                     |
                     |
                     |
   -5                +  C3
                     |
                     |
                     |
                     |
   -6                +  14
               <less>|<frequent>
 EACH "#" IS 3: EACH "." IS 1 TO 2
```

Figure 5.26. Item and person map for the equated tests (original analysis*)*.

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

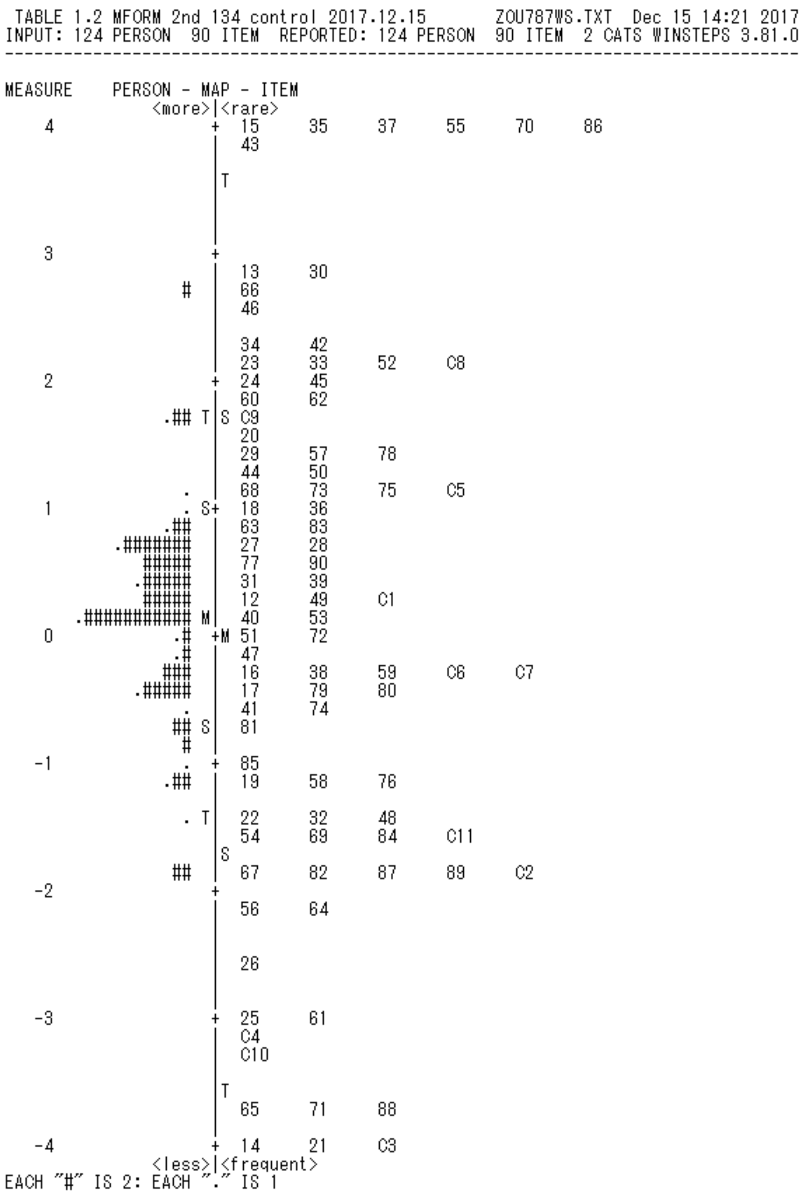
Of the 134 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 12 (Infit MNSQ: 2.20, ZSTD: 2.37), 64 (Infit MNSQ: 0.56, ZSTD: -3.08), 68 (Infit MNSQ: 1.65, ZSTD: 3.33), 75 (Infit MNSQ: 1.65, ZSTD: 3.35), 81 (Infit MNSQ: 1.44, ZSTD: 2.38), 93 (Infit MNSQ: 1.46, ZSTD: 2.42), 96 (Infit MNSQ: 1.77, ZSTD: 3.86), 100 (Infit MNSQ: 1.54, ZSTD: 2.84), 111 (Outfit MNSQ: 4.50, ZSTD: 4.77), and 121 (Infit MNSQ: 1.45, ZSTD: 2.18)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 12, 64, 68, 75, 81, 93, 96, 100, 111, and 121 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.

1st revision

As shown in Table 5.14, there were 90 items and 124 persons in the 1st revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 1st revision.

The separation measure concerning the 1st revision of the equated tests is low (0.79–0.93), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.38), which is due to the relatively small number of test items. The reliability of item is relatively high (0.90) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.95–2.97 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.27 shows that the persons mean (0.14) is located just above the items mean that is set to 0.00 by default, indicating that, on average, items are at an appropriate level of difficulty for persons. Some item difficulty estimates, however, fall far below/above the person ability estimates for the equated tests.



```
 TABLE 1.2 MFORM 2nd 134 control 2017.12.15       ZOU787WS.TXT  Dec 15 14:21 2017
INPUT: 124 PERSON  90 ITEM  REPORTED: 124 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
---------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +  15     35     37     55     70     86
                     |  43
                     |
                     |T
                     |
                     |
                     |
    3                +
                     |  13     30
                  #  |  66
                     |  46
                     |
                     |  34     42
                     |  23     33     52     C8
    2                +  24     45
                     |  60     62
                .## T|S C9
                     |  20
                     |  29     57     78
                     |  44     50
                  .  |  68     73     75     C5
    1             . S+  18     36
                .##  |  63     83
           .#######  |  27     28
              #####  |  77     90
             .#####  |  31     39
              #####  |  12     49     C1
       .########### M|  40     53
    0            .#  +M 51     72
                 .#  |  47
                ###  |  16     38     59     C6     C7
             .#####  |  17     79     80
                  .  |  41     74
                 ## S|  81
                  #  |
   -1             .  +  85
                .##  |  19     58     76
                     |
                  . T|  22     32     48
                     |  54     69     84     C11
                     |S
                 ##  |  67     82     87     89     C2
   -2                +
                     |  56     64
                     |
                     |
                     |  26
                     |
                     |
   -3                +  25     61
                     |  C4
                     |  C10
                     |
                     |T
                     |  65     71     88
                     |
   -4                +  14     21     C3
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.27. Item and person map for the equated tests (1st revision*)*.

Note. This figure is WINSTEPS$^{\text{R}}$ version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

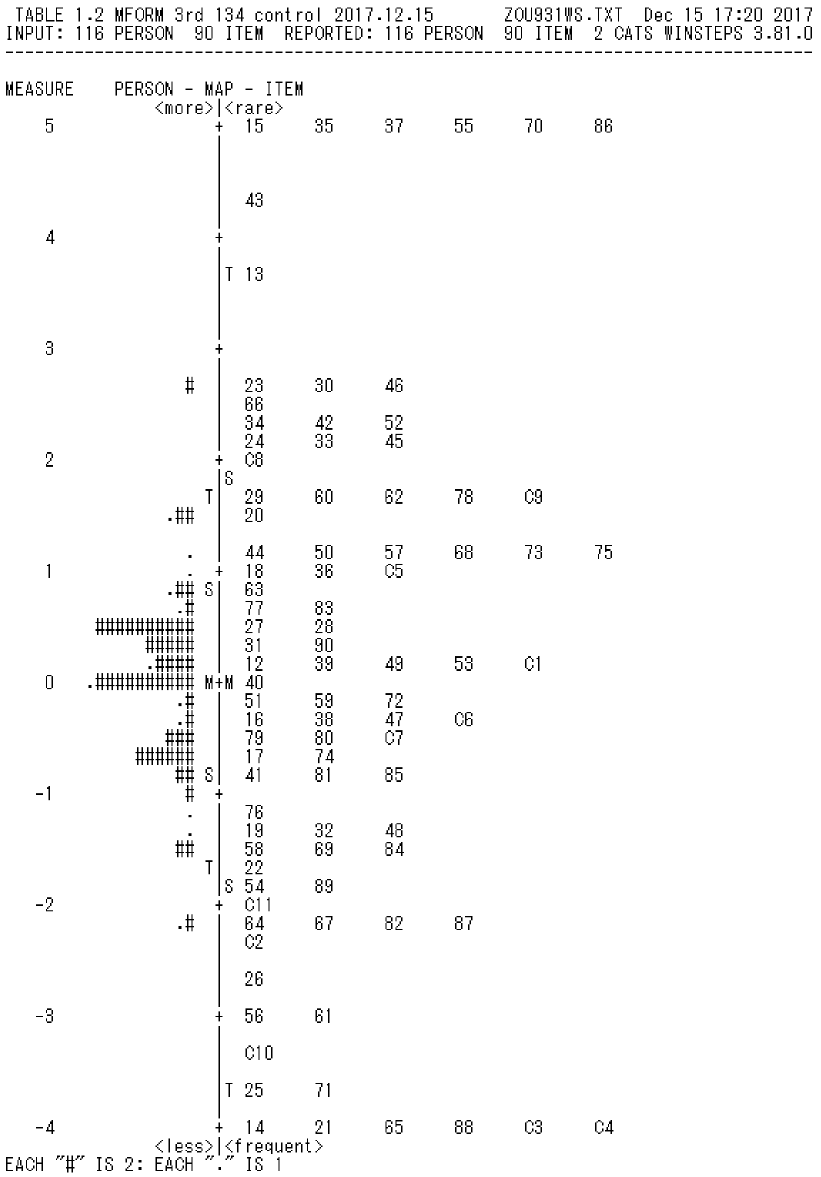
Of the 124 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 25 (Infit MNSQ: 2.28, ZSTD: 2.02), 59 (Infit MNSQ: 1.41, ZSTD: 2.19), 67 (Infit MNSQ: 1.45, ZSTD: 2.32), 76 (Outfit MNSQ: 1.97, ZSTD: 2.32), 87 (Outfit MNSQ: 1.93, ZSTD: 2.37), 108 (Infit MNSQ: 1.50, ZSTD: 2.54), 127 (Outfit MNSQ: 2.51, ZSTD: 2.94), and 132 (Infit MNSQ: 1.41, ZSTD: 1.98)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 25, 59, 67, 76, 87, 108, 127, and 132 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.

2nd revision

As shown in Table 5.14, there were 90 items and 116 persons in the 2nd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of the 2nd revision.

The separation measure concerning the 2nd revision of the equated tests is relatively low (.77–0.91), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is very low (0.37), which is due to the relatively small number of test items and a narrow range of person measure. The reliability of item is relatively high (0.89) which indicates that, if the items were given to other comparable groups of test takers, there is a moderate probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.86–2.87 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.28 shows that the persons mean (0.03) is located just above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively easy for persons. Moreover, there are several item difficulty estimates fall above/below the person ability estimates for the equated tests.



```
 TABLE 1.2 MFORM 3rd 134 control 2017.12.15       ZOU931WS.TXT  Dec 15 17:20 2017
INPUT: 116 PERSON  90 ITEM  REPORTED: 116 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
    5               +  15     35     37     55     70     86
                    |
                    |
                    |
                    |  43
                    |
    4               +
                    |
                    |T 13
                    |
                    |
                    |
    3               +
                    |
                 #  |  23     30     46
                    |  66
                    |  34     42     52
                    |  24     33     45
    2               +  C8
                    |S
                   T|  29     60     62     78     C9
               .##  |  20
                    |
                 .  |  44     50     57     68     73     75
    1            .  +  18     36     C5
               .## S|  63
                .#  |  77     83
        ##########  |  27     28
             #####  |  31     90
            .####   |  12     39     49     53     C1
    0  .########## M+M 40
                .#  |  51     59     72
                .#  |  16     38     47     C6
               ###  |  79     80     C7
            ######  |  17     74
                ## S|  41     81     85
   -1            #  +
                 .  |  76
                 .  |  19     32     48
                ##  |  58     69     84
                   T|  22
                    |S 54     89
   -2               +  C11
                .#  |  64     67     82     87
                    |  C2
                    |
                    |  26
                    |
   -3               +  56     61
                    |
                    |  C10
                    |
                    |T 25     71
                    |
   -4               +  14     21     65     88     C3     C4
              <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.28. Item and person map for the equated tests (2nd revision*).*

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

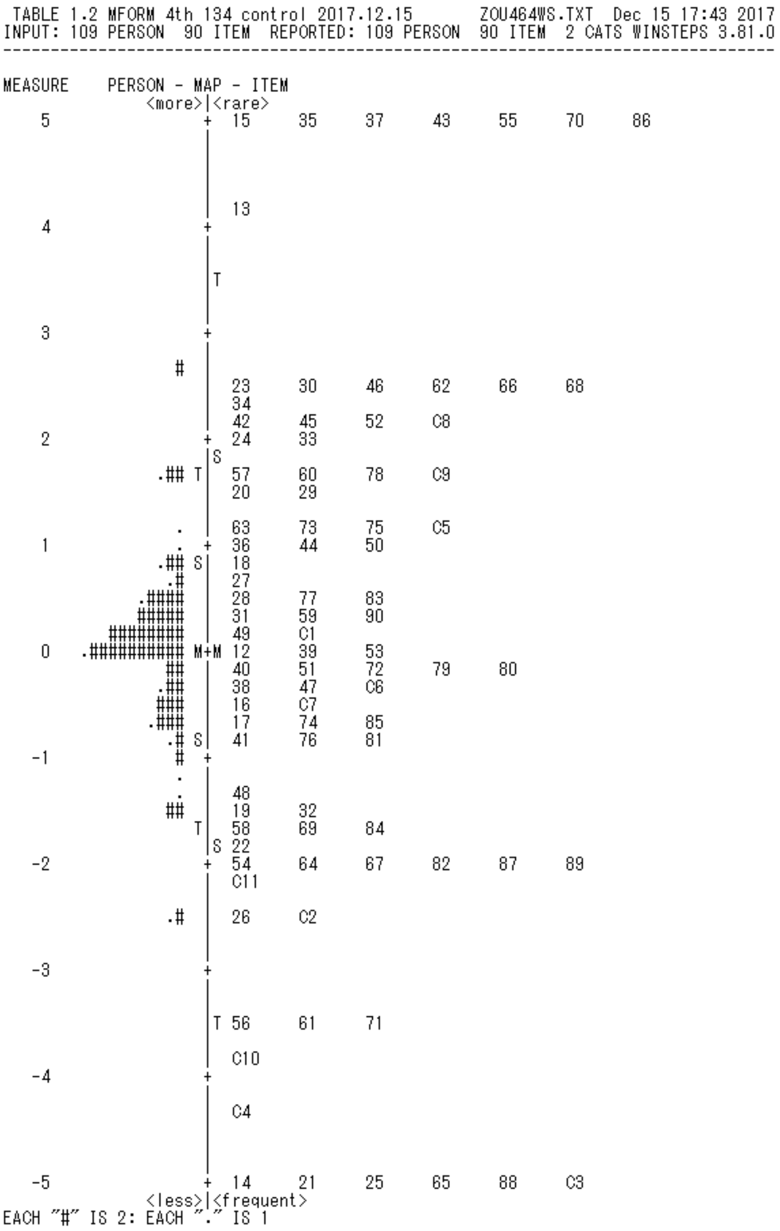
Of the 116 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 42 (Infit MNSQ: 2.17, ZSTD: 2.07), 46 (Outfit MNSQ: 2.36, ZSTD: 2.56), 69 (Infit MNSQ: 1.46, ZSTD: 2.35), 94 (Infit MNSQ: 1.45, ZSTD: 2.19), 109 (Outfit MNSQ: 2.00, ZSTD: 2.03), 114 (Infit MNSQ: 1.49, ZSTD: 2.26), and 117 (Outfit MNSQ: 2.03, ZSTD: 2.51)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 42, 46, 69, 94, 109, 114, and 117 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.

3rd revision

As shown in Table 5.14, there were 90 items and 109 persons in the 3rd revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 3rd revision.

The separation measure concerning the 3rd revision of the equated tests is relatively low (0.77–0.92), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.37), which is due to the relatively small number of test items. The reliability of item is relatively high (0.88) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.73–2.75 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.29 shows that the persons mean (0.02) is located just above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively at the right level for persons. Moreover, there are some items which item difficulty estimates fall above/below the person ability estimates for the 3rd revision of the equated tests.



```
 TABLE 1.2 MFORM 4th 134 control 2017.12.15       ZOU464WS.TXT  Dec 15 17:43 2017
INPUT: 109 PERSON  90 ITEM  REPORTED: 109 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    5                +  15     35     37     43     55     70     86
                     |
                     |
                     |
                     |
                     |  13
    4                +
                     |
                     |
                     |T
                     |
                     |
    3                +
                     |
                  #  |
                     |  23     30     46     62     66     68
                     |  34
                     |  42     45     52     C8
    2                +  24     33
                     |S
                .## T|  57     60     78     C9
                     |  20     29
                     |
                  .  |  63     73     75     C5
    1             .  +  36     44     50
                .## S|  18
                 .#  |  27
              .####  |  28     77     83
              #####  |  31     59     90
           ########  |  49     C1
    0   .########## M+M 12     39     53
                 ##  |  40     51     72     79     80
                .##  |  38     47     C6
                ###  |  16     C7
               .###  |  17     74     85
                 .# S|  41     76     81
   -1             #  +
                  .  |
                  .  |  48
                 ##  |  19     32
                    T|  58     69     84
                     |S 22
   -2                +  54     64     67     82     87     89
                     |  C11
                     |
                 .#  |  26     C2
                     |
                     |
   -3                +
                     |
                     |
                     |T 56     61     71
                     |
                     |  C10
   -4                +
                     |
                     |  C4
                     |
                     |
                     |
   -5                +  14     21     25     65     88     C3
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.29. Item and person map for the equated tests (3rd revision*).*

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

Of the 109 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 22 (Infit MNSQ: 2.53, ZSTD: 2.19), 63 (Infit MNSQ: 1.42, ZSTD: 2.16), 95 (Infit MNSQ: 1.47, ZSTD: 2.27), 123 (Outfit MNSQ: 1.93, ZSTD: 2.04), and 134 (Outfit MNSQ: 2.14, ZSTD: 2.40)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 22, 63, 95, 123, and 134 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.

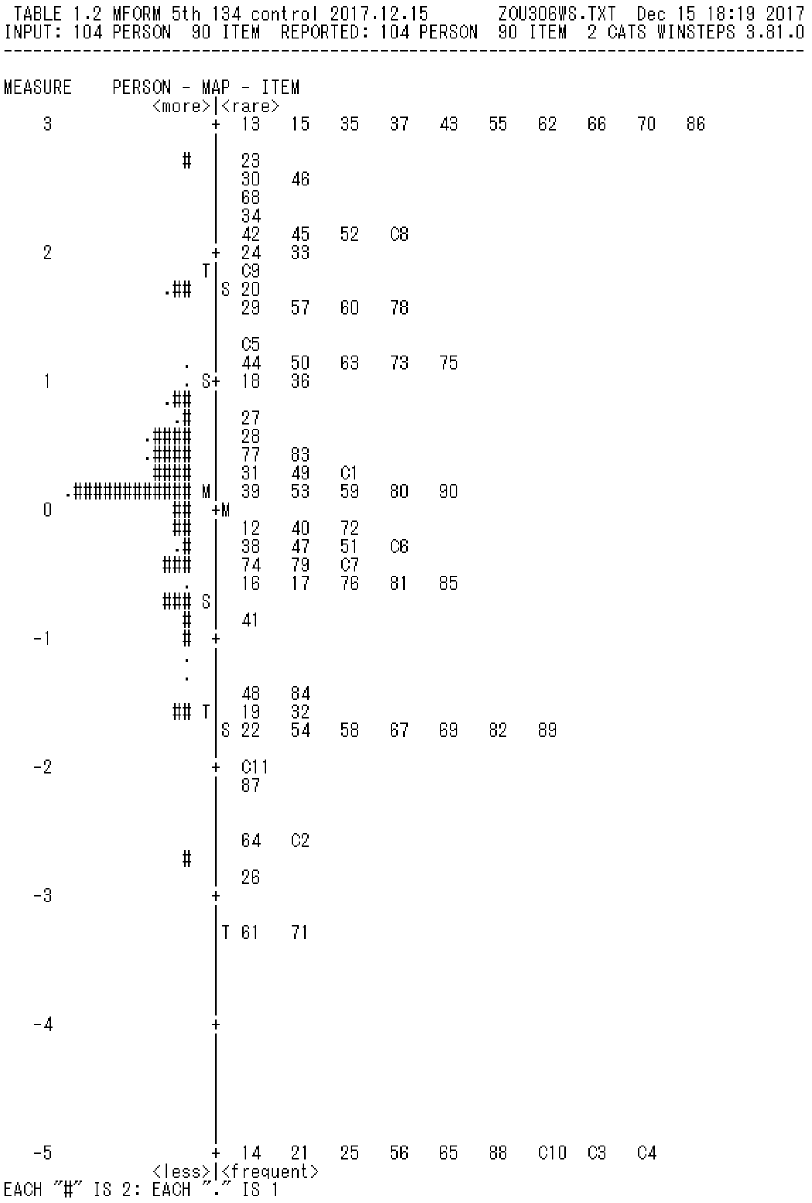
4th revision

As shown in Table 5.14, there were 90 items and 104 persons in the 4th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items 4th revision.

The separation measure concerning the 4th revision of the equated tests is relatively low (0.74–0.89), which indicates that the number of items used is rather small to distinguish persons. Its person reliability is low (0.35), which is due to the relatively small number of test items. The reliability of item is relatively high (0.87) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.59–2.60 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.30 shows that the persons mean (0.10) is located slightly below the items mean that is set to 0.00 by default, indicating that, on average, items are relatively appropriate levels of difficulty for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the equated tests.

Of the 104 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 30 (Outfit MNSQ: 5.88, ZSTD: 2.19), 41 (Outfit MNSQ: 8.23, ZSTD: 3.05), and 48 (Infit MNSQ: 1.41, ZSTD: 2.08)’s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 30, 41, and 48 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.



```
 TABLE 1.2 MFORM 5th 134 control 2017.12.15       ZOU306WS.TXT  Dec 15 18:19 2017
INPUT: 104 PERSON  90 ITEM  REPORTED: 104 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +  13   15   35   37   43   55   62   66   70   86
                     |
                  #  |  23
                     |  30   46
                     |  68
                     |  34
                     |  42   45   52   C8
    2                +  24   33
                    T|  C9
                .##  |S 20
                     |  29   57   60   78
                     |
                     |  C5
                  .  |  44   50   63   73   75
    1             . S+  18   36
                .##  |
                 .#  |  27
              .####  |  28
              .####  |  77   83
               ####  |  31   49   C1
      .############ M|  39   53   59   80   90
    0            ##  +M
                 ##  |  12   40   72
                 .#  |  38   47   51   C6
                ###  |  74   79   C7
                  .  |  16   17   76   81   85
                ### S|
                  #  |  41
   -1             #  +
                  .  |
                  .  |
                     |  48   84
                 ## T|  19   32
                     |S 22   54   58   67   69   82   89
                     |
   -2                +  C11
                     |  87
                     |
                     |
                     |  64   C2
                  #  |
                     |  26
   -3                +
                     |
                     |T 61   71
                     |
                     |
                     |
                     |
   -4                +
                     |
                     |
                     |
                     |
                     |
                     |
   -5                +  14   21   25   56   65   88   C10  C3   C4
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.30. Item and person map for the equated tests (4th revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

5th revision

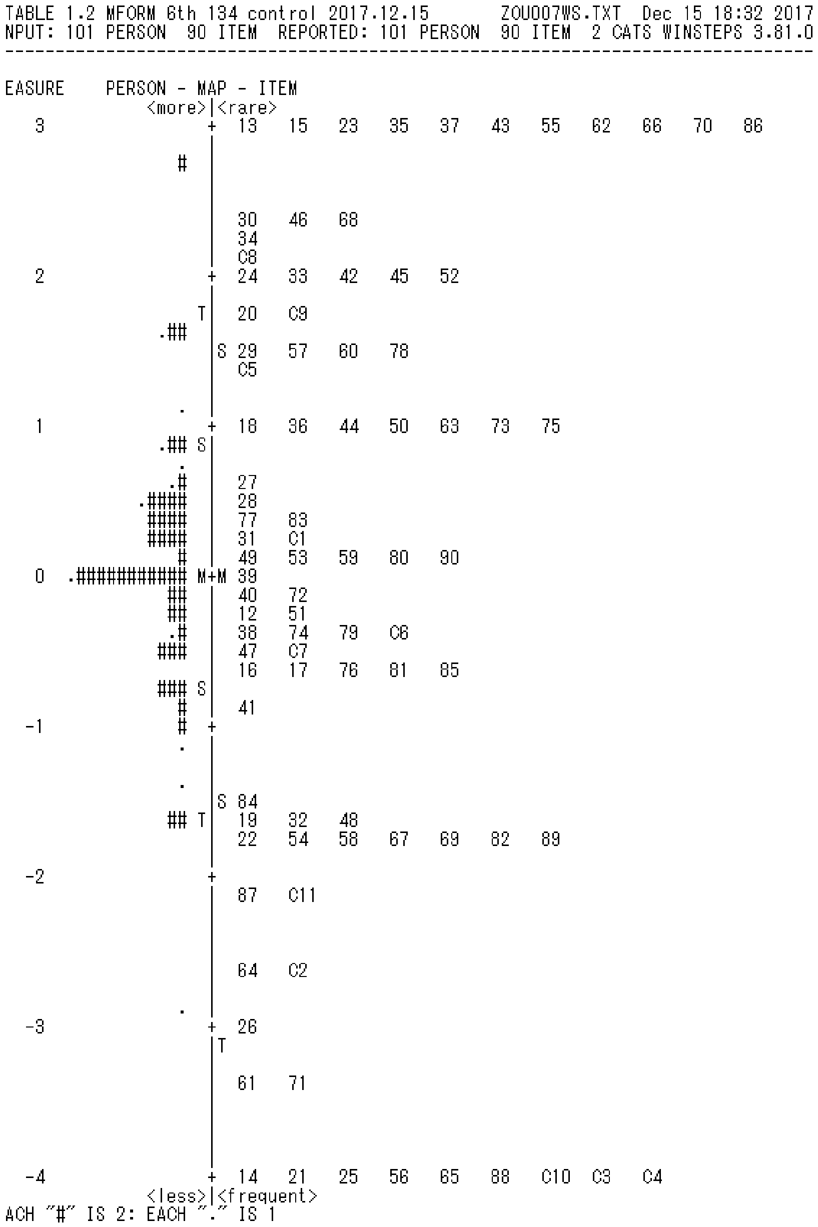
As shown in Table 5.14, there were 90 items and 101 persons in the 5th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 5th revision.

The separation measure concerning the 5th revision of the equated tests is relatively low (0.69–0.83), which indicates that the number of items used is rather

small to distinguish persons. Its person reliability is low (0.32), which is due to the relatively small number of test items. The reliability of item is relatively high (0.87) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.57–2.58 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.31 shows that the persons mean (0.06) is located slightly above the items mean that is set to 0.00 by default, indicating that, on average, items are relatively appropriate levels of difficulty for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the equated tests.

Of the 101 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 4 (Outfit MNSQ: 5.65, ZSTD: 2.14), 15 (Outfit MNSQ: 4.86, ZSTD: 2.56), 37 (Outfit MNSQ: 3.38, ZSTD: 2.15), 74 (Infit MNSQ: 1.40, ZSTD: 2.02), 78 (Outfit MNSQ: 2.11, ZSTD: 2.47), 82 (Outfit MNSQ: 0.50, ZSTD: -2.02), 119 (Outfit MNSQ: 1.79, ZSTD: 2.21), and 126 (Outfit MNSQ: 1.88, ZSTD: 2.18),'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 4, 15, 37, 74, 78, 82, 119, and 126 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.



```
TABLE 1.2 MFORM 6th 134 control 2017.12.15       ZOU007WS.TXT  Dec 15 18:32 2017
NPUT: 101 PERSON  90 ITEM  REPORTED: 101 PERSON  90 ITEM  2 CATS WINSTEPS 3.81.0
--------------------------------------------------------------------------------

EASURE    PERSON - MAP - ITEM
              <more>|<rare>
   3                +  13   15   23   35   37   43   55   62   66   70   86
                    |
                 #  |
                    |
                    |
                    |  30   46   68
                    |  34
                    |  C8
   2                +  24   33   42   45   52
                    |
                   T|  20   C9
               .##  |
                    |S 29   57   60   78
                    |  C5
                    |
                 .  |
   1                +  18   36   44   50   63   73   75
               .## S|
                 .  |
                .#  |  27
             .####  |  28
              ####  |  77   83
              ####  |  31   C1
                 #  |  49   53   59   80   90
   0  .########### M+M 39
                ##  |  40   72
                ##  |  12   51
                .#  |  38   74   79   C6
               ###  |  47   C7
                    |  16   17   76   81   85
               ### S|
                 #  |  41
  -1             #  +
                 .  |
                    |
                 .  |
                    |S 84
                ## T|  19   32   48
                    |  22   54   58   67   69   82   89
                    |
  -2                +
                    |  87   C11
                    |
                    |
                    |
                    |  64   C2
                    |
                 .  |
  -3                +  26
                    |T
                    |
                    |  61   71
                    |
                    |
                    |
                    |
  -4                +  14   21   25   56   65   88   C10  C3   C4
              <less>|<frequent>
ACH "#" IS 2: EACH "." IS 1
```

Figure 5.31. Item and person map for the equated tests (5th revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

6th revision

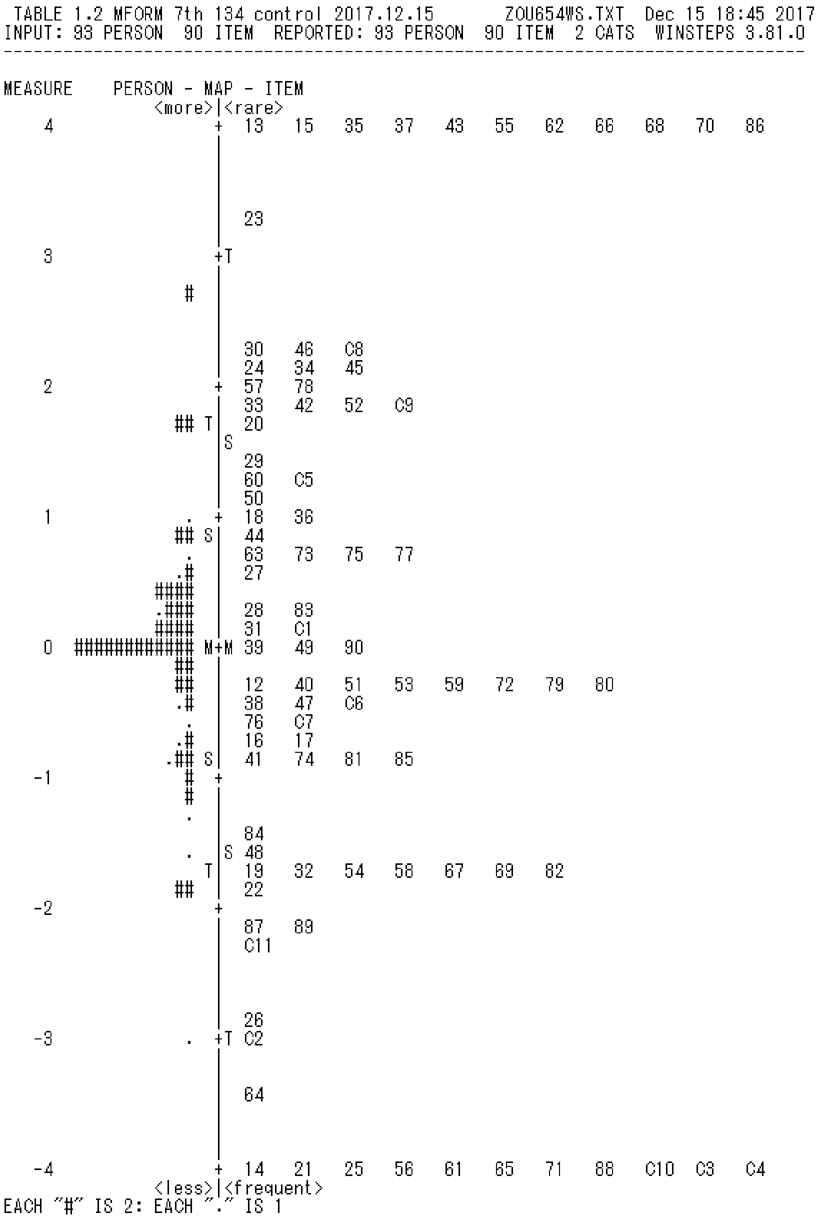
As shown in Table 5.14, there were 90 items and 93 persons in the 6th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 6th revision.

The separation measure concerning the 6th revision of the equated tests is relatively low (0.77–0.91), which indicates that the number of items used is rather

small to distinguish persons. Its person reliability is low (0.38), which is due to the relatively small number of test items. The reliability of item is moderately high (0.86) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.46–2.47 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.32 shows that the persons mean (−0.01) is located slightly below the items mean that is set to 0.00 by default, indicating that, on average, items are relatively appropriate levels of difficulty for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the equated tests.

Of the 93 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 27 (Outfit MNSQ: 5.27, ZSTD: 2.06), and 124 (Outfit MNSQ: 2.35, ZSTD: 2.45),'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that persons 27 and 124 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.



```
 TABLE 1.2 MFORM 7th 134 control 2017.12.15        ZOU654WS.TXT  Dec 15 18:45 2017
INPUT: 93 PERSON  90 ITEM  REPORTED: 93 PERSON  90 ITEM  2 CATS  WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +  13   15   35   37   43   55   62   66   68   70   86
                     |
                     |
                     |
                     |
                     |  23
                     |
    3                +T
                     |
                  #  |
                     |
                     |
                     |  30   46   C8
                     |  24   34   45
    2                +  57   78
                     |  33   42   52   C9
                 ## T|  20
                     |S
                     |  29
                     |  60   C5
                     |  50
    1             .  +  18   36
                 ## S|  44
                  .  |  63   73   75   77
                 .#  |  27
               ####  |
               .###  |  28   83
               ####  |  31   C1
    0  ############ M+M 39   49   90
                 ##  |
                 ##  |  12   40   51   53   59   72   79   80
                 .#  |  38   47   C6
                  .  |  76   C7
                 .#  |  16   17
                .## S|  41   74   81   85
   -1             #  +
                  #  |
                  .  |
                     |  84
                  .  |S 48
                    T|  19   32   54   58   67   69   82
                 ##  |  22
   -2                +
                     |  87   89
                     |  C11
                     |
                     |
                     |
                     |  26
   -3             .  +T C2
                     |
                     |
                     |  64
                     |
                     |
                     |
   -4                +  14   21   25   56   61   65   71   88   C10  C3   C4
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.32. Item and person map for the equated tests (6th revision*)*.

Note. This figure is WINSTEPS® version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

7th revision

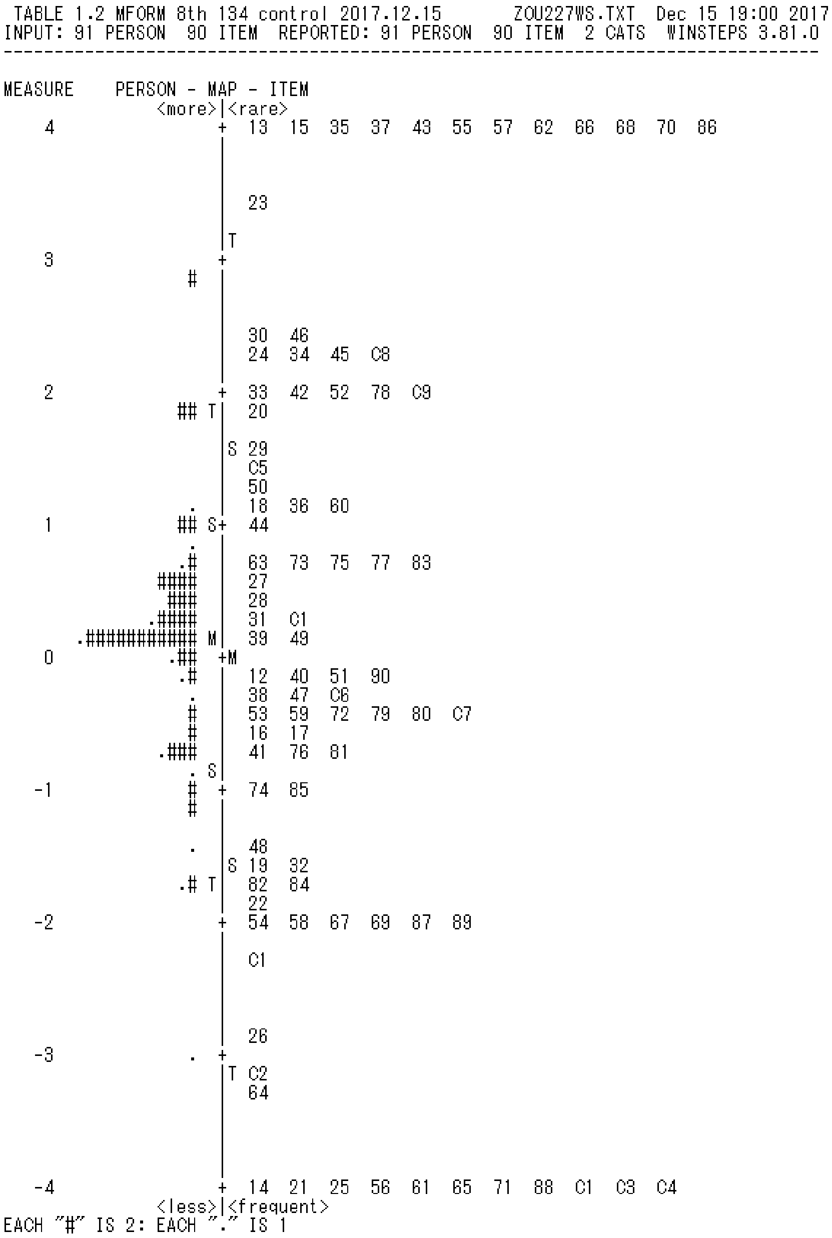
As shown in Table 5.14, there were 91 items and 93 persons in the 7th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 7th revision.

The separation measure concerning the 7th revision of the equated tests is relatively low (0.79–0.92), which indicates that the number of items used is rather

small to distinguish persons. Its person reliability is low (0.38), which is due to the relatively small number of test items. The reliability of item is moderately high (0.85) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.43 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.33 shows that the persons mean (0.10) is located slightly below the items mean that is set to 0.00 by default, indicating that, on average, items are relatively appropriate levels of difficulty for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the equated tests.

Of the 91 persons measured, all MNSQ ranges fell between 0.6-1.4, except for persons 112 (Outfit MNSQ: 1.75, ZSTD: 2.12)'s MNSQ value with a ZSTD value, possibly indicating mismatched persons. As Jarl, Heinemann, and Hermansson (2012) point out, an MNSQ value exceeding 1.4 with a ZSTD value more than 2.0 or an MNSQ value less than 0.6 with a ZSTD value less than −2.0, indicates a possibly mismatched person, suggesting that person 112 should be deleted from the list. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.



```
 TABLE 1.2 MFORM 8th 134 control 2017.12.15       ZOU227WS.TXT  Dec 15 19:00 2017
INPUT: 91 PERSON  90 ITEM  REPORTED: 91 PERSON  90 ITEM  2 CATS  WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +  13  15  35  37  43  55  57  62  66  68  70  86
                     |
                     |
                     |
                     |  23
                     |
                     |T
    3                +
                  #  |
                     |
                     |
                     |  30  46
                     |  24  34  45  C8
                     |
    2                +  33  42  52  78  C9
                 ## T|  20
                     |
                     |S 29
                     |  C5
                     |  50
                  .  |  18  36  60
    1            ## S+  44
                  .  |
                 .#  |  63  73  75  77  83
               ####  |  27
                ###  |  28
              .####  |  31  C1
       .########### M|  39  49
    0           .##  +M
                 .#  |  12  40  51  90
                  .  |  38  47  C6
                  #  |  53  59  72  79  80  C7
                  #  |  16  17
               .###  |  41  76  81
                  . S|
   -1             #  +  74  85
                  #  |
                     |
                  .  |  48
                     |S 19  32
                 .# T|  82  84
                     |  22
   -2                +  54  58  67  69  87  89
                     |
                     |  C1
                     |
                     |
                     |
                     |  26
   -3             .  +
                     |T C2
                     |  64
                     |
                     |
                     |
                     |
   -4                +  14  21  25  56  61  65  71  88  C1  C3  C4
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.33. Item and person map for the equated tests (7th revision*).*

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

8th revision

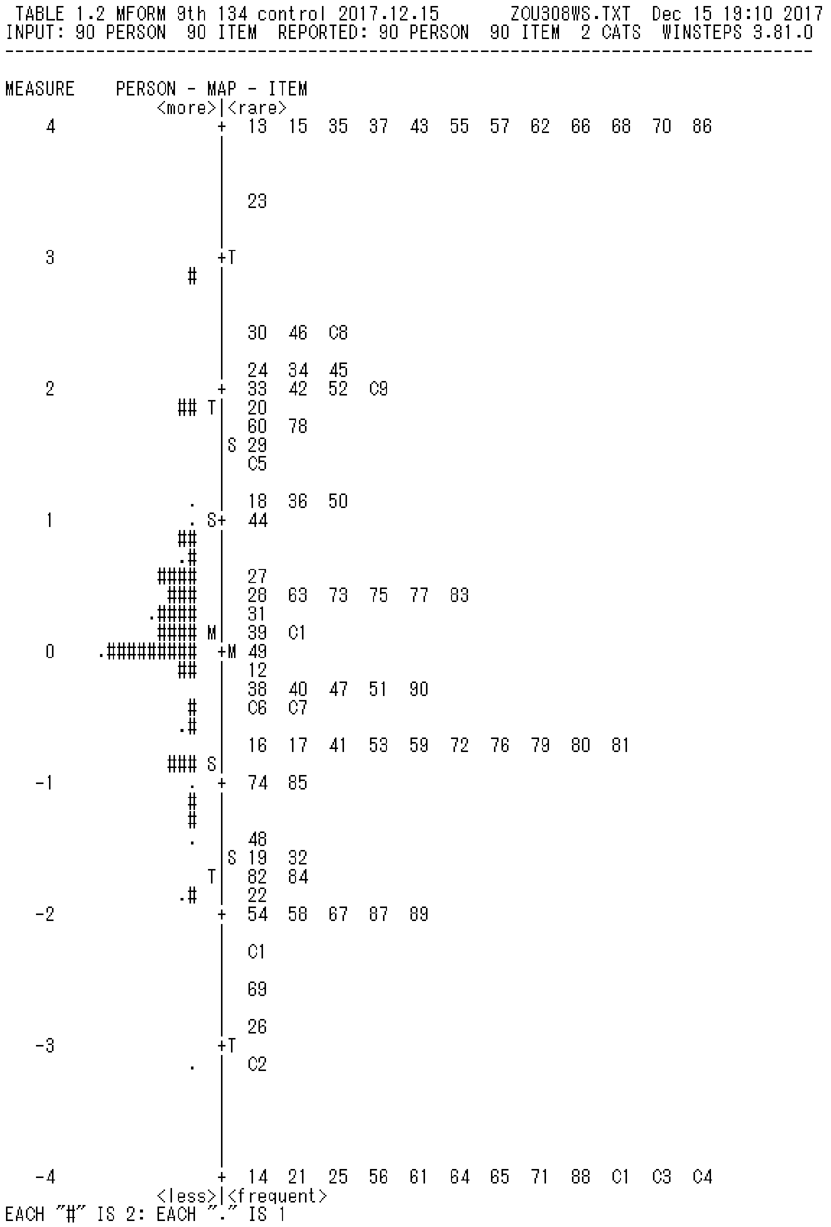
As shown in Table 5.14, there were 90 items and 90 persons in the 8th revision. WINSTEPS® version 3.81.0 provides the statistical values of the test items of 8th revision.

The separation measure concerning the 8th revision of the equated tests is relatively low (0.81–0.94), which indicates that the number of items used is rather

small to distinguish persons. Its person reliability is low (0.39), which is due to the relatively small number of test items. The reliability of item is moderately high (0.85) which indicates that, if the items were given to other comparable groups of test takers, there is a probability that the test would reproduce a similar order of item hierarchy. The item separation measure of 2.39–2.40 indicates that the items can be separated into more than two strata of difficulty (Karim, Shah, Din, Ahmad, & Lubis, 2014).

Figure 5.34 shows that the persons mean (0.08) is located slightly below the items mean that is set to 0.00 by default, indicating that, on average, items are relatively appropriate levels of difficulty for persons. Moreover, there are items which item difficulty estimates fall below the person ability estimates for the equated tests.

Of the 90 persons measured, all MNSQ ranges fell between 0.6-1.4. Of the 90 items measured, all MNSQ ranges fell between 0.6-1.4.



```
 TABLE 1.2 MFORM 9th 134 control 2017.12.15       ZOU308WS.TXT  Dec 15 19:10 2017
INPUT: 90 PERSON  90 ITEM  REPORTED: 90 PERSON  90 ITEM  2 CATS  WINSTEPS 3.81.0
--------------------------------------------------------------------------------

MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    4                +  13  15  35  37  43  55  57  62  66  68  70  86
                     |
                     |
                     |
                     |  23
                     |
                     |
    3                +T
                  #  |
                     |
                     |
                     |  30  46  C8
                     |
                     |  24  34  45
    2                +  33  42  52  C9
                 ## T|  20
                     |  60  78
                     |S 29
                     |  C5
                     |
                  .  |  18  36  50
    1             . S+  44
                 ##  |
                 .#  |
               ####  |  27
                ###  |  28  63  73  75  77  83
              .####  |  31
               #### M|  39  C1
    0    .#########  +M 49
                 ##  |  12
                     |  38  40  47  51  90
                  #  |  C6  C7
                 .#  |
                     |  16  17  41  53  59  72  76  79  80  81
                ### S|
   -1             .  +  74  85
                  #  |
                  #  |
                  .  |  48
                     |S 19  32
                    T|  82  84
                 .#  |  22
   -2                +  54  58  67  87  89
                     |
                     |  C1
                     |
                     |  69
                     |
                     |  26
   -3                +T
                  .  |  C2
                     |
                     |
                     |
                     |
                     |
   -4                +  14  21  25  56  61  64  65  71  88  C1  C3  C4
               <less>|<frequent>
EACH "#" IS 2: EACH "." IS 1
```

Figure 5.34. Item and person map for the equated tests (8th revision*).*

Note. This figure is WINSTEPS^{R} version 3.81.0 output Table 1.2. Letters on left side represent examinees, while number on the right side indicate items. *M*, *S*, and *T* indicate the mean, 1- and 2-standard deviations respectively.

5.8.4 Unidimensionality

In this section, the unidimensionality of the final revision (8th revision) will be described in terms of principal component analysis of residuals. Table 5.15 indicates that observed raw variance explained by measures (34.2%) almost fits the expected raw variance explained by measure (34.4%), indicating that explainable variance fits the Rasch model. Rasch, however, explained only 34.2% of the 90 items, leaving more than half of variance (66.0%) unaccounted for by the model. This is due to the fact that the ability range of the test takers is relatively narrow. A wider proficiency level of the test takers would result in a greater explained variance. The strongest secondary dimension is named the first contrast, while the following dimensions are named the second, third, fourth and fifth respectively. The largest secondary dimension (first contrast) in the equated tests data had strength of 6.9 units (10.5%) while the variance explained by measures was larger at 34.3 units (34.2%), indicating that the secondary contrast does not create multidimensionality.

Table 5.15

Standardized Residual Variance (in Eigen-value Units) for the equated tests

	Observed					Expected	
Total raw variance in observations	100.3	100.0	%			100	%
Raw variance explained by measures	34.3	34.2	%			34.4	%
Raw variance explained by persons	6.6	6.5	%			6.6	%
Raw variance explained by items	27.7	27.7	%			27.8	%
Raw unexplained variance (total)	66.0	65.8	%	100	%	65.6	%
Unexplained variance in 1st contrast	6.9	6.9	%	10.5	%		
Unexplained variance in 2nd contrast	4.4	4.4	%	6.7	%		
Unexplained variance in 3rd contrast	4.1	4.1	%	6.2	%		
Unexplained variance in 4th contrast	3.4	3.4	%	5.1	%		
Unexplained variance in 5th contrast	3.0	3.0	%	4.6	%		

Note. This figure is from WINSTEPS^{R} version 3.81.0 output Table 23.0.

5.9.1 Pre-test prediction

Before administering the test, the researcher developed test items according to two features: test type and loanwords. Test type represented six groups:

- Group JJ, in which the grapheme–phoneme relationship of both words in the test item is the same as that in Japanese;
- Group JN, in which the grapheme of one of the two words has a Japanese grapheme–phoneme relationship, whereas the other’s grapheme has a phoneme Japanese speakers can pronounce, but a grapheme–phoneme relationship unlike the ones in Japanese;
- Group JJN, in which the grapheme of one of the two words has a Japanese grapheme–phoneme relationship, whereas the other’s grapheme is pronounced in an English-language phoneme that does not exist in Japanese pronunciation;
- Group NN, in which the grapheme of the two words has a phoneme that Japanese speakers can pronounce, but a grapheme–phoneme relationship unlike the ones in

Japanese (NN);

- Group NNN, in which the grapheme of the two words has a phoneme that can be pronounced by Japanese speakers but a grapheme–phoneme relationship that differs from that in Japanese, whereas the other word’s grapheme is pronounced in an English-language phoneme that does not exist in Japanese pronunciation; and
- Group NNNN, in which the grapheme of the two words is pronounced in an English-language phoneme that does not exist in Japanese pronunciation.

The intuitive prediction was that Japanese learners of English would find it most difficult to identify a grapheme pronounced in an English-language phoneme that does not exist in Japanese pronunciation and it easiest to identify a grapheme with a Japanese grapheme–phoneme relationship. Accordingly, the items in Group JJ should pose the least difficulty, whereas items in Group NNNN should pose the most. Table 5.16 shows the number of vowel items based on test type in the equated tests.

Table 5.16

Number of Items Based on Test Type

	JJ	JN	JNN	NN	NNN	NNNN	Total
Equated tests	6	9	3	29	21	22	90

Meanwhile, loanwords can be categorised into three groups:

- Group NN, in which the two words in a test item do not have loanwords in Japanese;
- Group NY, in which one of the two words in a test item has a loanword in Japanese, whereas the other does not; and
- Group YY, in which the two words in a test item have loanwords in Japanese.

The intuitive prediction was that Japanese learners of English would find identifying words with Japanese loanwords more difficult than words without them, because, as described in the previous chapter, Japanese loanwords tend to oversimplify English-language grapheme–phoneme relationships. Therefore, items with two loanwords should have the highest item difficulty, whereas items without loanwords should have the least.

Table 5.17 shows the number of test items based on loanwords in the equated tests.

Table 5.17

Number of Items Based on Japanese Loanwords

	NN	NY	YY	Total
Equated tests	73	7	10	90

5.9.2 Empirical results

Figure 5.35 shows the difficulty estimates of items based on test type in the equated tests. Although the pre-test prediction estimated that test types involving a grapheme with an English grapheme–phoneme relationship would be most difficult, the results of Rasch analysis contradict that prediction, since two groups representing a grapheme with an English grapheme–phoneme relationship, Group NN (−0.07) and NNNN (0.09), were ranked in the least according to item difficulty. Results thus suggest that although test type may not appear to be a strong factor in item difficulty on equated tests, the fact that Group JJ (1.3), which was ranked highest, and Group JN (1.11) had only six and nine items, respectively, needs to be considered.

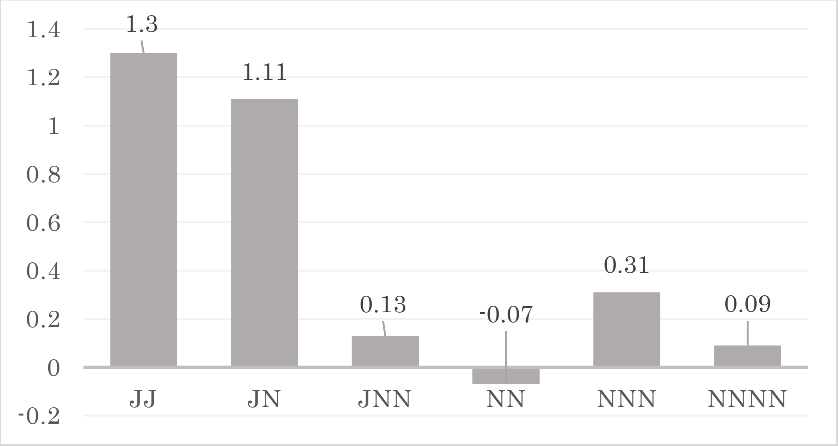


Figure 5.35. Item difficulty by test type on equated tests.

Figure 5.36 shows the difficulty estimates of items based on loanwords on the equated tests. The pre-test prediction estimated that items with loanwords would have the highest item difficulty, whereas items without loanwords would have the least. The results of Rasch analysis contradict that prediction, however, Group NY (0.30) was ranked highest in terms of item difficulty, whereas Group YY (0.11) was ranked in the middle. The fact that Group NY/YN, which ranked highest, and Group YY had only 10 and 7 items respectively should be considered.

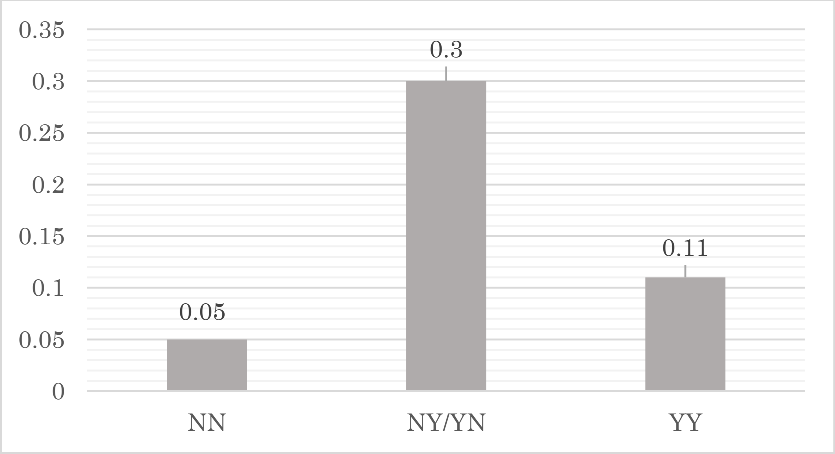


Figure 5.36. Item difficulty by loanwords on the equated tests.

5.10 Item banking

As a result of WINSTEPS' analysis, 90 items that assess orthographic and phonological processing skills were divided into three levels, beginner, intermediate, and advanced level according to item difficulty and test type. Table 5.18 shows the number, the average and S.D. of item difficulty measure. Items with item difficulty value less than −1.0 were categorized as beginner level, item with item difficulty value between −1.0 and 1.0 as medium level, and items with more than 1.0 were categorized as advanced level. More than 82% (74) of items were categorized as medium level while, approximately 10% for both beginner and advanced level. One of the purposes of developing item bank is to place most achieving students into an advanced class while others are all placed into medium class, the development of more advanced-level items is one of the important future task.

Table 5.18

Number and Difficulty Measures Based on Item Levels

Level	Number	M	S.D.
Beginner (<−1.0)	9	-1.66	0.60
Medium	74	0.10	0.40
Advanced (>1.0)	7	2.02	0.60

With regards to the test type, the items are categorized into six groups; the categorization referred to in the earlier section. Figure 5.37 shows the test type ratio of beginner level group. Only 11 percent of the items at least includes one grapheme that has Japanese grapheme–phoneme relationships in beginner level items.

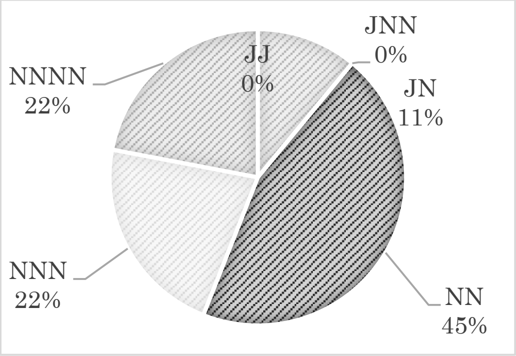


Figure 5.37. Beginner level items' ratio of test type

Figure 5.38 shows the test type ratio of medium level group. Unlike beginner's level, more than a fifth (23%) of the items at least includes one grapheme that has a Japanese grapheme–phoneme relationship.

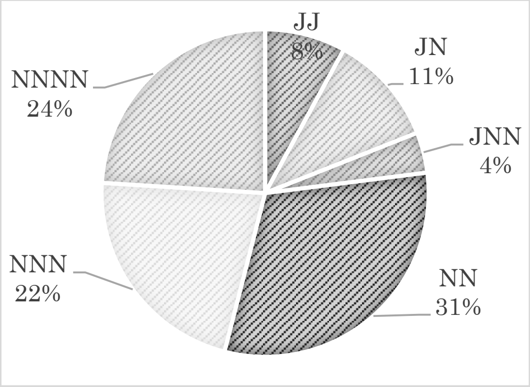


Figure 5.38. Medium level items' ratio of test type

Figure 5.39 shows the test type ratio of advanced-level group. Unlike medium level items, all items do not include a grapheme that has a Japanese grapheme–phoneme relationship. Although the total number of advanced-level items is limited (7) compared to medium level (74), the item ration of advanced level may suggest that items without a grapheme that has a Japanese grapheme–phoneme relationship have higher item difficulty values.

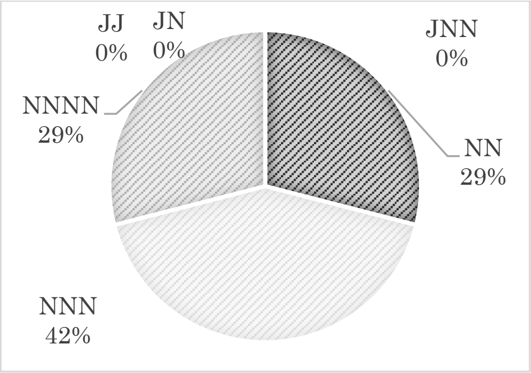


Figure 5.39. Advanced-level items' ratio of test type

5.11 Summary findings and conclusion

In this chapter, the researcher analysed the results of five tests—namely the Spring 2014 final test (14SF), Fall 2014 final test (14FF), 2015 placement test (15PP), Spring 2015 final test (15SF), and Fall 2015 final test (15FF)—that include orthographic and phonological test items administered to students of the nursing faculty during 2014–2015. WINSTEPS version 3.81.0 was used to analyse the tests' separation, reliability, targeting, item fit, and unidimensionality. Based on Linacre (2013) and Jarl, Heinemann, and Hermansson (2012)'s definition, items or persons with MNSQ values greater than 1.4 with ZSTD in excess of 2.0 or MNSQ values less than 0.6 with ZSTD values less than −2.0 were deleted until all items and person values improved.

After analysis, three tests—namely 14SF, 15SF, and 15FF—that have 16 common items were calibrated in order to organise test items into the same frame of reference. Firstly, Differential Test Functioning (DTF) anaylsis was performed to investigate whether the test items for two groups of test takers functioned similarly. Secondly, outlier items were removed until all items functioned similarly for all groups. Through the process five items were deleted. Using remaining eleven common items, 90 items—41 items from 15SF and 38 items from 15FF—, concurrent equating was performed. After calibration, equated test items were analysed for test separation, reliability, targeting, item fit, and unidimensionality. The items were, then, analysed for how item difficulty estimates, test type, and loanword related to one another. In the intuitive prediction, items with English-language grapheme–phoneme relationships were predicted to have the greatest difficulty estimates. That prediction, however, was contradicted by the fact that no items with English-language grapheme–phoneme relationships showed higher difficulty estimates. Moreover, the items with loanword did not differ greatly from items without loanword.

With regards to the development of item bank, 90 test items are categorized into three level groups—beginner, medium, and advanced. The analysis of test type reveals the advanced-level group has less ratio of a grapheme with a Japanese grapheme–phoneme relationship. The result moderately fit the intuitive prediction that Japanese learners of English would find it most difficult to identify a grapheme pronounced in an English-language phoneme that does not exist in Japanese pronunciation and it easiest to identify a grapheme with a Japanese grapheme–phoneme relationship. In the future item development, especially when devising advanced-level test items, more items that has a grapheme with an English grapheme–phoneme relationship should be included.

Chapter 6: Discussion

The purpose of this chapter is to discuss the three primary research questions outlined earlier by focusing on the data reported in the Results section.

6.1 Research question 1

Research Question (RQ) 1 asked what kinds of abilities are required in the faculty's curriculum and English materials in terms of orthographic knowledge and phonological awareness. The English subjects and textbooks from the Faculty of Nursing' syllabus, along with its list of medical vocabulary in a reference book were adopted as material for the research.

The syllabus analysis shows that the faculty requires students to acquire both communication and academic skills in English. To this end, students are encouraged to acquire reading fluency in both medical and everyday English. The selection of English course textbooks is left up to each lecturer to make, as long as the textbooks reflect the aims of the faculty's curriculum.

The textbook analysis shows that although all textbooks contained a separate section for learning vocabulary in each lesson, none of the sections include supports for learning orthographic and phonological awareness. Just three textbooks of 12 included the phonetic symbol for each word in the vocabulary section, while the majority of the sections only provide an accurate meaning of the words.

As shown in Table 3.1, the textbooks selected for the classrooms are inconsistent in their purposes and goals. Nearly half of the textbooks focus on acquiring communication skills, while the others contain medical and health issues. Therefore, the vocabulary list in the medical vocabulary book, "*Igakueitango* [*Medical English vocabulary*]" is included as material in this research in order to investigate the orthographical and phonological features of medical terms. The vocabulary analysis shows that the terms are all ranked as off-list words in Laufer and Nation's (1995) lexical frequency profile. Moreover, of 493 words, more than three-quarters of the words are categorized as words that contain more than four syllables. With regard to the grapheme–phoneme relationship of medical terms, the percentage of graphemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes and whose phonemic symbols do not exist in Japan exceeds the other two categories in all five graphemes, /a/, /e/, /i/, /o/, and /u/. While 71% of grapheme /i/ is pronounced as [i], which is the same as Japanese grapheme–phoneme correspondence, less than 5% of graphemes /a/, /o/, and /u/ are pronounced in the ways the graphemes /a/, /o/, and /u/ are pronounced in Japanese. Moreover, more than 80% of graphemes /a/ and /u/ are pronounced differently in Japanese.

The recapitulation of the research and main findings indicate that although the faculty encourages the students to acquire reading fluency in medical English, the textbooks do not provide sufficient support to raise awareness of or teach orthographical and phonological skills. Textbooks are highly influential in English classrooms because teaching is conducted based on the content of the textbook. The absence of regular word exercises that lead to awareness of the complexities of English grapheme–phoneme relationships in the textbooks indicates classroom practices will also follow this trend. Students may acknowledge that they should pay less attention to the orthographical and phonological aspects of English words than the meanings of those words.

The level of medical terms, however, appears more demanding for the students who have been learning English with less attention on orthographical and phonological awareness. This is not only due to the fact that medical terms are categorized as difficult in Laufer and Nation’s (1995) word list, but the irregularity of English grapheme–phoneme relationships is prevalent in most of the English vowels except /i/. As shown in Chapter 1, the *Course of Study* is relatively reluctant to teach orthographical and phonological awareness in junior high schools and high schools in Japan. Moreover, none of the textbooks used in junior high schools contained orthographic or phonological exercises or activities in the textbooks. The English-Japanese dictionaries available for students show inconsistencies in presenting the phonemes. With the absence of appropriate learning opportunities of English grapheme–phoneme inconsistencies, both in class and in textbooks, along with L1 interference, which is one-to-one correspondence between grapheme and phoneme in Japanese, the loanwords, and open-ended syllables, the teaching and learning of medical terms seems formidable for both teachers and students.

6.2 Research question 2

Research Question (RQ) 2 asked what kinds of test items are used and which constructs are measured in practice tests of commercially produced English-language proficiency tests and how well those abilities reflect the content of the faculty's curriculum. The official guides and the official collection of practice test items of the two most prominent and frequently adopted commercially produced English-language tests for placement purposes in Japanese universities are the Test of English as a Foreign Language (TOEFL), and the Test of English for International Communication (TOEIC), which were adopted as material for the research.

The analysis of official guides shows that neither the TOEFL nor TOEIC explicitly states that their test items assess orthographic and phonological awareness. As shown in Table 4.2, the TOEFL guide offers the following advice for low-level test takers: "increase your vocabulary by analysing word parts; study roots, prefixes and suffixes; [and] study word families" (ETS, 2008). The statement appears to estimate orthographic and phonological awareness as prerequisites for gaining reading proficiency in English. Moreover, the facts that the advice is administered towards low-level achievers and that the TOEFL aims to assess the test takers' academic reading abilities indicate that orthographic and phonological processing skills are acknowledged as basic skills for reading academic materials.

The analysis of orthographic and phonological features of the TOEFL and TOEIC's reading passages' vocabulary shows that more than 70% of words contain no more than three syllables. As Figure 4.9 shows, the medical reference book has 4.19 syllables per word, while both the TOEFL and TOEIC contain approximately two syllables per word. As Figure 4.10 shows, the percentages of medical term's graphemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes, and whose phonemic symbols do not exist in Japanese (NNE) are largest of the three: one, phonetic symbols that have the same graphemes as the Japanese one-to-one correspondences between graphemes and phonemes (JE); two, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes but whose phonetic symbols exist in Japanese pronunciation (NE); and, three, phonemes that do not follow Japanese one-to-one correspondences between graphemes and phonemes, and whose phonemic symbols do not exist in

Japanese (NNE).

The recapitulation of the research and main findings indicate that although the faculty encourages the students to acquire reading fluency in medical English, the two commercially produced English-language proficiency tests may not be appropriate tools to measure the students' orthographic and phonological processing skills. As shown in Table 1.6, the number of universities that use English-language placement tests is increasing in Japan. Moreover, commercially produced, norm-referenced, English-language proficiency tests are favoured because of their efficacy and effectiveness in administering the test. The lack of test items that directly assess orthographic and phonological processing skills in the two tests, however, may prevent the test administrators from making a valid judgement when placing students into an appropriate level of English classes. Yet, the faculty encourages students to acquire high-level medical terms as a result of participating in English classes. In order to achieve this goal, the accurate assessment of students' orthographic and phonological processing skills at the beginning of their academic life is important. To this end, the development of an item bank that stores orthographic and phonological processing skills is necessary. Moreover, the test items should reflect the orthographic and phonological features of medical terms; relatively high-level vocabulary in Laufer and Nation's (1995) lexical frequency profile; relatively large amount of syllables per words; and relatively complex grapheme–phoneme correspondence.

6.3 Research question 3

Research Question (RQ) 3 asked whether the development of an item bank for the English-language placement test is a valid process and how well the test items reflect the content of the faculty's curriculum.

All items aimed to assess students' ability to identify relationships between English graphemes and phonemes. To validate the item development process, Rasch analysis, including separation, reliability, test targeting, and unidimensionality, was performed with all five tests, for a total of 147 items, which yielded 90 equated test items. The validation process included the deletion of items and participants that did not fit the Rasch model, which resulted in one to four revisions for each test. Moreover, the equated test items were also evaluated for whether they fit the Rasch model. As a result, the test items showed sufficient spreads: 9 (10%) were grouped at the beginner level, 74 (82%) at the intermediate level, and 7 (8%) at the advanced level. The pre-test prediction that students would find the test items with an English grapheme–phoneme relationship that does not exist in Japanese pronunciation most difficult, however, was not confirmed mainly due to the fact that the number of test items that assess Japanese grapheme-phoneme relationships was relatively limited.

6.4 Overall discussion

The combined findings from each part of the study enabled the current research to identify the contexts and importance of developing test items that assess orthographic and phonological processing skills early on during university study. The analysis of the Faculty of Nursing's syllabus and the medical vocabulary list revealed challenges that nursing students face when required to read texts with complex grapheme–phoneme relationships, but that are nevertheless seldom explored. The fact that students were not prepared to read words with English grapheme–phoneme relationships in high school, as well as the variation of phonetic symbols in English-Japanese dictionaries frequently used in English-language learning in Japan, contributes to those students' challenges. Although providing appropriate assistance at the beginning of university life is necessary, many English proficiency tests frequently used as placement tests at Japanese universities do not meet those demands. The lack of test items that explicitly assess orthographic and phonological processing skills in turn negatively affects students' motivation to learn those skills. Students might view the skills as unnecessary to learn, and, therefore spend more cognitive energy on identifying graphemes with English grapheme–phoneme relationships. Developing test items that assess orthographic and phonological processing skills and presenting them with a common frame of reference are necessary to assess students' English proficiency upon their matriculation at university.

Chapter 7 Conclusion

7.1 Contribution of the thesis

While the importance of acquiring orthographic and phonological processing skills when learning to read fluently is widely acknowledged, especially in L1 reading, less has been addressed in the L2/FL reading context. In Japanese tertiary-level institutions, this trend is highly applicable with the increase in the numbers of universities over the past twenty years, as well as with the relaxation of admission policies, especially in the area of the English language, the neglect of teaching English grapheme–phoneme relationships has been fostered in primary and secondary schools. Moreover, Japanese language systems—most of which have one-to-one relationships between graphemes and phonemes—have posed a challenge to university entrants who are required to read medical texts with complex grapheme–phoneme relationships. Little has been said, however, about what is necessary for the university to meet the English needs of the university entrants in the Faculty of Nursing, particularly for the assessment of their orthographic and phonological processing skills. Moreover, there has always been concern over using commercially produced English proficiency tests to assess students' orthographic and phonological processing skills.

The analysis of the reading texts used in two commercially produced English proficiency tests and the features of medical terms have provided a valuable insight into the fact that there is an evident discrepancy between the two. The medical terms require more complex processing skills regarding grapheme–phoneme relationships than that of the TOEFL and the TOEIC. The study has also contributed to the development of placement test items that assess students' orthographic and phonological processing skills. The items that have been developed in the study can be used for students with diverse orthographic and phonological processing skill levels, since item difficulty values were estimated in advance using Rasch analysis. The use of our own items with item difficulty estimates enables the Faculty of Nursing to provide students to take English placement test at their appropriate levels. Moreover, the study offers insights into the process of developing a unique item bank in a relatively small and less resourceful institution that needs to assess orthographic and phonological processing skills. Finally, the study provides an insight into the type of test item in which the English grapheme does not exist in Japanese pronunciation and the grapheme–phoneme relationship differs from that of Japanese; these test items might give test takers more difficulty.

7.2 Implications

The study highlights several issues regarding the assessment of orthographic and phonological processing skills. First, there is a need to provide a placement test that meets the demands of the faculty's English learning context. For example, like the Faculty of Nursing in this study, the two frequently used commercially produced English proficiency tests, the TOEFL and the TOEIC, were not suitable for fulfilling the faculty's need to know students' ability to read English medical texts with complex grapheme–phoneme relationships. To meet those demands, the development of a unique in-house test and item bank using Rasch analysis is encouraged.

Second, more extensive study on the effect of test types on item difficulty estimates is necessary for the enlargement and improvement of an item bank that meets the diverse proficiency levels of students. Although the study has partially revealed that certain test types (such as questions on English grapheme–phoneme relationships) might affect item difficulty, there are still more elements that contribute to the difficulty of test items. For example, there is a need to explore whether graphemes such as /a/ or /u/, which have various phonemes, are more difficult than a graphemes such as /i/, which have limited phoneme variation.

Finally, the study has revealed the importance of reviewing test items not only to analyse the quality of the item but also to use the information to create new test items in the future. As has been shown in the study, test items can be used in the same frame-of-reference by Rasch equating. Since the creation of new test items is demanding for test providers (which appears to have contributed to the hesitance of developing in-house placement test and thus dependence on commercially produced English proficiency tests), there is a need to develop a feasible and sustainable in-house placement development process. Test development process will benefit greatly from the using and reviewing of information from past tests.

7.3 Limitations

While the study has provided several insights into the less-addressed study area of English reading assessment, it has various limitations. First, the materials and instruments used in this study might have limited the outcomes of the study. For example, the study used only one medical vocabulary list to be compared to the vocabulary of commercially produced tests, which might have contributed to the distortion of the results. Moreover, the test items ask students to choose one answer from two, which might have contributed to the creation of chance score. A wider triangulation of measures would have been preferable had time and resources permitted.

Second, in the study, items were created based on grapheme–phoneme relationships and loanwords. There might have been several other factors contributing to the estimation of item difficulty, such as types of letters. Since tests were administered within the regular test period, there is a limitation on the number of test items that assess orthographic and phonological processing skills. With more time, the item analysis might have provided more insights.

7.4 Suggestions for future study

While the test items were validated only by Rasch analysis, there are several other ways to validate the test. For example, external tests can be used to assess its criterion, and questionnaires and observations of teachers and students would have contributed to increasing the face validity of the test.

Second, there is a need in the future study to explore the reason why the pretest estimate regarding the difficulty of test items was not confirmed. In order to do so, there is a need to increase the number of test items that assess the grapheme that has a Japanese phoneme-grapheme relationship since only 13 (14.4%) items included a word with a Japanese grapheme-phoneme relationship in the study.

As Linacre (2000) points out, that the development of computer-adaptive testing (CAT) should also be considered in the future study because of its potential benefits. For example, CAT can accommodate diverse proficiency levels of students because a wide range of tests items are stored in an item bank. Moreover, because CAT can flexibly provide well-targeted questions based on the individual student's response to each test item, test times can be shorter than with fixed-item tests, leading to less test fatigues and fewer careless mistakes. While researchers such as Sato (2015) provides a useful insight into the use of CAT for the preparation of TOEIC, less has been explored regarding CAT for measuring orthographic and phonological processing skills in Japan.

Finally, the future study should consider the practical use of the results of this study that contribute to the development of students' orthographic and phonological processing skills in actual classrooms. For example, the study sorted 90 test items into three levels: 7 advanced level items (8%), 74 intermediate level items (82%), and 9 beginner level items (10%). Words used in the test items were all to be found in the 2,000 most frequently used words in VocabProfiles section of the Compleat Lexical Tutor website, indicating that the majority of the words have been taught in Japanese junior and high school English classes. Therefore, the test items are appropriate elements to be used in the development of orthographic and phonological processing skills activities and worksheets in English-language classrooms at tertiary-level institutions. The direct teaching and learning of orthographic and phonological processing skills are judged to contribute to the acquisition of fluent reading skills in EFL.

Bibliography

Abe, M. Wistner, B., & Sakai, H. (2008). Analyzing an English language proficiency test using classical item analysis and Rasch modeling. *The Economic Journal of Takasaki City University of Economics, 50*(3-4), 125-134.

Alderson, J.C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Aline, D., & Churchill, E. (2006). Analyzing entrance exam item types with Rasch. *Kanagawa Daigaku Gengo Kenkyuu, 28*, 125-142. Retrieved on October 2014 from: http://hdl.handle.net/10487/3846

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats , R. Taft, , R. A. Heath, & S. H. Lovibond (Eds.), *Proceedings of the XXIVth International Congress of Psychology, 4*, 7-16. North Holland: Elsevier Science.

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Badian, N. A. (2001). Phonological and orthographic processing: Their roles in reading prediction *Annals of Dyslexia, 51*(1), 177-202.

Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham, & D. Corson (Eds.), *Encyclopaedia of language and education Vol. 7: Language testing and assessment* (pp. 275-287). Dordrecht, the Netherlands: Kluwer Academic.

Barker, T. A., Torgesen, J. K., & Wagner, R. K. (1992). The role of orthographic processing skills on five different reading tasks. *Reading Research Quarterly, 27*, 335–345.

Beeston, S. (2000). The UCLES EFL item banking system. *Research Notes 2*, 8-9.

Benthusysen, R. V. (2005). Japanese EFL students' awareness of English loanword origins. *Journal of Bunkyo Gakuin University*, 169-174.

Biemiller, A. (1970). The development of the use of graphic and contextual information as children learn to read. *Reading Research Quarterly, 6*, 75-96.

Bloom, B. S., Hasting, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill Book Co.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Dordrecht, the Netherlands: Springer.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement for the human sciences (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education, Inc.

Brown, J. B., & Williams, C. J. (1985). Gairaigo: a latent English vocabulary base? *Tohoku Gakuin University Review: Essays and Studies in English Eibungaku, 76,* 129-146.

Carrell, P. L. (1988). *Interactive text processing. Implications for ESL/second language reading classrooms*. Cambridge: Cambridge University Press.

Carroll, B. J. (1985). Second language performance testing for university and professional contexts. *Second Language Performance Testing*, 73-88.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge, United Kingdom: Cambridge University Press.

Chabot, R. J., Zehr, H. D., Prinzo, O. V., & Petros, T. V. (1984). The speed of word recognition subprocesses and reading achievement in college students. *Reading Research Quarterly, 19,* 147-161.

Chang, C-H. (2012). *Instruction on pronunciation learning strategies: Research findings and current pedagogical approaches*. (Unpublished master's thesis) The Faculty of Graduate School, The University of Texas at Austen.

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

Chapman, M., & Newfields, T. (2008). The new TOEIC. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(2), 30-34.

Cheng, L., & Watanabe, Y. with Curtis, A. (Eds.). (2004). W*ashback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cobb, T. *Web Vocabprofile* [accessed 10 November 2014 from http://www.lextutor.ca/vp/], an adaptation of Heatley, Nation & Coxhead's (2002) *Range*.

Cowles, M. (1989). *Statistics in psychology: A historical perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses of old and new. *Psychometrika*, *53*, 63-70.

Culligan, B., & Gorsuch, G. (1999). Using a commercially produced proficiency test in a one-year core EFL curriculum in Japan for placement purposes. *JALT Journal, 21*, 7-27.

Davis, F. (1995). *Introducing reading*. London: Penguin.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords.* Clevedon & Philadelphia: Multilingual Matters.

English Testing Service (ETS). (2008). *A guide to understanding TOEFL iBT scores*. Retrieved on October 2014 from: www.ets.org/s/toefl/pdf/performance_feedback_brochure.pdf

English Testing Service (ETS). (2000). *TOEFL Monograph series: TOEFL 2000 reading framework: A working pape*r. Princeton, NJ: Author.

English Testing Service (ETS). (2007). *TOEFL iBT Tips- How to prepare for the TOEFL iBT*. Retrieved on December 2014, from www.ets.org/toefl/tips

English Testing Service (ETS). (2008). *TOEIC Can-Do guide executive summary listening & reading*. Retrieved on October 2014 from: www.ets.org/Media/Tests/Test_of.../TOEIC_Can_Do.pdf

English Testing Service (ETS). (2012a). *TOEFL official guide to the TOEFL test with CD-ROM, 4th Edition*. Princeton, NJ: Author.

English Testing Service (ETS). (2012b). *TOEIC shin koshiki mondaishu Vol 5*. [New official collection of past questions]. Tokyo: IIBC.

English Testing Service (ETS). (2013a). *Report on test-takers worldwide: The TOEIC listening and reading test*.

English Testing Service (ETS). (2013b). *TOEIC user guide: Reading and listening*. Retrieved on October 2014 from: www.ets.org/Media/Tests/.../TOEIC_User_Gd.pdf

English Testing Service (ETS). (2014). *TOEIC reading score descriptors*. Retrieved on October 2014 from: http://www.toeic.or.jp/english/toeic/guide04/guide04_02/score_descriptor.html

Eskey, D. E., & Grabe, W. (1988). Interactive models for second language reading: Perspectives on instruction. In P. L. Carrell, J. Devine, & D. E. Eskey, (Eds.), *Interactive approaches to second language reading*. Cambridge: Cambridge University Press.

e-Stat. (2014b). *Gakkou Kihon Chosa* [Basic School Statistics]. Retrieved on September 2014, from http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001015843

Embreston, S. E., & Hershberger, S. L. (1999). The new rules of measurement: What every psychologist and educator should know. In Embreston, S. E., & Hershberger, S. L. (Eds.), *Summary and future of psychometric methods in testing* (pp. 243-254). Mahwah, Educational Testing Service (ETS). Retrieved on January 2011 from http://www.ets.org/toeic/test_takers/listening_reading/about.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language, 30*, 210-233.

Fischer, W. P. J. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction*, *21*(1). 1095.

Ford, K. (2009). Principles and practices of L1/L2 use in the Japanese university EFL classroom. *JALT Journal*, *31*(1), 63-80.

Frost, R. (1994). Prelexical and postlexical strategies in reading: Evidence from a deep and shallow orthography. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 116–129.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 104-115.

Fulcher, G. (1997). An English-language placement test: issues in reliability and validity. *Language Testing*, *14*, 113-138.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessmen*t. London & New York: Routledge.

Funahashi, S. (1996). Prefrontal cortex and working memory. In K. Ishikawa, J. M. McGaugh & H. Sakata (Eds.). *Brain processes and memory* (pp.397-410).

Amsterdam: Elsevier.

Furnes, B., & Samuelsson, S. (2009). Phonological awareness and rapid automatized naming predicting early development in reading and spelling: results from a cross-linguistic longitudinal study. *Learning and Individual Differences, 21*(1), 85-95.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist, 6*, 126-135.

Gough, P. B. (1972). One second of reading. In J. F. Kavanagh, & I. G. Mattingly (Eds.), *Language by ear and by the eye* (pp.331-358). Cambridge, MA: MIT Press.

Gough, P. B., & Wren, S. (1999). Constructing meaning: The role of decoding. In J. Oakhill, & R. Beard (Eds.), R*eading development and the teaching of reading: A psychological perspective.* Malden, MA: Blackwell.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406.

Grabe, W. (2000). Reading research and its implications for reading assessment. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 226-260). Cambridge: Cambridge University Press.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow: Longman.

Hamada, K., & Koda, K. (2010). The role of phonological decoding in second language word-meaning inference. *Applied Linguistics, 31*, 513-531.

Hanna, P. R., Hanna, J. S., Hodges, R. E., & Rudorf, E. H. Jr. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement (USDOE Publication NO, 32008).* Washington, D.C.: U.S. Government Printing Office.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx .

Hei, K. C. (2009). Common loanwords identified in Japanese music magazines. *Polyglossia, 17*, 1-16.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing & Evaluation SIG Newsletter.* 6(3), 2-8.

Hirai, T. (2010). *Tesutomonndai kyouzai sairiyounosusume: TEASY rironhen* [The reuse of test items and materials: TEASY theoretical issues]. Tokyo: Maruzenpuranet.

Hirose, K. (2004). A study of test items on English terminologies for dentistry 7: Statistical item analysis in the B-index. *Meirinshishi, 7*(1), 3-6.

Hirose, K. (2005). A study of test items on English terminologies for dentistry 8: Distracter efficiency analysis. *Meirinshihi, 8*(1), 12-16.

Hoover, W., & Gough, P. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal, 2*, 127–160.

Hsiao, Y.Y., Shih, C. L., Yu, W. H., Hsieh, C. H., & Hsieh, C. L. Examining unidimensionality and improving reliability for the eight subscales of the SF-36 in opioid-dependent patients using Rasch analysis. *Quality of Life Research, 24*(2), 279-85

Hudson, T. (1998). Theoretical perspective on reading. *Annual Review of Applied Linguistics, 18*, 43-60.

Hughes, A. (2003). *Testing for language teachers*.(2nd ed.). Cambridge: Cambridge University Press.

Ikeda, O. (1994). *Gendai testoriron* [Modern test theory]. Tokyo: Asakurashoten.

Ichiyama, Y. (2014). A small-scale preliminary study of item bank development to include test items assessing orthographic and phonological processing skills. *Kitasato Review: Annual Report of Studies in Liberal Arts and Sciences, 19*, 101-112.

Ishikawa, S. (2008). Reaction time for processing vocabulary stimuli in the case of Japanese learners of English: A study on the relation between phonology and semantics. *Annual Review of the Chubu English Language Education Society, 37*, 17-24.

Ishikawa, S., & Ishikawa, Y. (2008). L2 proficiency and word perception. An fMRI-based study. *ARELE, 19*, 131-140.

Jarl, G. M., Heinemann A. W., & Hermansson L. M. N. (2012). Validity evidence for a modified version of the orthotics and prosthetics users' survey. *Disability and Rehabilitation: Assistive Technology, 7*, 469–478.

Jung, J. (2010). Second language reading and the role of grammar. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 9*(2), 29-48.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston, MA: Allyn & Bacon.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kadota, S. (2001). Yominiokeru jyouhousyoritanniha donoyounamonoka: tyannkinngunituite [The information process unit while reading: About chunking]. In S. Kadota, & C. Noro (Eds.), *Eigo readinno ninchi mechanism* [The cognitive process of English reading]. (pp.99-111). Tokyo: Kuroshio Shuppann.

Kadota, S. (2012). *Shadowing to ondoku no kagaku* [Science of shadowing and reading aloud].Tokyo: Cosmopia.

Kadota, S., Noro, T., & Shiki, M. (2010). *Eigo reading shidou handobook* [A manual to the teaching of English reading]. Tokyo: Taisyukan Shoten.

Kameyama, T. (1992). A study of pronunciation-spelling relation in the domain of English education. *Kikou Periodicals, 27*, p.55-62.

Karim, A. A., Shah, P. M., Din, R., Ahmad, M., & Lubis, M. A. (2014). Developing information skills test for Malaysian youth students using Rasch analysis. *International Education Studies, 7*(13), 112-122.

Katzir, T., Kim, Y., Wolf, M., Kennedy, B., Morris, R., & Lovett, M. (2006). The relationship of spelling recognition, RAN and phonological awareness to older poor readers and younger reading-matched control. *Journal of Reading and Writing, 19* (8), 845-872.

Kawaijyuku Educational Institution. (2014). *2014 nendo nyushi jyohou* [FY2014 Report on entrance examination]. Retrieved on September 2014, from http://www.keinet.ne.jp/topics/14/20140811.pdf

Kawasaki, M. (2013). A comparison of the decoding skills of children and adolescents: An examination of automaticity and error types. *Language Education & Technology, 50*, 1-21.

Kay, G. (1995). English loanwords in Japanese. *World Englishes, 14*(1), 67-76.

Kenneth, M. W. (2000). An exploratory dimensionality assessment of the TOEIC Test. *ETS Research Report, 14*, 1-28.

Kim, Y. H., & Goetz, E. (1994). Context effects on word recognition and reading comprehension of poor and good Readers: A test of the interactive-compensatory hypothesis. *Reading Research Quarterly, 29*(2).

Kimura, M. (1989). *The effect of five Japanese loanwords on the acquisition of the correct range of meanings of English words*. (Unpublished master's thesis). Department of Linguistics, Brigham Young University.

Kluitmann, S. (2008). *Testing English as a Foreign Language: Two EFL-tests used in Germany*. (Unpublished master's thesis). Albert-Ludwigs-Universitat Freiburg, Germany.

Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.

Koike, I. (1990). *A general survey of English language teaching in Japan*. Eigo Kyouiku Jittai Chousa Kenkyuukai.

Koike, I., Kinoshita, K., Narita, M., & Terauchi, M. (2004). *Dainigenngo syuutokukennkyuuno gennzai* [The present state of research in second language acquisition].Tokyo: Taisyukann-shoten.

Koizumi, R. (2005). *Nihonjintyukosei niokeru happhyougoitisiki no hirosatofukasa no kannkei* [Relationship between the vocabulary width and the depth of Japanese junior high school level students]. *STEP Bulletin, 17*, 63-80.

Koizumi, R., & Iiduka, H. (2010). Characteristics of neural test theory: comparison with classical test theory and Rasch modeling. *JLTA (Japan Language Testing Association) Journal, 13,* 91-109.

Kojima, M. (2010). Effects of word recognition speed, accuracy, and automaticity on reading ability. *ARELE: Annual Review of English Language Education in Japan, 21*, 151-160.

Koshimizu, I. (2010). Which is the better predictor of reading comprehension, lexical route or non-lexical route? *ARELE: Annual Review of English Language Education in Japan, 21,* 101-110.

Koyama, Y. (2013). Construction of an Item-bank Based on Science & Technology Corpora and Trial of a Computerized Adaptive Test Using Latent Rank Theory. *Research Report: Analysis of Domain Specific Expressions from Science & Technology Corpora and their Pedagogical applications. (The institute of Statistical Mathematics cooperative research report), 295*, 29-49.

LaBerge, D., & Samuels, S. J. (1974*).* Towards a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

Lawrence, R. (1998). Item banking. *Practical Assessment, Research & Evaluation, 6*(4), 1-3.

Lee, S., Yoshizawa, K., & Shimabayashi, S. (2006). The content analysis of the TOEIC and its relevancy to language curricula in EFL contexts in Japan. *JLTA Journal, 9*, 154-173.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. *Chae, S.-Kang, U.–Jeon, E.–Linacre, JM (eds.): Development of Computerised Middle School Achievement Tests, MESA Research Memorandum, 69.*

Linacre, J. M. (2006). *WINSTEPS Rasch measurement computer program (version 3.81.0).* Chicago: Winsteps.com

Linacre J. M. (2012). *Winsteps® Rasch Tutorial 4.* Available at: http://www.winsteps.com/a/winsteps-tutorial-4.pdf

Linacre, J. M. (2013). *A User's Guide to WINSTEPS®. Ministep: Rasch-model computer programs [program manual 3.80.0]*. Chicago: IL.

Lynch, B. K. (2003). *Language assessment and program evaluation*. Edinburgh: Edinburgh University Press.

Maeda, H. (2003). *Toutatumokuhyoukyouikuni muketa eigotesutono kaizen: kotenntekitesutoriron to koumokuoutourironn ni motozuite* [The effects of improvements in an English test on educational achievement: Based on classical test theory and item response theory]. *Hiroshimagaikokugokyouikukennkyu 6*, 131-140.

Martin, M. O., Mullis, I. V. S., & Foy, P (with Olsen J. F., Preuschoff, C., Erberber, E., & Galia, J.) (2008). *TIMSS 2007 international science study report: Findings from IEA's trends in international mathematics and science study at the fourth grade and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5-11.

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report Series, 1996*(1), i-18.

Miller, P. (2005a). Reading comprehension and its relation to the quality of functional hearing: Evidence from readers with different functional hearing abilities. *American Annals of the Deaf, 150,* 305-323.

Miller, P. (2005b). What the word recognition skills of prelingually deafened readers tell about the roots of dyslexia. *Journal of Development & Physical Disabilities, 17*, 369-393.

Minai, Y. (2000). *Kisokarawakaru eigo riiding kyohon* [English reading basic]. Tokyo: Kenkyusha.

Minai, Y. (2003). *Gakkoude oshietekurenai eibunnpou* [English grammar that is not taught in school]. Tokyo: Kenkyusha.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (1989). *Elementary and secondary education: The Course of Study*. Tokyo: Author.

Ministry of Education, Culture, Sports, Science and Technology MEXT. (1996). *Nijuisseikiwo tenboshita wagakunino kyouikuno arikatanituite: Daiichiji toushin* [Prospect of 21 century's education: First report]. Retrieved on September 2014 from
http://www.mext.go.jp/b menu/shingi/12/chuuou/toushin/960701.htm

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (1999). *Elementary and secondary education: The Course of Study*. Tokyo: MEXT.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2003a). *Elementary and secondary education: The Course of Study for foreign languages*. Retrieved on September 2014, from http://www.mext.go.jp/english/shotou/030301.htm

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2003b). *Eigogatukaerunihonnjinn no Ikuseintoameno Koudoukeikaku* [The action plan to cultivate "Japanese with English abilities"]. Retrieved on September 2014, from http://www.e-jes.org/03033102.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2003c). *Senior high school education: The Course of Study*. Tokyo: MEXT.

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2009). *The Course of Study for senior high schools guidelines explanation: Foreign languages (English)*. Retrieved from http://www.mext.go.jp/component/a_menu/education/micro _detail/ __icsFiles/afieldfile/2010/01/29/1282000_9.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2005). *Wagakuni no Koutoukyouiku no Souraizo* [Future of our tertiary-level education]. Retrieved on September 2014, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/attach/1335621.htm

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2006). *FY2006 White Paper on Education, Culture, Sports, Science and Technology*. Retrieved on August 2010, from http://www.mext.go.jp/b_menu/hakusho/html/hpac200601/index.htm

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2011). *The Course of Study for senior high schools guidelines explanation*. Retrieved on September 2014, from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/__icsFiles/afieldfile/ 2011/04/11/1298353_9.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2012a). *AO nyushi tou jishijyokyo nituit*e [Reports on AO nyushi]. Retrieved on September 2014, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo12/shiryo/__icsFiles/afieldfile/2013/01/09/1329266_1.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2012b). *Daigakunonyuugakuteiinn /nyuugakusyasuu tou no suiii* [Change of the intake of students and the number of university entrants]. Retrieved on September 2014, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo4/siryo/attach/__icsFiles/afieldfile/2012/06/28/1322874_2.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2014a). *Daigakunyugakusyasenbatu, daigakukyoiku no genjyo* [Report on university entrance and education]. Retrieved on September 2014, from http://www.kantei.go.jp/jp/singi/kyouikusaisei/dai11/sankou2.pdf

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2014b). *Daigakunyugakusyasenbatsu no kaizen* [Reform of university admission selection system]. Retrieved on September 2014, from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/gijiroku/attach/__icsFiles/afieldfile/2014/04/03/1346148_3.pdf

Ministry of Health, Labour and Welfare (MHLW) (2014). *Vital Statistics of 2013.* Retrieved on September 2014, from http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei13/dl/05_h2-2.pdf

Mitchell, J. (1999). *Measurement in psychology: Critical history of methodological concept*. Cambridge: Cambridge University Press.

Miyasako, N., & Takatsuka, S. (2004). What relationships do the efficiencies of phonological coding and lexical access have with reading comprehension for Japanese learners of English? *ARELE: Annual Review of English Language Education in Japan, 15,* 159-168.

Miyoshi, M., Naito, H., & Tozawa, R. (2011). *ESP Kyozai no dankaitekiensyu: Goigakusyu no kannten kara* [About ESP reading materials: From the perspective of vocabulary learning]. *ESP Hokkaido Journal, 1*, 14-36.

Moats, L. C. (2005). How spelling supports reading: And why it is more regular and predictable than you may think. *American Education*, *23*(4), 12-43.

Montgomery, P., & Connolly, B. (1987). Norm-referenced and criterion-referenced tests: Use in pediatrics and application of task analysis of motor skill. *Physical Therapy, 67*(12), 1873-1876.

Mullis, I. V .S., Martin, M. O., & Foy, P. (with Olsen J. F., Preuschoff, C., Erberber, E. Arora, A., & Galia, J.) (2008). *TIMSS 2007 international mathematics report: findings from IEA's trends in international mathematics and science study* at the *fourth grade and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nakamura, Y. (2007). A Rasch-based analysis of an in-house English placement test. *Proceedings of the 6th Annual JALT Pan-Sig Conference*, 97-109.

Nakamura, Y., & Ohtomo, K. (2002). *Testode gengonouryokuha hakareruka: genngotesuto detabunnseki nyuumon* [Whether a test can measure language ability: An introduction into language test data analysis]. Tokyo: Kirihara-shoten.

Nakano, M., Ueda, S., Oya, M., & Tsutui, E. (2004). Koumoku outou riron ni motozuita eigo kurasuwake testo kaihatsu niokeru ichikousatsu [A consideration into the development of English placement test using item response theory]. *Waseda Kyoiku Hyoron, 18*(1), 25-41.

Narita, K. (2007). Eigono tuduriji to hatuon no taiou [English orthographic and phonological correspondence]. *Niigata Studies in Foreign Languages and Cultures,12,* 55-68.

Narita, K. (2009). *Handbook of English spelling*. Tokyo: Sankeisya.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal*. *87*(2), 261-276.

Nation, I. S. P. (2003). The role of the first language in foreign language learning. *Asian EFL Journal, 5,* 1-8.

Noguchi, J. (2014). Contrastive analysis between Japanese and American English sound systems: from an articulatory setting perspective. *The Journal of Kanda University of International Studies 26,* 293-309.

Nuttall, C. (1996). *Teaching reading skills in a foreign language*. (2nd ed.). Oxford: Heinemann.

Obermeir, A. (2009). Improving English test questions through Rasch analysis. *Journal of Educational Research: Center for Educational Research and Training, Kyoto University of Education, 9,* 107-113.

O'Brien, B. A., Wolf, M., Miller, L.T., Lovett, M.W., & Morris, R. (2011). Orthographic processing efficiency in developmental dyslexia: An investigation of age and treatment factors at the sublexical level. *Annals of Dyslexia, 61*, 111-135.

Ohata, K. (2004). Phonological differences between Japanese and English: Several potentially problematic areas of pronunciation for Japanese ESL/EFL learners. *The Asian EFL Journal, 6*(4), 1-19.

Ohtomo, K. (1996). *Koumoku ouou riron nyuumon*. [An introduction to item response theory]. Tokyo: Taishukan Shoten.

Organisation for economic co-operation and development (OECD) (2007). *PISA 2006 science competencies for tomorrow's world*. Retrieved on December 2009, from http://www.oecd.org.dataoecd/30/17/ 39703267.pdf

Organisation for economic co-operation and development (OECD) (2009). *PISA data analysis manual: SPSS second edition*. Paris: OECD.

Organisation for economic co-operation and development (OECD) (2010a). *Learning trends: Changes in student performance since 2000*. Paris: OECD.

Organisation for economic co-operation and development (OECD) (2010b). *Learning to learn: Student engagement strategies and practices volume III*. Paris: OECD.

Organisation for economic co-operation and development (OECD) (2010c). *PISA 2009 results: What students know and can do*. Paris: OECD.

Oller, J. W. Jr. (1979). *Language tests at school*. London: Longman.

Otani, A., Yokoyama, H., & Bradford-Watts, K. (2014). *Pureisumennto testo niyoru syuujyukudobetu kurasuhennse nikannsuru houkokusyo* [Reports on placement test.] *Bulletin Paper of Kyoto Women's College, 62*, 27-50.

Otsu, Y. (2005). *Shogakko ni eigokyoiku wa iranai* [No English education is needed in primary school]. Tokyo: Keio University Press.

Portal Site of Official Statistics of Japan (e-Stat) (2014a). *Population estimates by age (5 year age group) and sex - June 1, 2014 (Final estimates), November 1, 2014 (Provisional estimates)*. Retrieved on September 2014, from http://www.stat.go.jp/english/data/jinsui/tsuki/index.htm

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Rogerson-Revell, P. (2011). *English phonology and pronunciation teaching*. London: Continuum.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (ed.), *Attention and performance* IV. NY: Academic Press.

Rumelhart, D.E. (1980). Schemata: the building blocks of cognition. In R. J. Spiro (Ed.) *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum.

Sang, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*(1), 31-57.

Samuels, S. J. (2006). Towards a model of reading fluency. In S. J. Samuels, & A. E. Farstrup (Eds.), *What research has to say about fluency instruction*. (pp.24-46). Newark, DE: International reading Association.

Sato, Y. (2015). An analysis of effectiveness of TOEIC practice with computer adaptive testing (u-CAT) as a web-based CALL system. *NUIS Journal of International Studies*, 91-96.

Sato, T., Nakagawa, T., & Yamana, T., (2007). The basic research of college-level English learners: what motivates them and how do they learn. *Bulletin of Tsukuba International University, 14,* 43-59.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York: Lang.

Schedl, M. A. (2010). Background and goals of the TOEIC listening and reading test redesign project. *TOEIC Compendium Report, 10*(2), 1-18.

Shepherd, J. W. (1996). Loanwords: A pitfall for all students. *The Internet TESL Journal, 2*. Retrieved on September 2014, from: http://iteslj.org/Articles/Shepherd-Loanwords.html

Shimizu, Y. (2003). Analyses of placement test. *Ritsumeikan Studies in Language and Culture, 14*(4), 181-188.

Shimizu, Y., Kimura, S., Sugino. N., Yamakawa, K., Ohba, K., & Nakano, M. (2003). Eigobunnpounouryoku hyoujyunntesuto no datousei shinraisei no kensyou to shineigobunnpounouryoku tesuto: Measure of English grammar [An investigation into the validity and reliability of standardized tests of English grammar and a new English grammar test, the Measure of English Grammar (MEG)]. *Seisakukagaku, 10*(3), 59-68.

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24,* 99-128 .

Shizuka, T. (2007) *Rasch modelling for objective measurement*. Osaka: Kansai Daigaku Syuppan.

Silberstein, S. (1994). *Techniques and resources in teaching reading*. New York: Oxford University Press.

Skaggs, G., & Wolfe, E. W. (2010). Equating designs and procedures used in Rasch scaling. *Journal of Applied Measurement, 11*(2), 182-195.

Smith, B. (2012). Pronunciation patterns of Japanese learners and their implications or teaching. *Polyglossia, 23,* 199-206.

Smith. F. (1971). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. New York: Holt. Rinehart and Winstone.

Smith. F. (1975). *Comprehension and learning: A conceptual framework for teachers*. New York: Holt. Rinehart and Winston.

Smith, R. M., Linacre, J. M., & Smith, Jr. E. V. (2003). Guidelines for manuscripts. *Journal of Applied Measurement, 4*, 198-204.

Souji, Y. (2007). *Jouhoukyouikuniokeru gendaitesutorironnno tekiyou: Koumokuoutourironnwo mochiita gakusyuutoutatudo hyoukasakuseino kokoromi* [The use of modern test theories in information education: An experiment on the development of achievement evaluation using IRT]. (Unpublished master's thesis). Keio University, Japan.

Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity and modularity. *Developmental Review, 10,* 72–100.

Stanovich, K. E. (1991). Changing models of reading and reading acquisition. In L. Rieben, & C. Perfetti (Eds.), *Learning to read: Basic research and its implications*. (pp.19-31). Hillsdale, NJ: Lawrence Erlbaum.

Sugimori, M. (2003). A study on the realities of enforcing placement tests, proficiency grouping, establishing achievement levels, and measuring attainment in Japanese colleges. *Policy Science Association, 10*(3), 3-26.

Sugiyama, Y. (2004). *Gakeppuchi Jyakushodaigaku monogatari* [Tale a small college of in a perilous position]. Tokyo: Chukosinnsyo Rakure.

Suski, P. M. (1931). *The phonetics of Japanese language*. Routledge Library Edition. Vol. 59. London: Routledge.

Takahashi, M. (2005). *The efficacy of grammar instruction in EFL classes in Japan*. (Unpublished doctoral dissertation). Kobe Shoin Women's University, Japan.

Takano, Y. (1995). Gengo to shikou [Language and Thought]. In Y. Ootsu (Ed.), *Ninchisinrigaku 3: Gengo* [Cognitive psychology] (pp.245-255). Tokyo: Tokyo Daigaku Shuppannkai.

Takebayashi, S., & Saito, H. (2008). *Eigo onsei gaku nyuumon* [Introduction to English phonology].Tokyo: Taishyushya.

Takemae, F. (2009). Daigaku eigokyoiku ha dokohei kunoka [The direction of university-level English education.]. *Gunsei, 2,* 5-10.

The Times Higher Education, *The Times Higher Education World University Rankings 2015-2016.* Retrieved on October 2016 from, http://www.timeshighereducation.co.uk/world-university-rankings/

Tokunaga, M. (2007). Students' assumptions for TOEIC classes. *JALT 2007*

Conference Proceedings, 257-271.

Toyoda, H. (2002). *Koumoku hanou riron nyumonhen* [An introduction to item response theory]. Tokyo: Asakura-Shoten.

Urquhart, S. W., & Weir, C. (1998). *Reading in a second language: Process, product and practice*. London: Longman.

Vellutino, F., Tunmer, W., Jaccard, J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading,11*(1), 3–32.

Vollmer, H., & Sang, F. (1983). Competing hypotheses about second language ability: a plea for caution. In J. Oller (Ed.), *Issues in language testing research* (pp. 29-79). Rowley, MA: Newbury House.

Watanabe, N., & Noguchi, Y. (1999). *Soshiki shinri sokutei ro*n [Psycholinguistic measurement theory]. Tokyo: Hakuto-shobou.

Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27*(1), 71-92.

Wei, M., & Zhouh, Y. (2013). Transfer of phonological awareness from Thai to English among grade three students in Thailand. *The Reading Matrix 13*(1), 1-13.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wistner, B., & Sakai, H. (2007). Rasch analyses of English language placement tests. *JALT 2007 Conference Proceedings*. 1045-1055. Tokyo: JALT.

Wistner, B., Sakai, H., & Abe, M. (2009). An analysis of the Oxford placement test and the Michigan English placement test as L2 proficiency tests. *Bulletin of the Faculty of Letters, Hosei University, 58*, 33-44.

Woodford, P. E. (1982). An introduction to TOEIC: The initial validity study. *TOEIC Research Summary*.

Wray, D., & Lewis, M. (1997). *Extending literac*y. London: Routledge.

Wright, B. D. (1997). Fundamental measurement for outcome evaluation. *Physical Medicine and Rehabilitation: State of the Art Reviews, 11*(2), 261-288.

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

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Yoshida, H. (2009). Analyzing English placement test: From perspectives of language testing. *Osaka Keidai Ronsyu, 60*(2), 103.